

NATURAL LANGUAGE SUMMARIZATION OF TEXT
AND VIDEOS USING TOPIC MODELS

by

Pradipto Das

February 1, 2014

A dissertation submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

Department of Computer Science and Engineering

Copyright by
Pradipto Das
2014

The dissertation of Pradipto Das was read and approved¹ by the following committee members:

Rohini K. Srihari

Professor of Computer Science and Engineering

State University at New York at Buffalo, USA

Thesis Adviser

Chair of Committee

Sargur Srihari

Distinguished Professor of Computer Science and Engineering

State University at New York at Buffalo, USA

Aidong Zhang

Professor of Computer Science and Engineering

State University at New York at Buffalo, USA

Chair of the Department of Computer Science and Engineering

¹Signatures on file in the Graduate School.

Acknowledgements

Acknowledgements go out to all *entities* mentioned below. Many more have been instrumental in my journey and deserve a mention. However, in the interest of keeping the list short, I have selected only the top few from each category.

[i] Mentors and faculty:

- ▶ Suprabhath Chakraborty (*Late*), my tutor of English during my middle and high school years. He was a literally a *wise*, man a parallel of whom I have not yet met till now. Although he was my English tutor, we never practised any school texts but rather texts and poems in literature and history which are profound. Without his mentoring, I probably would not have taken risks to do what is right.
- ▶ Arun K. Sanyal (*Late*), Emeritus Professor of Mathematics, IIT Kharagpur. I received mentoring from him only for a few months during my late high school years during which he had a profound influence on me and on my quest for pursuing the right kinds of educational avenues in life. I am and always will be reminded of his quote “I can solve this problem in many ways, but what is the most elegant and right way?” Thanks also goes out to Dr. D. P. Mandal, now Associate Professor at Indian Statistical Institute, Kolkata without whom my journey into UB would not have been possible. He also inculcated in me a sense of iterative perfection in technical writing.
- ▶ Matthew J. Beal, then Assistant Professor of Computer Science at SUNY Buffalo, for introducing me to a field of study which ultimately had a major influence on the development of the materials in this thesis.
- ▶ Jordan Boyd-Graber, now Assistant Professor of Computer Science at University at Maryland, College Park, for giving me a key hint on the use of regularization in topic models in one of our personal communications.
- ▶ Carl Alphonse, now Associate Teaching Professor of Computer Science at SUNY Buffalo, not only for demonstrating the best practices of teaching computer science to younger people, but also giving me a treat which included his fabulous home-made goat *biryani*, the best one I have had in over a decade.
- ▶ Jason J. Corso, now Associate Professor of Computer Science at SUNY Buffalo, for introducing me to topic models on text and images and generously supporting me as an RA for the last two years but more importantly for introducing me to *Crème brûlée* which to my mind is infinitely better than *Tiramisu*.
- ▶ Matthew D. Jones, now Associate Director and Lead Computational Scientist, CCR, SUNY Buffalo, for being that “cool physicist” amidst computer scientists.

[ii] Colleagues:

- ▶ Smruthi Mukund, now applied data scientist at EBay: We worked together on a couple of projects and one of the key things she intrinsically preached was the need to develop a quick and simple baseline. It has helped me later on since in the beginning I was quite blind to the elements of simplicity.
- ▶ Chenliang Xu and Richard F. Doell, my colleagues and co-authors in our 2013 CVPR paper: They can redefine slogging!
- ▶ John Chen, now consultant for AT&T Labs-Research: Simply put, he is like a *navy seal* whose mission is to process Natural Language. He is an extremely polite person and mentor to talk to as well.
- ▶ Scott McCloskey from Honeywell ACS Research Labs and Sangmin Oh from Kitware Inc.: I sincerely appreciate their help in processing videos which made our TRECVID submission possible and also their numerous comments over regular tele-conferences.
- ▶ Greg Mori's group at SFU, particularly Arash Vahdat and Kevin Cannons: Without their help in video processing, the final chapter of this thesis would not have been complete.

[iii] Family: Special thanks goes out to brother Debanjan Das and sister-in-law Tania Ghosh. I have been very fortunate to receive both moral and financial support as and when needed without which this journey would have been far from complete; indeed, the old car which he had given me has proven to be the most valuable gift; my parents, Sudipta and Prabir Das, for their dedicated and unconditional long-distance support; my late paternal grandparents, Suniti Das and Pramatha Lal Das; my late maternal grandmother Priti Das for her culinary skills in showing how the statistics of proportions can be used to create magic for the taste buds and my late maternal grandfather Sudhir Kumar Das for his passion to tinker with anything mechanical, specially automobiles and lathe machines, and an unparalleled ability to lead a balanced everyday life consistently. I have often reminded myself of his feats when he once helped build a mechanical instrumentation workshop named "Blue Earth Workshop" at Jadavpur University where he helped assemble parts of a small aircraft given to him as a gift from the US Army then camped at South Kolkata towards the end of World War II.

[iv] Adviser: Rohini K. Srihari, now Professor of Computer Science, SUNY Buffalo and also Prof. Sargur Srihari, Distinguished Professor of Computer Science, SUNY Buffalo for helping me out at a certain key low point in my life and not giving up on me.

[v] Life and its obstacles: For teaching me that when you cannot turn the tables sometimes, take a step back and turn the room itself.

I also like to take this opportunity to thank Lucy Vanderwende of Microsoft Research and Enrique Alfonseca of Google Research for several useful discussions on the applicability of bulleted list summaries during a meeting of the Text Analysis Conference, 2011 including the latter author's permission to re-use their new scores on a dataset and useful comments on a first initial draft of Chapter 5.

Finally, I thank Edward J. Sobczak, Kevin P. Cleary, Ken Smith, Mary J. Gallo and Eugenia Smith and the CSE administrative staff. People like them tirelessly work behind the scenes to make sure that we don't worry about software and systems maintenance or the periodic flow of paychecks or reimbursements and they have done a remarkable job in that respect.

Contents

Acknowledgments	iv
Table of contents	vi
List of figures	xi
List of tables	xvi
Abstract	xix
1 Introduction	1
1.1 Background	1
1.1.1 Summarization: Distilling Information	4
1.1.2 Probabilistic Browsing Models	9
1.2 Contributions of this thesis	17
1.2.1 Chapters 1 and 2	17
1.2.2 Chapter 3	17
1.2.3 Chapter 4	17
1.2.4 Chapter 5	18
1.2.5 Chapter 6	18
2 Introductory Concepts	20
2.1 Exponential Family Distributions	20
2.2 Maximum Likelihood, Sufficient Statistics and Conjugate Priors	26
2.2.1 Maximum Likelihood	27
2.2.2 Sufficient Statistics	29
2.2.3 Conjugate Priors	31
2.2.4 Asymptotics and MLE	32
2.3 How much training data is necessary?	33
2.4 Bayes Estimator and its Relation to Posterior	36
2.5 Bayesian vs. Frequentist	38
2.6 Expectation Maximization (EM) and variational Bayesian EM (VBEM)	39
2.6.1 Finding a lower bound to the log likelihood	43
2.6.2 EM for <i>Exact</i> Unconstrained Optimization	44
2.6.3 EM for <i>Approximate</i> Constrained Optimization	45

2.6.4	EM for Maximum-A-Posteriori Learning and its Connection with VBEM	47
2.6.5	EM for Bayesian Learning using Variational distributions	48
2.6.6	Mean Parameters	49
2.6.7	Significance of Mean Parameters on Inference Problems	50
2.6.8	What does Forward Mapping of Canonical to Mean Parameters mean?	51
2.6.9	Conjugate Duality	52
2.6.10	Mean Field and Tractable Families	54
2.6.11	Mean Field Procedure	57
2.7	Latent Dirichlet Allocation and Variational Bayesian EM	61
2.7.1	The Success Behind Latent Dirichlet Allocation	62
2.7.2	LDA: How much data is necessary to learn the model parameters?	63
2.7.3	Exponential Family Representation, Mean Field and Variational Bayes	64
2.7.4	Kullback-Leibler divergence in LDA	67
2.7.5	The E Step Inner Loop in the Mean Field Optimization of LDA	67
2.7.6	Gibbs Sampling versus Variational Bayes in Topic Models	69
3	Learning to Summarize using Sparse Coherence Flows	75
3.1	Introduction	75
3.2	Related Work	77
3.3	Centering Theory and Sparse Coherence Flows	79
3.3.1	Discourse Analysis: Centering Theory	79
3.3.2	Sparse Coherence Flows	82
3.4	Learning to Summarize using Utterance Topic Models	84
3.4.1	Utterance Topic Model	84
3.4.2	Parameter Estimation and Inference	85
3.4.3	Latent variable inference	86
3.4.4	Maximum Likelihood Parameter estimation	87
3.5	The Learning To Summarize model—LeToS	87
3.5.1	Parameter Estimation and Inference in the Extended Model	89
3.6	Experimental Setup and Results	89
3.6.1	Description of the Datasets	89
3.6.2	Qualitative Topic Analysis and Summarization Performance	90
3.7	Summary	97
3.8	Appendix	98
3.8.1	Derivations for the UTM model	98
3.8.1.1	Inference on Variational Parameters	99
3.8.1.2	Model Parameter Estimation	100
3.8.2	Derivations for the LeToS Model for Summarization	102
3.8.2.1	Inference on Variational Parameters	103
3.8.2.2	Maximum Likelihood Parameter Estimation	105

4	Bi-Perspective Topic Models	107
4.1	Introduction	107
4.1.1	Descriptions of Annotations in Datasets	109
4.1.2	Improving Existing Tag Topic Models	110
4.1.3	Applications and Quantitative Measures	111
4.2	Related Work	113
4.3	The Proposed Tag Squared LDA Models	114
4.3.1	Latent variable inference	115
4.3.2	Maximum Likelihood Parameter estimation	118
4.3.3	Algorithms for Implementation	119
4.4	Results and Discussions	121
4.4.1	Model Loglikelihoods on Held-out Test Data	121
4.4.2	Automatically Evaluating Suggested Tags From Image Captions	126
4.4.3	Automatically Evaluating Named Entity Relationship Discovery	128
4.4.4	TagLDA Revisited	129
4.4.5	Evaluating Tag-Topic Models through Extractive Multidocument Summarization	130
4.5	Summary	132
4.6	Appendix	133
4.6.1	Inference on Variational Parameters	134
4.6.2	Model Parameter Estimation	136
5	Using Bi-Perspective Topic Models and Rhetorical Structure Trees to generate Bullet Lists	139
5.1	Introduction	139
5.1.1	Datasets	142
5.1.2	Global Topic Models and Local Sentential Models	143
5.1.3	Rhetorical Structure Trees as a Local Model	144
5.2	Related Work	147
5.2.1	Existing Topic Model Based Summarization Approaches	147
5.2.2	Existing Linguistic and Vector Space Model Based Summarization Approaches	148
5.3	The Tag-Topic Models	149
5.3.1	Data Preparation for the Tag-Topic Models	150
5.3.2	Descriptions of the Bi-Perspective Tag-Topic Models	151
5.3.3	Mean Field Inference	152
5.3.4	Parameter Estimation	153
5.3.5	The Need for Informative Priors for Topic Proportion Distributions	154
5.3.6	Model Log Likelihoods	157
5.3.7	Tag-Topic Model Evaluation through Multi-document Summarization	158
5.4	The Local Models	161
5.4.1	Document Set Models—Bags of Key Terms	162
5.4.2	Event Classification Performance	162
5.4.3	Sentence Dependency Graphs and RS-trees	163
5.4.4	Sentence Compression using RS-tree Spans	164
5.5	Summarization Experiments	168

5.5.1	Basic Summarization Algorithms	168
5.5.2	Evaluation Settings	170
5.5.3	Results	170
5.5.3.1	Performance of Baseline Models on TAC 2010A and 2011A Datasets	170
5.5.3.2	Proposed Model Performance on TAC 2010A Dataset	172
5.5.3.3	Proposed Model Performance on TAC 2011A Dataset	175
5.5.4	Performance on Update Summarization	178
5.6	Summary	180
5.7	Acknowledgements	182
6	Summarizing Videos into Natural Language Text	183
6.1	Introduction	183
6.1.1	Dataset Description	186
6.1.2	Evaluation Measures	187
6.2	Related Work	187
6.3	The Proposed Models	189
6.3.1	Inference on Latent Variables	191
6.3.2	Model Parameter Estimation	195
6.4	Experimental Setup and Results	196
6.4.1	Held-out Log Likelihoods and Topics	197
6.4.2	Translating Related Words to Videos	199
6.4.3	Translating/Summarizing Videos To Text	201
6.4.4	Event Classification	202
6.4.4.1	Natural Language Generation	202
6.5	A Thousand Frames in just a Few Sentences – Enhancing Summary Relevancy	204
6.5.1	An Information Extraction and Summarization System Framework	206
6.5.1.1	Low level: Topic Model	206
6.5.1.2	Middle level: Visual Concept Extraction to Language	208
6.5.1.3	High level: Final Lingual Descriptions	210
6.5.2	Further experiments and Results	211
6.5.2.1	Datasets and Features	211
6.5.2.2	Quantitative Evaluation	213
6.5.2.3	Qualitative Examples	214
6.6	Summary	214
6.7	Acknowledgements	216
6.8	Appendix	216
6.8.1	Some Important Derivations	216
6.8.2	Object Bank Vocabulary	222
6.8.3	Implementation	223
6.8.4	Algorithm and Pseudocodes	223
6.8.5	Important low level features from videos	228
6.8.5.1	HOG3D Features	228
6.8.5.2	Color histogram features	228

6.8.6 Code snippets	229
7 Conclusion and Talking Points	233
Bibliography	235
Index	251

List of Figures

1.1	Do we speak all that we see? Human summaries of a short video on rock climbing	5
1.2	Some topics from patent and legal documents about rockets and propulsion	6
1.3	Cartoon illustration of exploratory topic analysis	7
1.4	Snapshot of an article on “Fog” from Wikipedia	8
1.5	Cartoon illustration of the central modeling questions answered in this thesis	9
1.6	Faceted topics from Wikipedia (see Chapter 4)	11
1.7	Some examples of topic models used in the various chapters of this thesis.	12
1.8	Mapping topics on Wikipedia articles on common visual objects to fMRI patterns. Un- seen fMRI patterns are then used to predict words that appear to be semantically related through latent topics.	14
1.9	Changes in the BOLD (Blood Oxygen Level Dependent) patterns in a small area at the back of the visual cortex of the human brain (the region inside the green ellipses shown in the flattened brain scans) when the same subject is shown different movies. Reproduced here with permission from the authors in [Nishimoto et al., 2011]	15
1.10	Some examples of keyword predictions for test videos using MMLDA (Figure 1.7c) with action features	16
2.1	Graphs of the objective function of a two component Gaussian-mixture model over a set of mean parameters and over a set of sample points. The surface normals are shown as tiny blue arrows	26

2.2	Simple illustration on forward mapping from canonical parameters to mean parameters and vice versa. We consider a distribution over the discrete random variable $\mathbf{Z} = \{z_1, z_2, z_3\}$ as a product of three independent Bernoulli distributions. The realizations of the binary random variables z_i form the corners of the marginal polytope (shown here without any constraint cutting planes which would typically arise out of the constraints on each of the μ_i s $\in [0, 1]$). Given a biased coin with probability of heads being 0.8, we are more likely to observe realizations of \mathbf{Z} which have more ones. The red shadow bubbles beneath each of the red nodes in the $[0, 1]$ cube on the left are indicative of this. The forward mapping is shown at the bottom right half of the illustration where the darker rows are indicative of more probable configurations of \mathbf{Z} . The value of the mean parameter $\boldsymbol{\mu}$ will tend to the true mean with a value of $(0.8, 0.8, 0.8)$ of the generating distribution as we observe an infinite number of samples with more and more configurations of two or more ones. This problem of forward mapping to find the mean parameters from the observations generated from the distribution with canonical parameters is equivalent to the problem of finding $\boldsymbol{\theta}_\zeta(\boldsymbol{\mu})$ through the backward mapping which in this illustrative case has a closed form solution. The Bernoulli nature of the $q(z_i)$ s is also verified by the form of the log partition function of $\boldsymbol{\mu}$ which in this case is the negative of the entropy of the Bernoulli distribution with mean parameter $\boldsymbol{\mu}$. Note that if we observe all configurations of \mathbf{Z} only once then the forward mapping of $\boldsymbol{\mu}$ yields $(0.768, 0.768, 0.768)$ for which the backward mapping causes $\boldsymbol{\theta}_\zeta(\boldsymbol{\mu})$ to be greater than one. To avoid this possibility, Lagrange multipliers are used to constrain $\boldsymbol{\mu}$ thereby making $\boldsymbol{\theta}$ valid.	54
2.3	Graphical depiction of the hidden-variable / parameter factorization. Fig. 2.3a: The original generative model for N hidden and observed variables. Fig. 2.3b: The exact posterior graph for $p(\mathbf{Z}, \boldsymbol{\theta} \mathbf{X})$. The z_i and z_j pairs are not directly coupled, but interact through $\boldsymbol{\theta}$ [Shachter, 1998]. By De Finetti’s theorem, the hidden variables are conditionally independent of one another, but only given the parameters. Fig. 2.3c: the posterior graph after the variational approximation between parameters and hidden variables where the arcs between parameters and hidden variables are removed. Due to this type of factorization the hidden variables become independent because of i.i.d assumption. A similar kind of “product” factorization is shown in Fig. 2.5	56
2.4	Cartoon illustrations highlighting the key features of mean field optimization	60
2.5	The graphical model for the Latent Dirichlet Allocation	62
2.6	LDA with hyperparameters over topic multinomials	70
3.1	Utterance Topic Model: Extending LDA to include coarse coherence properties of discourse segments	86
3.2	Extending the UTM model to the Learning To Summarize model (LeToS) by assuming that sentences are distributions over coarse coherence properties of discourses which in our case are GSRts	88
3.3	Empirical proportions of GSRts based on a maximum of three-sentence window for the Yahoo! Answers and TAC08 Newswire datasets	90

3.4	Strengths of the multinomial parameters of the topic distributions over GSRts for the in-house Yahoo! Answers dataset. Three sample queries are shown just beneath the plot that highlight the words which are highly probable for the topics over GSRts: ρ_1 , ρ_2 and ρ_3	92
3.5	Strengths of the multinomial parameters of the topic distributions over GSRts for the TAC 2008 newswire dataset	93
3.6	Empirical proportions of GSRts in the summaries obtained by various models on the Yahoo! Answers and TAC08/09 newswire datasets. Summaries from LeToS have each sentence coming from a different document.	95
4.1	Graphical model representations of one supervised topic model, two existing tag topic models, one extended tag topic model and two new tag squared topic models	110
4.2	Mean field representation of the METag ² LDA model	116
4.3	Cross-Validation results on DUC 2005 newswire data (higher is better in 4.3a and 4.3c): (4.3a) ELBO-Validation DUC 2005 GSRTNe; (4.3b) Minimum of differences in ELBO across topics of Corr-METag ² LDA to corrLDA and METag ² LDA for GSRTNe tagging; (4.3c) ELBO-Validation DUC 2005 GSRTPos; (4.3d) Minimum of differences in ELBO across topics of Corr-METag ² LDA to corrLDA and METag ² LDA for GSRTPos tagging	122
4.4	Training and test negative ELBO plots of tag topic models on the Wiki data (Lower is better). In each K -group, the models from left to right are MMLDA, TagLDA, Corr-MMLDA, METag ² LDA and Corr-METag ² LDA	123
4.5	Training and test negative ELBO plots of tag topic models on the Amazon Review (AR) data (Lower is better). In each K -group, the models from left to right are MMLDA, TagLDA, sLDA, Corr-MMLDA, METag ² LDA and Corr-METag ² LDA	124
4.6	(4.6a) The best ontological inverse path length measure between suggested DL tags from image captions and ground truth Wikipedia categories for the test set in Fig. 4.4b and (4.6b) PERSON Named Entity-pair coverage ratio to baseline from DUC 2005 Newswire data	126
4.7	Cross-Validation results on DUC 2005 (DUC05) newswire data (higher is better in all figures); Fig. 4.7a: ELBO-Validation on DUC05 with GSRTNe tagging - not showing TagLDA; Fig. 4.7b: Better ELBO for TagLDA on DUC05 with Ne and GSRT tagging; Fig. 4.7c: ELBO-Validation on DUC05 with GSRTPos tagging - not showing TagLDA; Fig. 4.7d Better ELBO for TagLDA on DUC05 with Position tagging. X-axis represents the values of K	129
4.8	Fig. 4.8a: Optimum value of the prior parameter α for Wikipedia dataset; Fig. 4.8b: Optimum value of the prior parameter α for DUC 2005 dataset with GSRTNe/prioritized GSRT tagging (TagLDAGSRt). X-axis represents the values of K	130

4.9	ROUGE SU4 scores (higher is better) of 250 word summaries for the models in fig. 4.1 - the scores are averaged over all docsets in the DUC 2005 dataset (GSRTNe/Ne/prioritized GSRt tagging); Fig. 4.9a: ROUGE SU-4 scores when each sentence in the summary is atleast 20 words (GSRTNe/Ne/prioritized GSRt tagging); Fig. 4.9b: ROUGE SU-4 scores when each sentence in the summary is atleast 30 words (GSRTNe/Ne/prioritized GSRt tagging); Fig. 4.9c: ROUGE SU-4 scores when each sentence in the summary is atleast 20 words (GSRTPos/Position tagging); Fig. 4.9d: ROUGE SU-4 scores when each sentence in the summary is atleast 30 words (GSRTPos/Position tagging). X-axis represents the values of K	131
5.1	An article for the query “sleep deprivation” showing a document level and a word level perspective with some shallow and deep linguistic structures	143
5.2	Rhetorical Structure tree of the second sentence in Fig. 5.1	145
5.3	Our proposed summarization system architecture using global tag-topic models and local linguistic models.	146
5.4	Graphical model representations of the tag-topic models used in modeling the corpus	149
5.5	Some negative examples of topic anotation on <i>sentences</i> from news documents. The topic annotations are shown as color coded text. The text within the first row of bubbles indicate the topics which annotate the sentences. These are obtained by finding $k^* = \arg \max_{k \in \{1, \dots, K\}} \lambda_{d,m,1:K}$ for the TagLDA and METag ² LDA models and $\arg \max_{k \in \{1, \dots, K\}} \sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,1:K}$ for the Corr-METag ² LDA model. The text in the second row of bubbles for the Tag ² LDA family of models denote the topic of the sentence obtained directly from $\arg \max_{k \in \{1, \dots, K\}} (\gamma_{d,1:K} - \alpha_{1:K})$	156
5.6	Evidence Lower BOunds (ELBO)s of the tag-topic models on the TAC 2010A and 2011A datasets – Lower is better.	158
5.7	Five-fold Cross-validation accuracies of the local models (bags-of-key terms) on event category classification of TAC 2010A/2011A documents. The legend is read from left to right and from top to bottom corresponding to the bar groups for each of the TAC base collections.	163
5.8	Event classification cross-validation accuracies on TAC 2010A and TAC 2011A dataset using per-document latent topic proportions from the tag-topic model as features, i.e. $\gamma_d - \alpha$, for different number of topics - Higher is better. Note that for symmetric prior over θ_{dS} , the vector α becomes the scaler α	164
5.9	Choosing thresholds for selecting RS-tree spans using Parzen density estimates of the cosine similarity values of the words in the spans to those in the feature set.	167
6.1	An example of the task of video summarization	183
6.2	Do we speak all that we see? Human summaries of a short video on rock climbing	184
6.3	An example of vocabulary intrusion in the task of video summarization. Best viewed with magnification	185

6.4	Graphical model representations of existing topic models and proposed extensions— Figs. 3d and 3e. In this paper, we extend the model in Fig. 3c i.e. the Corr-LDA model in [Blei and Jordan, 2003] with Normal-Wishart priors over parameters for real valued observations as well.	190
6.5	GIST features projected on to 15 (left), 30 (middle) and 60 (right) dimensions & visualized in two dimensions using t-SNE [Maaten and Hinton, 2008]	197
6.6	Test ELBOs on events E001-E005 in the Dev-T set. Lower is better.	197
6.7	Test ELBOs on events E006-E015 from Dev-T set. Lower is better	198
6.8	Prediction ELBOs on first 5 event for Dev-T set. Lower is better.	198
6.9	Prediction ELBOs on events E006-E015 on Dev-T set—lower is better. Best viewed with magnification	199
6.10	Average test ELBOs on all events in the Dev-T set for different topics. Lower is better	200
6.11	Topic 10 for the “Flash mob” event from a ten-topic MMGLDA	200
6.12	Topic 6 for the “Flash mob” event from a ten-topic Corr-MMGLDA	200
6.13	Topic 10 for the “Flash mob” event from a ten-topic Corr-MMGLDA	200
6.14	Event detection accuracies for cross-validation (light gray bars) and test (dark gray bars) with different features	202
6.15	Bag of keywords and sentence translations from our proposed MMGLDA ($K=20$) for some clips from the first five events from the Dev-T set	203
6.16	A framework of our hybrid system showing a test video being processed and lingually described through top-down concept detection and bottom-up keyword summarization	204
6.18	Examples of DPM based concept detector.	208
6.17	Prediction ELBOs from the two topic models for the videos in TRECVID dataset. Lower is better	208
6.19	Lingual descriptions from tripartite template graphs consisting of concepts as vertices	210
6.20	Qualitative results from MER12 test and our “YouCook” dataset. Only top 5 sentences from our system are shown.	215

List of Tables

2.1	Key differences between EM and VBEM. The variable θ is the parameter of the model which we wish to find. The observations are denoted by \mathbf{X} and the hidden state variables are denoted by \mathbf{Z} .	56
3.1	Transition relations holding between pair of adjacent sentences due to the centers	80
3.2	Snapshot of a sentence-word GSR grid view of a document on “Health and Safety” category	83
3.3	Snapshot of a sentence-word GSR grid view of a document on “Attacks” category	84
3.4	Some topics from LDA for TAC 2008	90
3.5	Some topics from UTM for TAC 2008 matching those in Table 3.4	91
3.6	Comparison of DUC 2005 ROUGE Results	94
3.7	A snapshot of sentences that focuses on mudslides killing men	95
3.8	Different short ≈ 120 words summarized answers for “Are Sugar substitutes bad for you?”	96
4.1	Document with word level “Position” tags and document level image caption word tags. The ellipses (...) indicate substantial skip of paragraphs [source: http://en.wikipedia.org/wiki/Syringa]	107
4.2	Document with word level “Named Entity” annotations and document level “Named Entity as well as semantic and syntactic role transition” tags	108
4.3	Document with word level emotion tags and document level product feature tags	109
4.4	Model features and their comparison	111
4.5	Topics and correspondences from the Corr-METag ² LDA for the Wikipedia data for $K = 200$	112
4.6	Symbols used in this chapter and their meaning	114
4.7	Topics and correspondences from the Corr-METag ² LDA for the Wikipedia data for $K = 200$	125
4.8	Sample evidence Chains for DL Tag suggestions from image captions to ground truth category labels from the Tag ² topic models	126
4.9	Three sample topics from the Corr-METag ² LDA for the Amazon Product Review (AR) data for $K = 200$. Topic 49 highlights the problem with correspondence when there are more than a few competing topics for explaining the DL metadata	127
4.10	DUC 2005 dataset: Related PERSON named entity pairs and evidence from documents	128
4.11	DUC 2005 docset, latent topic and generated summary. Sentences are at least 20 words long	132

5.1	Sample Docset IDs, their corresponding information needs and categories, important nouns and important verbs from the TAC 2011 Guided Summarization dataset. The nouns and verbs are obtained using an automatic part-of-speech tagger.	140
5.2	Latent topics from the TAC 2011A dataset for $K = 80$ using asymmetric Dirichlet prior α over θ_d in TagLDA. Terms within square brackets [] are Named Entity phrases that are treated as single concatenated tokens.	155
5.3	ROUGE-SU4 scores and confidence intervals of the summaries from the MEAD system for all base and update collections from TAC Summarization tasks.	159
5.4	ROUGE-SU4 scores for TAC 2010A/2010B datasets obtained from sentence ELBO based summarization using tag-topic models. K is the number of topics.	160
5.5	ROUGE-SU4 scores for TAC 2011A/2011B datasets obtained from sentence ELBO based summarization using tag-topic models. K is the number of topics.	161
5.6	RS-tree spans and their importance.	165
5.7	ROUGE-SU4 and ROUGE-2 scores for summaries from baselines and human summarizers for TAC 2010A and TAC 2011A base collections.	171
5.8	ROUGE-SU4 and ROUGE-2 scores for summaries from topic model baselines for TAC 2010A and TAC 2011A base collections.	173
5.9	ROUGE-SU4 and ROUGE-2 scores from our proposed models for TAC 2010A.	174
5.10	ROUGE-SU4 and ROUGE-2 scores from our proposed models for TAC 2011A.	175
5.11	100-word summaries for the harder information need on “Sleep Deprivation” in TAC 2011A dataset. Individual sentences are square bracketed. A (*) indicates that the bullets or sentences belong to the same document. Notice how the CLASSY summary is drawn towards a “cardiovascular” bias while our full sentence summary is drawn towards a “napping” bias. Incidentally, “nap” has a strong focus in D1127E-A.	177
5.12	ROUGE-SU4 and ROUGE-2 scores of summaries from local model baselines and top performing peer systems for TAC 2010B and 2011B update collections.	178
5.13	ROUGE-SU4 and ROUGE-2 scores from our proposed models for TAC 2010B.	179
5.14	ROUGE-SU4 and ROUGE-2 scores from our proposed models for TAC 2011B.	180
5.15	Few sample docset IDs, queries and categories from the TAC 2011 dataset. The docset-tfistf-noun+verb feature set is shown for base and update collections.	181
6.1	Meanings of the variables used in the models	189
6.2	Three latent topics for two events from proposed five-topic Corr-MMGLDA and Corr-MGLDA-PDS models. The topics from Corr-MGLDA-PDS are similar as a result of high values of α_k obtained after running Corr-MGLDA-PDS on scaled i.e. normalized data. The topics from Corr-MMGLDA are qualitatively far superior and indicates sub-events of the “Wedding ceremony” and the “Boarding” events	198
6.3	$\ln \frac{\alpha_k}{ \Lambda_k }$ values for topics in event 8	200
6.4	Individual and average ROUGE-1 scores on the events—best results from 10/20 latent topics are shown. The value of n represents the top- n most probable keywords. A (*) means significantly worse performance at 95% confidence to {MM,MMG}LDAs. These results are only reported for the same hyperparameter settings.	201

6.5	Average word prediction 1-gram recall for different topic models with 200 topics when the full corpus is used. The numbers are slightly lower for lower number of topics but not statistically insignificant.	208
6.6	ROUGE-1 precision and recall scores for MER12 test set. A (−) for the R-T-KW column means significantly lower performance than the next 2 columns. The bold numbers <i>in the last column</i> is significantly better than the previous 3 columns in terms of recall. The bold numbers <i>in P-D-S column</i> are significantly better than those in P-T-KW column. A (*) in columns 3, 4 or 5 means significantly lower than P-T-KW. A 95% confidence interval is used for significance testing.	213
6.7	ROUGE scores for our “YouCook” dataset	214

Abstract

Probabilistic topic models have recently become the cornerstone of unsupervised exploratory analysis of text documents using Bayesian statistics. The strength of the models lie in their modularity—random variables can be introduced or modified to suit the requirements of the different applications. Many of these models however consider modeling only one particular view of the observations such as treating documents as a flat collection of words ignoring the nuances of the different classes of annotations which may be present in an implicit and/or explicit form.

We extend a few existing unsupervised topic models such as Latent Dirichlet Allocation (LDA) to model documents which are annotated from two different perspectives. The perspectives consist of both a word level (e.g. part-of-speech, affect, positional etc.) tag annotation and a document level (e.g. crowd-sourced document labels, captions of embedded multimedia) highlighting. The new models are dubbed as the Tag²LDA class of models whose primary goal is to combine the best aspects of supervised and unsupervised modeling learning under one framework. Additionally, the correspondence class of Tag²LDA models explored in this context are state-of-the-art among the family of parametric tag-topic models in terms of predictive log likelihoods. These models are presented in Chapter 4.

The field of automatic summary generation is increasingly gaining traction and there is a steady rise in demand of the summarization algorithms that is applicable to a wide variety of genres of text and other kinds of data as well (e.g. video). Producing short summaries in a human readable form is very attractive particularly for very large datasets. However, the problem is NP-Hard even for smaller domains such as summarizing small sets of newswire documents. We use the Tag²LDA class of models in conjunction with local models (e.g. extracting syntactic and semantic roles of words, Rhetorical Structure trees, etc.) to do multi-document summarization of text documents based on information needs that are *guided* by a common information model. The guided summarization task, as laid out in recent text summarization competitions, aims to cover information needs by asking questions like “who did what when and where?” We also have successfully applied multi-modal topic models to summarize domain specific videos into natural language text directly from low level features extracted from the videos. The experiments performed for this task are described in detail in Chapter 5.

Finally, in Chapter 6, we show that using topic models it is possible to outperform keyword summaries generated by annotating videos through state-of-the-art object recognition techniques from computer vision. Summarizing a video in terms of natural language generated from such keywords in context removes the laborious frame-by-frame drawing of bounding boxes surrounding objects of interest—a scheme which is required for annotating videos to training a large number of object detectors. The topic models that we develop for this purpose instead use easily available short lingual descriptions of entire videos to predict text for a given domain specific test video. The models are also novel in handling both text and video features particularly with regards to multimedia topic discovery from captioned videos whose features can belong to both discrete and real valued domains.

Chapter 1

Introduction

“There is no great mystery in this matter,” he said, taking the cup of tea which I had poured out for him; “the facts appear to admit of only one explanation.”

“What! you have solved it already?”

*“Well, that would be too much to say. I have discovered a suggestive fact, that is all. It is, however, very suggestive.” – **Sherlock Holmes: The Sign of Four***

1.1 Background

The problem of compressing text documents into short summaries has been studied from an early time with one of the seminal works authored by Luhn [Luhn, 1958] in the 1950s where he noticed that “the significance factor of a sentence is derived from an analysis of its words.” It has been generally observed that the saliency induced by high frequency of occurrences of the words and their positions in an article can be used to rank sentences to generate an abstract or summary. These two important concepts of salience and position have played a significant role in the development of the Natural Language Processing (NLP) field in recent years particularly in the development of sophisticated probabilistic models. Furthermore identifying salience of words has been immensely important in complementary research areas such as Information Retrieval (IR) [Butcher et al., 2010], Machine Learning [Bishop, 2006] with positions of words and their semantic information playing very important roles in structured predictions for various NLP tasks [Jurafsky and Martin, 2000].

On the other hand, in the realm of probabilistic graphical models [Jordan, 2004], there has been an abundance of models which capture the underlying structure of the data through some assumed representation of data generation with observed and latent (or indicator) variables and parameters tied through causal arcs. The assumptions of the structural representation arise out of the specific problem being solved and rarely there is any value to address a single unified model which addresses multiple loss functions.

The value of these probabilistic models is to compute posterior distributions over the latent variables that have the power to summarize the data in a way which can be interpretable by an end user. The modes of these distributions, when applied to mixed corpora which includes text, reveal an inherent semantic clustering of words closely resembling a distinct topical structure. Such exploratory models, although

rigorous and sophisticated, have an inherent problem of capturing higher order dependencies within observations such as complete parses of sentences. Capturing such complex dependencies incurs very high computational complexity and thus are not suitable for producing descriptions which are *complete* natural language summaries.

Although solving the problem of multi-document summarization exactly in polynomial time is hard particularly with unsupervised techniques and for arbitrary domains, the advantages of having a multi-document summary cannot be overlooked particularly in this age of both new and repetitive information overload. Obvious applications include: summarizing books and novels for essays, scientific papers etc.; summarizing answers in an online question answering service like Yahoo! Answers; summarizing news in an online news service like Google News; summarizing discussion threads in an e-learning framework that may improve teaching effectiveness of the instructors. To give a feel for the impact of the multi-document summarization problem, a few real world application scenarios are highlighted where a readable summary is often better than a ranked list of objects.

- (i) **Online search service:** Consider a scenario where a user is using an internet search engine, searching for certain topics of interest. The information retrieval system returns a ranked list of documents by maximizing their relevance to the query. Without having to check these links one by one, it is often useful to generate a unified summary of the information contained in these documents and present it to the reader. Short multi-document summaries are also advantageous in the mobile search setting, where a multi-document summary w.r.t. a query can greatly reduce the power consumed by clicking each individual link and checking whether the document truly reflected the information need. A sample static user interface prototype reflecting this idea can be found in [the author's website](#).
- (ii) **Online news evolution modeling:** In an online news service setting such as Google News or Yahoo! News, news stories are clustered i.e. labeled as “politics,” “sports,” “science,” etc. from different news agencies and are presented to the reader. The different news stories in one cluster, presumably on the same topic, have major redundancy in their contents due to way the news stories are aggregated from different news sources and so it is hard to find what is new and relevant information in this setting automatically. Multi-document summarization is very applicable in this setting providing users with the right information at their fingertips. However, biased human evaluation of what constitutes a perfect summary is a major bottleneck to generate the *best* summary automatically. A very recent supervised approach based on re-ranking of sentences using diversity modeling of M -best solutions has been proposed by Lin and Bilmes [Lin and Bilmes, 2012] and applied in this context. However, the results are not definitive at this point as to whether there exist simpler and more scalable solutions which perform just as well.

Another interesting problem is to model the summarization of news articles following the temporal dimensions i.e., given that an user had read a summary in one timeline, how would he be presented with a “newer” summary in the next timeline? This *modeling of evolution of summaries over time* allows us to predict how an event is going to unfold over time. Topic models with temporal dynamics have been in focus recently [Blei and Lafferty, 2006, Wang et al., 2008] where linear dynamical system principles are applied to couple the latent topics in one timeline to another. However, it remains to be seen if these models can be extended to rigorously formulate a model of summary evolution and also address scalability issues. Tackling this problem is a part of our

research which is to be pursued in the recent future.

- (iii) **Online market place and social networks:** In another scenario of an online marketplace like Amazon.com, customers write many reviews for many products. It is thus worthwhile not only to summarize the reviews based on their ratable aspects [Titov and McDonald, 2008] but also to summarize the products themselves. This problem differs from the generic multi-document summarization problem in that each review is typically very short and opinionated. Local coherence is generally not observed in such small amount of product review text. Some recent work on very concise summarization of reviews can be found in the recent papers by Ganesan et al. [Ganesan et al., 2012, Ganesan et al., 2010].

Summarization in a social network setting like FacebookTM, TwitterTM or a professional network setting like LinkedInTM is also important and is receiving some focus only very recently [Ramage et al., 2010]. However, the fundamental problem in this setting is that of entity-pair summarization—given two entities present a chain of evidence as a bulleted list summary that precisely answers the question of whether there is a connection which is “strong enough.”

- (iv) **Video summarization:** Natural language summarization of real-life videos is a vastly understudied problem where the goal is to summarize a video into free flowing natural language text. Although a lot of advances have been made in supervised and weakly supervised learning of object and action detectors in images [Li et al., 2010a, Felzenszwalb et al., 2010] these problems are still considered open in the computer vision research community [Makadia et al., 2008]. Further, the consideration of what objects and actions need to be trained and how they may contribute to generating a final summary is still more harder and subjective. An initial investigation into this latter problem is presented in Chapter 6. The problem of efficient semantic understanding of videos has immense significance in robotic applications. Solving this problem also opens up the possibility of improving searchability and automated semantic understanding of scenes (useful in driver-less car or in-home robot scenarios) with little or no textual metadata.

One of the major contributions of this thesis is the extension of a previously state-of-the-art topic model—Latent Dirichlet Allocation (LDA) [Blei et al., 2003] to model documents which are tagged from two different perspectives. The perspectives consist of both a word level (e.g. parts-of-speech, named entity classes, affect categories, positional etc.) annotation and a document level highlighting (e.g. crowd-sourced document labels, captions of embedded multimedia) [Das et al., 2011]. In all future references, we dub these models as Tag²LDA. At a higher level, such models combine the best aspects of supervised and unsupervised modeling under one unified *exploratory* framework.

Secondly, we apply these new topic models (i.e. global models looking at a tag-annotated corpus as a whole) in conjunction with local models (e.g. extracting syntactic and semantic roles of words, Rhetorical Structure trees [Mann and Thompson, 1988], etc.) to perform multi-document summarization of text documents that is *guided* by a set of relevant questions which are some predefined attributes of some event category [Das and Srihari, 2011]. Such “guidance” provide important clues to help in improving information coverage in the resulting summaries.

In the final part of our research we successfully use topic models to summarize videos of a specific domain into natural language text directly from quantization of low level filters. These new topic models involving multimedia are unique w.r.t joint modeling of both text and low visual features, the latter being

represented in both real valued and discrete domains—they help both in semantic clustering of video frames [Das et al., 2013a] as well as predicting unstructured text for an unknown domain specific video [Das et al., 2013b].

1.1.1 Summarization: Distilling Information

Document summarization is a fairly old problem dating back to the 1960s and 1970s [Luhn, 1958, Edmundson, 1968]. Since then major improvements in both storage and speed of computational hardware had led to tremendous growth of algorithms mimicking human intelligence and reasoning in the areas of natural language processing, computer vision, document processing and retrieval and more. However, in the field of multi-document summarization, the performances of the systems to date have not been able to catch up with the quality of human summaries when evaluated manually. Advances, though, have been made where system summaries score at par with human summaries when automatic evaluation with ROUGE [Lin and Hovy, 2003], a recall oriented metric to measure information need, is used.

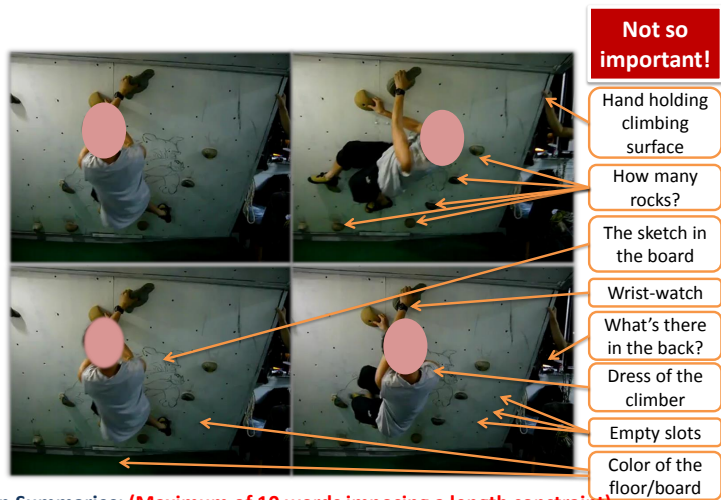
More recently it has been shown in [McDonald, 2007] that summarization consists of satisfying three separate objectives—i) relevance of the summary sentences to the query; ii) non-redundancy between the sentences in the summary and iii) the summary obeys a length constraint—and that optimizing all three of them simultaneously is an NP-Hard problem.

In general it is expected that all summarization systems respect an unified information model¹ which has recently been popularized by the Text Analysis Conference (TAC). The information model asks the system summaries to output summarized information along some more fine grained vertical aspects and across more coarser event categories. For example, there are several categories in which news documents can be categorized—i) Accidents and Natural Disasters; ii) Attacks (Criminal/Terrorist); iii) Health and Safety; iv) Endangered Resources; v) Investigations and Trials (Criminal/Legal/Other); vi) Science and Technology; vii) Entertainment and many more. These are the more coarser vertical categories. For each category, there are much more finer information nuggets or *aspects* which when properly identified and included in a summary vastly improves its relevance to the unified information model. Examples of such aspects for the Health and Safety category are: What is the issue; Who is affected by the health/safety issue; How they are affected; Why the health/safety issue occurs; Countermeasures, prevention efforts.

Clearly questions like “What is the issue?” are very subjective in nature and when expressed concretely through various choices of words, the same central issue can be paraphrased in many different ways. Although not all documents belong to such well defined categories, nor do they contain the well defined finer aspects as mentioned here, however, the question of what information makes the document relevant to a query is answered by the quality of content reflected in the gist of the document. This notion of a summary is extremely popular now-a-days in the setting of search engine result displays where users are more likely to click an url depending on the usefulness of the corresponding snippet generated as a function of the query – although this is more of a single document summarization problem where the goal is to predict which portions of the document are relevant ahead of query processing time since the full document is usually not stored in a search engine index.

The problem of paraphrasing is more evident from Figure 1.1 which shows a rock climbing video (only four manually chosen key frames are shown). Five human annotators have been asked to summarize the contents of this video in natural language text in about ten words. The ten word limit has

¹<http://www.nist.gov/tac/2011/Summarization/>

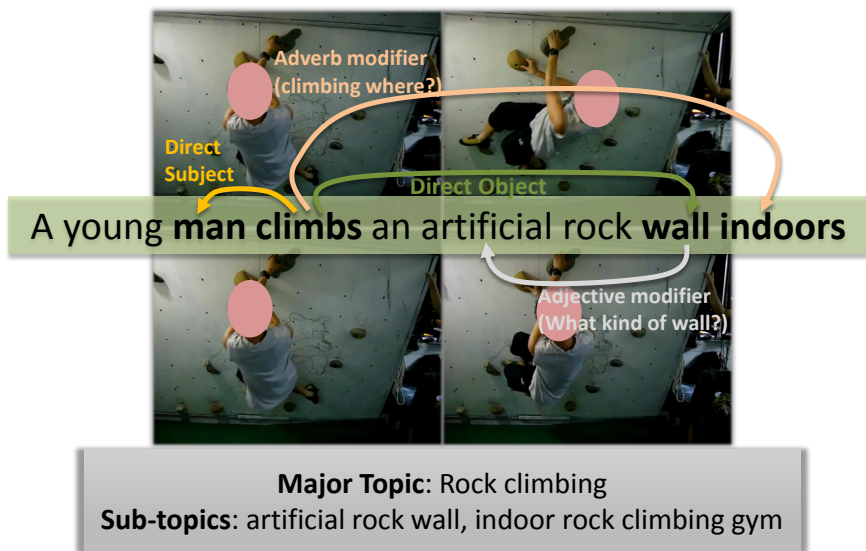


Multiple Human Summaries: (Maximum of 10 words imposing a length constraint)

- | | |
|--|---|
| 1. There is a guy climbing on a rock-climbing wall. | 4. A person is practicing indoor rock climbing. |
| 2. A man is bouldering at an indoor rock climbing gym. | 5. A man is doing artificial rock climbing. |
| 3. Someone doing indoor rock climbing. | |

Summaries point toward information needs!

(a) Short summaries from human annotators on an indoor rock climbing video



(b) Another ground truth short summary on the same video with complex semantic dependencies between the words in the sentence. The major and minor topics are also reflected here.

Figure 1.1: Do we speak all that we see? Human summaries of a short video on rock climbing

been used following some statistics of human annotations in the training dataset (Chapter 6) from which this video has been selected. However, we can hypothesize that the length constraint simulates a time constraint where subjects are asked to *speak* about the video in a *short amount of time*. Although the experiment do not specifically consist of a query on which to base the descriptions, the human summaries clearly reveal a highly probable set of words—“rock climbing”—which can easily be understood to be the query for which the video is most relevant.

Interestingly, none of the human subjects have put any importance on the concepts mentioned in

the right margin of Figure 1.1a within the orange bubbles such as the “color of the wall/board,” “the color of the vest which the climber is wearing,” “how many rocks are there in the artificial wall?” and so on. However, these objects appear in almost every frame of the video and yet recognizing them through computer vision techniques and subsequent use of frequency based text summarization techniques as in [Nenkova et al., 2006a] clearly leads to incorporation of extra information that is typical of the query drift phenomenon [Buttcher et al., 2010].

Intuitively, a lingual summary of a video is primarily focusing on the prominent actions and objects (or nominal concepts) associated with such actions. These concepts can be thought of as *active concepts*. Most other objects play the role of *passive concepts*. Our intuition is verified quantitatively in Chapter 6 where low level action features from videos improve video event classification significantly and in Chapter 5 where a few verbs and nouns selected based on *sf-isf* (*sf*: sentence frequency; *isf*: inverse sentence frequency) weights from a set of newswire documents on a particular event actually have the potential to recover the exact query for that event.

Even though the sentence shown in Figure 1.1b from a sixth annotator is simplistic in construction from the perspective of inferring the semantic dependency of the words in it, the human process by which low level visual signals get translated to high level language that respects the precise relative arrangement of objects is not well understood. It is implied that color, scenes, actions and objects all are mapped to some vocabulary that the subject has acquired over years and the clues about the arrangement of objects in the visual world and the grammar of the language together aid in the generation of a final fluent and coherent lingual description or summary. The coherence is more prominent when multiple sentences make up a single summary where the center of attention within sentences is maintained in a manner so as to aid in easier inference of the meaning of the whole passage while reading or listening to.

The summaries in Figure 1.1 highlight two important questions for multimedia (video) to text summarization—i) What are the prominent objects and actions

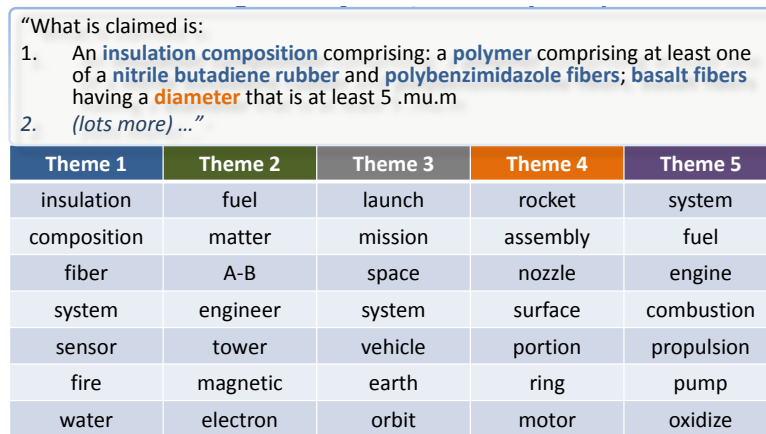


Figure 1.2: Some topics from patent and legal documents about rockets and propulsion

that are principle to describing the main event in the video and ii) Does an extended set of prominent concepts (akin to query expansion) truly increase the relevancy of the summaries in terms of information need coverage? Additionally answering the second question leads to better understanding of what queries can be generated for a corresponding test video.

We look at this general problem and answer these

questions in Chapter 6. We develop probabilistic topic models of unstructured corpora with some inherent domain knowledge and then use those models to emphasize parts of the learnt vocabularies which help in constructing relevant summaries.

Probabilistic browsing models are usually expressed through depiction of some assumed causal relationships between random variables (or groups of random variables). A directed graphical model visualizes these causalities and provides a succinct description of the factorization of a joint distribution: nodes denote random variables—shaded for observed variables and unshaded for hidden variables; edges denote possible dependencies between random variables; and plates denote replication of a substructure, with appropriate indexing of the relevant variables. The random variables which are drawn either outside of any plate which is not embedded in any other enclosing plates or in plates with a fixed multiplicity (which can be infinite for non-parametric models) are the parameters of a model.

An example of the exploratory power of such a model (named Latent Dirichlet Allocation by Blei et al. [Blei et al., 2003]) is shown in Figure 1.2 where a large set of patent documents on rockets and propulsion has been automatically annotated by ascribing individual words through possible latent themes which seems to group semantically related words quite effectively. Latent themes are represented as distributions over some observations—be they words from a text vocabulary or some other vocabulary such as codebooks of quantized features for images, videos or audio.

Topics for an unseen patent document fitted to a learnt LDA model are shown in Figure 1.2. The document is mostly about the “insulation” topic (topic 1) and less about the “rocket parts” topic (topic 4). Note that each word in the document implicitly has a distribution over topics and although a single mode of those distributions has been chosen for color coding and annotation, it is needless to say that each document is represented not by a single topic (as would typically happen in the case of a singular value decomposition of the word document count matrix) but by a mixed membership over the set of topics where only a few topics have very high probability. Figure 1.3 shows the essence of exploratory analysis of text documents w.r.t topic discovery.

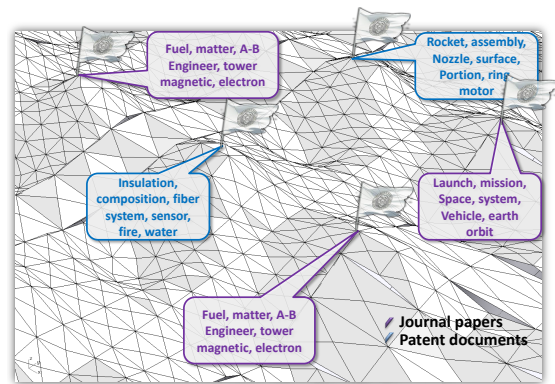


Figure 1.3: Cartoon illustration of exploratory topic analysis

In this cartoon illustration, we imagine that initially there is a two dimensional grid or graph where the nodes are labeled with words from some vocabulary and the edges encode the semantic closeness of the neighboring nodes. Our goal is then to traverse this grid and swap the nodes using the observed document partitioning as clues to reweight the edges. The confidence of this edge reweighting allows *peaks* to be formed which appears to put semantically related words closer together. Of course, we do not “see” the landscape in three dimensions from the top but we explore the space in two dimensions and “shade” regions of it similar in spirit to finding the map of a maze from within the maze itself.

Note that we have made up the labels “insulation” and “rocket parts” in the preceding paragraph and it is a non-trivial problem to find the best description of a latent topic that goes beyond the set of a few most probable terms. Although some work along this direction has been pursued in [Mei et al., 2007b, Blei and Lafferty, 2009], in this thesis, we go one step beyond by labeling topics through multimedia in terms of captioned images as well as video frames. To achieve such labeling of the latent structure of a corpora, we needed to retain as much relevant embedded and meta information (such as the multimedia and their captions, text annotations, document tags etc.) as possible into the model itself.

This leads to the incorporating domain knowledge which far surpass the ambiguous expressive power of words alone.

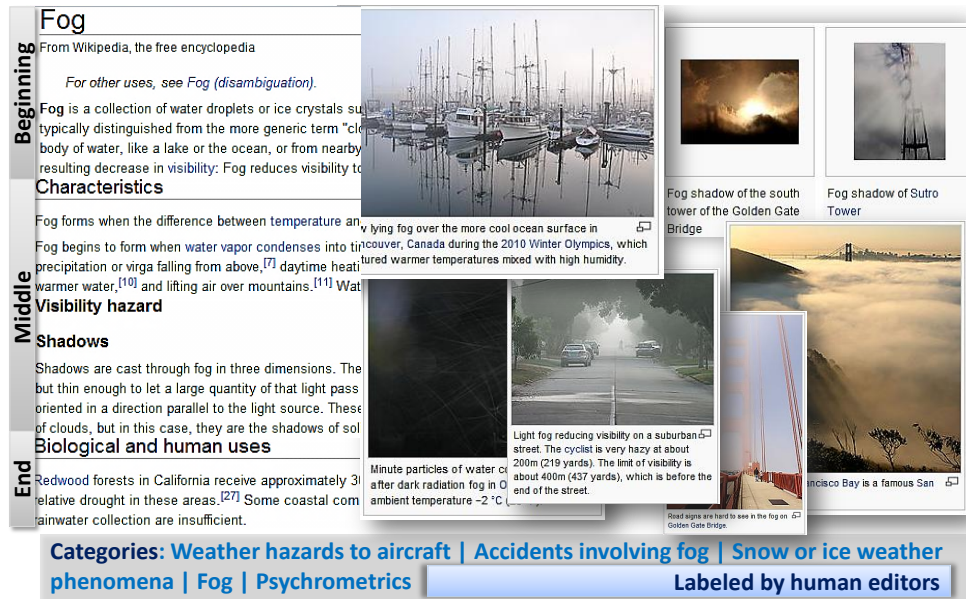
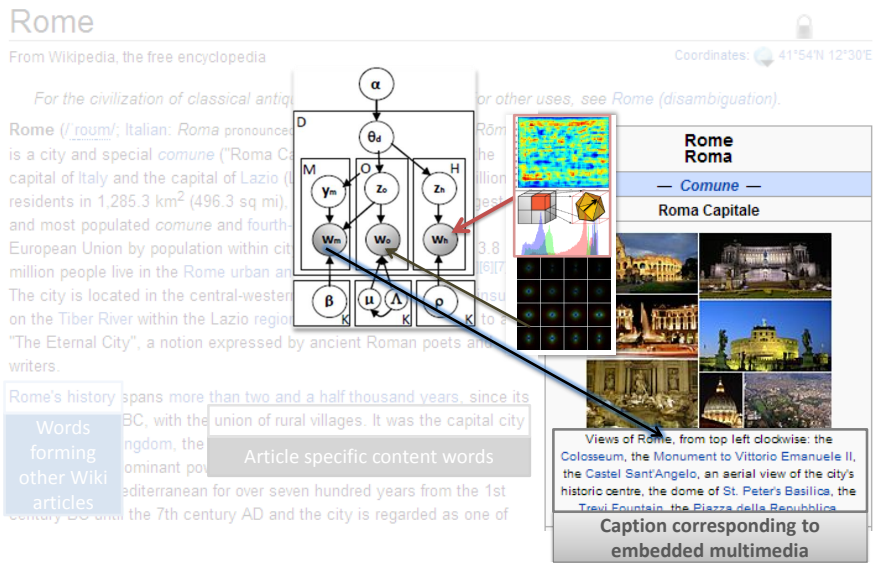


Figure 1.4: Snapshot of an article on “Fog” from Wikipedia

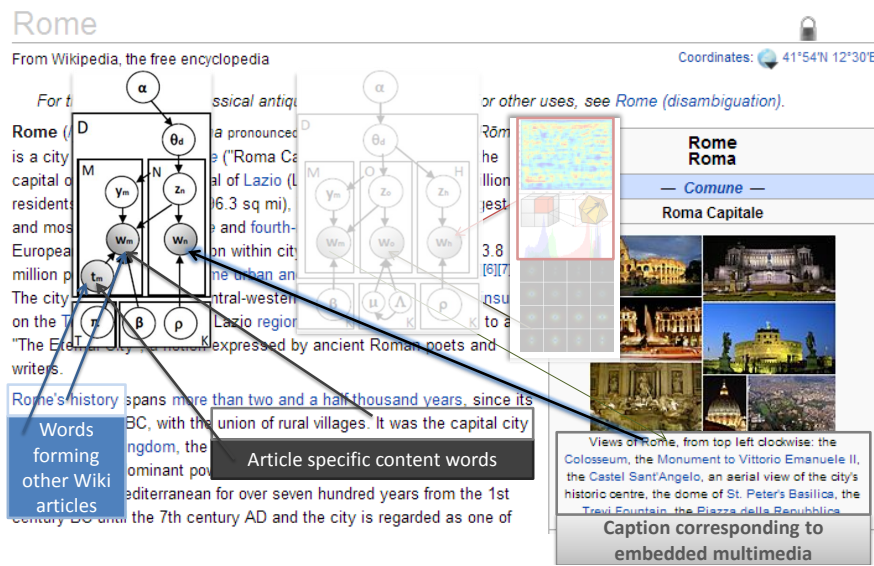
We now take a specific example of domain knowledge as incorporated into the models mentioned in this thesis (see Chapter 4) using a Wikipedia article on “fog” shown in Figure 1.4. Each of these Wikipedia documents have an inherent structure to them; for example, the document shown in Figure 1.4 shows words constituting sections, embedded images and captions highlighting the corresponding sections as well as category labels manually assigned by human editors that reflect the different ways in which the document can be classified. The main content can even further be annotated based on whether the words are part of an URL signifying a possible Wikipedia article title. The field of structured prediction in the area of Natural Language Processing (NLP) is devoted to such automatic annotations based on their contexts. Such annotations clearly suggest a “word level” perspective.

Alternately, the category labels and even the captions of the multimedia reflect a “document level” perspective. Often these labels at the document level carry an implicit topical structure which is complementary to the topical structure of the main contents. Thus, although individual documents can have semantic structure, the corpora as a whole may lack a data driven coherent organization which exploits the relevancy of document annotation from different perspectives. In Chapter 4, we address this problem by incorporating these perspectives into a family of single probabilistic browsing models.

This thesis answers the following central questions: i) How can the embedded multimedia in a document be translated to captions which represent a summary of the multimedia? ii) How well does the caption text (or document metadata in general) relate to the main document content? and iii) How do we build models which organize the many instances of i) and ii). Figures 1.5a and 1.5b show the depiction of the first two questions as a cartoon illustration using a Wikipedia article on “Rome.” We now briefly mention such models which also appear in subsequent chapters along the course of this thesis.



(a) Predicting caption as a summary from embedded multimedia in a Wikipedia article using a translation model (See Figure 1.7i and Chapter 6)



(b) Modeling a Wikipedia article document using word level annotations and document level multimedia captions (See Figure 1.7f and Chapter 4)

Figure 1.5: Cartoon illustration of the central modeling questions answered in this thesis

1.1.2 Probabilistic Browsing Models

Modeling Joint Distribution of Corpus through Topic Models: Topic models have become the corner-stone of unstructured document analysis [Blei, 2004, Griffiths and Steyvers, 2004]. A principle theme of such models is to find latent topics of interest where the latent topics are just some discrete distributions (often Multinomials) over some observations (usually textual words).

Often these latent topics are referred to as latent spaces and these latent spaces have very similar

properties of semantic re-organization of words which frequently co-occur together [Blei, 2004, Girolami and Kabán, 2003]. Models like LDA are often compared to Latent Semantic Analysis (LSA) [Landauer et al., 1998] with the latter as the former’s non-probabilistic version. Although LSA has recently been shown to be not so effective in information retrieval tasks [Atreya and Elkan, 2011] in terms of precision and recall, it is mostly due to the nature of the objective functions which each model optimizes. For information retrieval tasks, the *tf-idf* weighting scheme is an approximation to the probabilistic modeling of query relevance (See Chapter 10 in [Buttcher et al., 2010]).

One of the main advantages of topic models such as LDA are their modularity—basic models can be extended by introducing random variables which encode some more extended aspect of the observations. These aspects arise out of a need to introduce more structure to explain a certain phenomenon in the data. We also want to find latent thematic structures through the use of some statistical moments thereby eliminating the need for introducing a data dependent distance metrics. This is a major motivation to use distributions where the mean parameters of the models identify information which is “central” to the observations and also how well do new sample points deviate from this central tendency.

It is much harder to incorporate extensions into a fixed algebraic model such as LSA. LSA uses Singular Value Decomposition (SVD) and inherently such a decomposition assumes that a single topic is being allocated to a document. This assumption is the basis of poor generalization performance [Blei et al., 2003]. Additionally, using SVD, one usually obtains a span and not a basis over topic vectors which further diminishes generalization power. On the other hand, domain knowledge often demands extensions to the basic topic models – for an example see Figure 1.7f. A sample output from this extended model is shown in Figure 1.6. In the figure, we observe how the two distributions over words (topics on “artillery” and “tofu”) are conditional on the positions of the sections in which the words occur in the Wikipedia documents. The Figure also shows the same latent distributions over image caption vocabularies in the column named “Tag Suggestions.” The “Correspondence” column shows possible statistical associations of caption words to the words in the main document body. The rows showing $\beta_{(\cdot)}^{\text{learned}}$ are the learnt topics (with ids (\cdot)) which have been marginalized out of the influence of the tag distributions over the same set of words. Finally, images corresponding to the captions can be used to label the latent topics in a clear and visually pleasing way following the proverbial adage—“a picture is worth a thousand words.”

The exploratory view of data analysis involves the analysis of patterns based on some assumptions on how the patterns are generated. The patterns can be a pattern of annotated words in documents which gives rise to a semantic structure, energy patterns corresponding to a summarized representation of an image that is not conditioned upon the specific arrangements of objects, link structure of documents in the Web, degrees of separation graphs of users in a social network and so on. No matter what the dataset be or what the problem scenario be, three basic assumptions of machine learning are always satisfied—i) There is a pattern ii) The target function is not known and iii) There is data to learn from. In the supervised scenario the target function is only known for the training samples (which we usually refer to as labels) and in the unsupervised scenario it is completely unknown as is the case for the scenarios handled in this thesis.

Figure 1.7 shows the models that has been used in this thesis in the context of different problems. The square plates enclose the observed (shaded) and hidden variables as well as some of the parameters and denote repetitions of the random variables within. The parameters such as β , π , μ , Λ and ρ have a fixed multiplicity as they do not grow with the data.

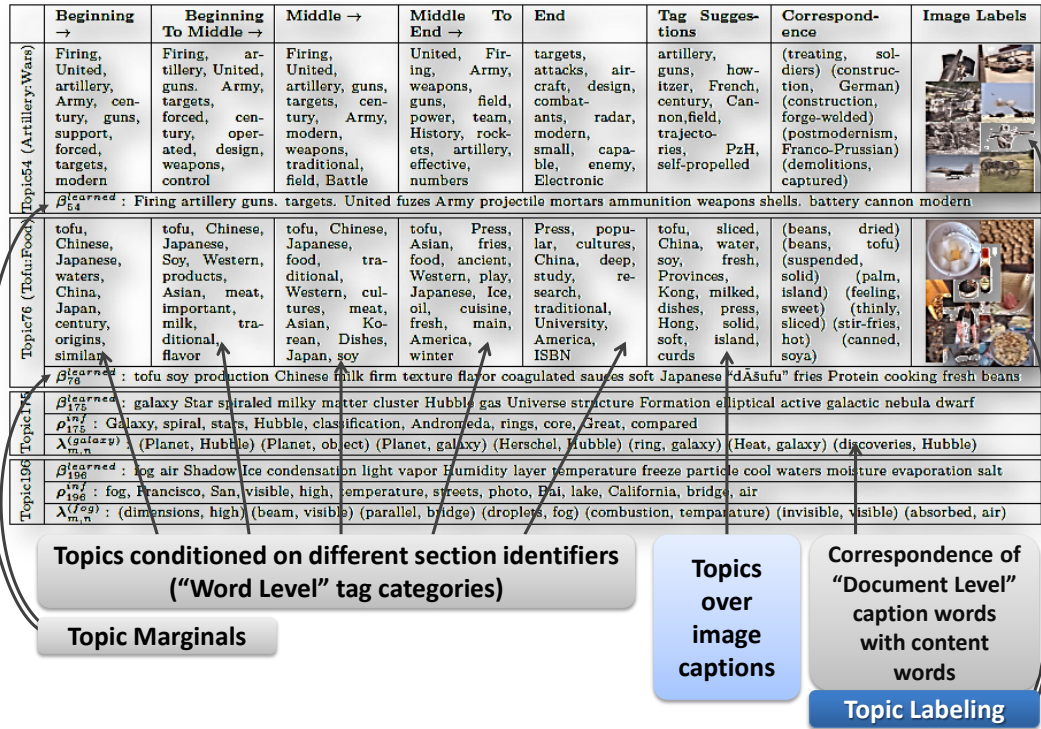


Figure 1.6: Faceted topics from Wikipedia (see Chapter 4)

The central problem of learning in this case is to find the *statistical estimators* for the mean parameters of the model that are consistent with both the current and future observations. By attacking that problem, we address the issue of document generalization which is the goal of probabilistic models like LDA [Blei et al., 2003]. It helps us answer the following questions: Given a new document, how similar it is to the previously seen documents? Where does it fit within them? What can one predict about it? Efficiently answering such questions is the main focus of any statistical analysis of large data collections.

We next write down the marginal probability of a corpus \mathcal{D} consisting of words $\mathbf{w}_{1:M}$ per document $d \in \mathcal{D}$ in the form of $p(\mathcal{D}|\Theta)$ for some of these mixed membership topic models. These models involve hidden state indicator variables corresponding to (usually) every observed variable $\mathbf{w}_m \in \{\mathbf{w}_{1:M}\}$ with Θ denoting the parameter set of the model and M denoting the number of words in document d . For judging predictive performance, we are interested in evaluating $p(\mathbf{w}|\mathcal{D}) = \int p(\mathbf{w}, \Theta|\mathcal{D})d\Theta = \int p(\mathbf{w}|\Theta) \times p(\Theta|\mathcal{D})d\Theta$ where \mathbf{w} is a unknown document never seen in the training set. The splitting over the new (test) and old (training) observations happen since the observations are conditionally independent of each other given the parameters of their common causal distributions [Shachter, 1998]—again a central assumption for the distribution over the observations for all of these models. We highlight at this point that in evaluating the expression $p(\mathbf{w}|\mathcal{D}) = \int p(\mathbf{w}, \Theta|\mathcal{D})d\Theta$, computing the posterior over *Theta* i.e. $\Theta|\mathcal{D})d\Theta$ plays a central role in all of predictive analysis involving unsupervised learning.

Latent Dirichlet Allocation (LDA): Figure 1.7a shows the basic model upon which we base our extensions which are explored along the course of this thesis. The initial development of this model happened around early 2000 and was published in 2003 [Blei et al., 2003]. The parameters of the model are α and $\beta_{1:K}$: the former represents a pseudo-count of how many observations do we expect to see in each

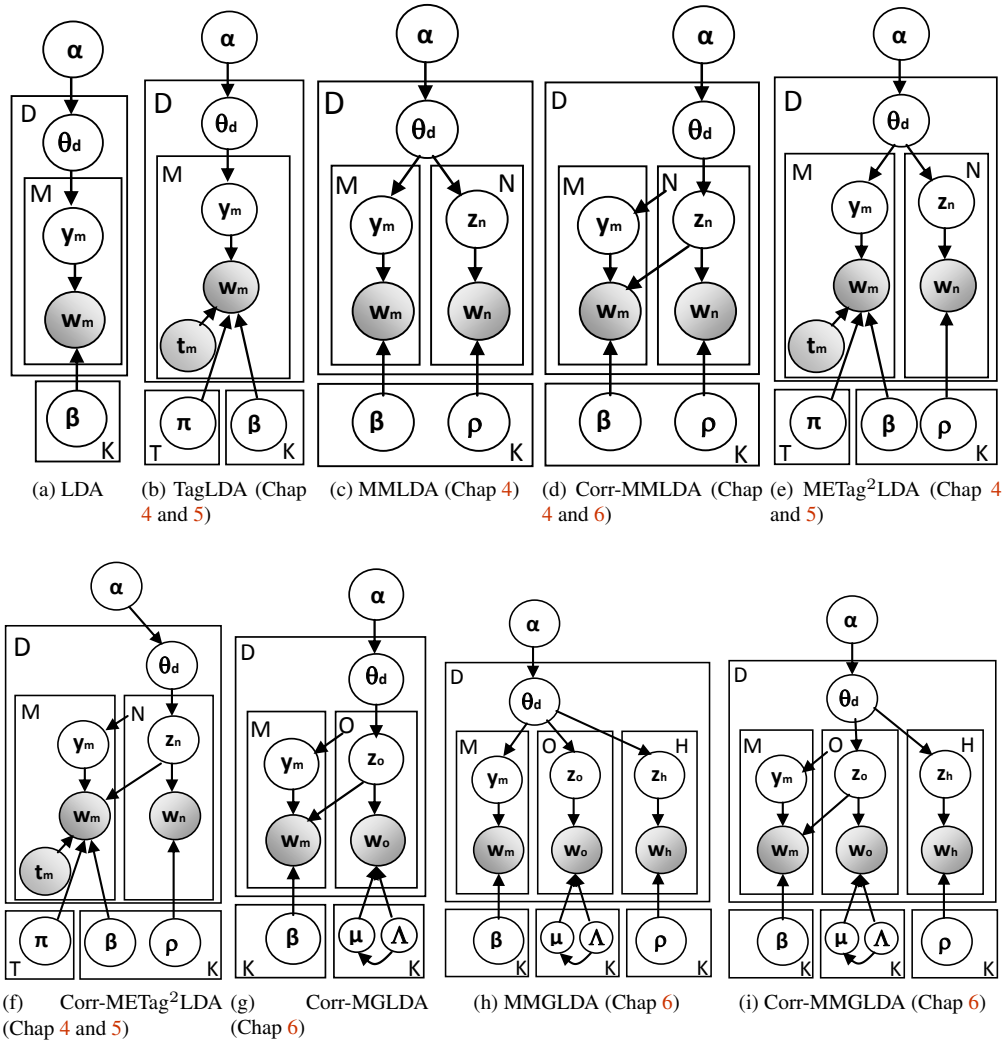


Figure 1.7: Some examples of topic models used in the various chapters of this thesis.

topic for each of the documents $d \in D$, D being the total number of all documents in a corpus \mathcal{D} . These pseudo-counts are in the absence of any observations. The K topics which are statistically interpreted as K independent multinomial distributions over words form the parameters, $\beta_{1:K}$, and are commonly referred to as the topics when visualized through samples from their modes. A corpus is a multiset of $D \times M_d$ words $\{w_{d,m}\}$ partitioned by $d \in \mathcal{D}$. The document level hidden variable θ representing topic proportions i.e. expected topic counts per document per topic is a matrix of positive reals of dimension $D \times K$ where each row is θ_d^T . The indicator variable $y_{d,m}$ answers “what is the topic for the current observation i.e. the word $w_{d,m}$?” The marginal probability of a corpus \mathcal{D} over the hidden variables is given by:

$$p(\mathcal{D}|\alpha, \beta_{1:K}) = \prod_{d=1}^D \int p(\theta_d|\alpha) \left(\prod_{m=1}^{M_d} \sum_{y_{d,m}=1}^K p(y_{d,m}|\theta) p(w_{d,m}|y_{d,m}, \beta_{1:K}) \right) d\theta_d \quad (1.1)$$

Regarding input, all that this model requires is simply the counts of the words in the corpus partitioned by each document i.e. $w_{d,m}$ and after processing this data of counts through some topic inference machinery (see Chapter 2), we can visualize the corpus through a set of topics as shown in Figure 1.2. Further computing probabilistic document similarity is also easily achieved through this machinery.

TagLDA: The TagLDA model [Zhu et al., 2006] shown in Figure 1.7b is an extension over the LDA model and was published as a technical report in 2006. The novelty of this model is the incorporation of word level annotations (dubbed tags in [Zhu et al., 2006]) so that the latent topics can be conditioned on such tags. Such a conditional distribution arises since the fundamental assumption is that the event space is not just document partitioned word ensembles but rather document partitioned (word, annotation/tag) ensembles. This means that there are two different sets of coupled distributions over the same set of words—one set representing latent topic distributions and the other set representing the tag distributions which can arise independent of any topical significance of the words. Each of these distributions are weighed inversely by its complementary distribution which intuitively means that the generative probability of a word depends on how much of it is explained by a topic and how much of it is explained by its associated *observed* annotation. The marginal probability of a corpus \mathcal{D} is given by:

$$p(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\pi}_{1:T}) = \prod_{d=1}^D \int p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \left(\prod_{m=1}^{M_d} \sum_{y_{d,m}=1}^K p(y_{d,m}|\boldsymbol{\theta}) p(w_{d,m}|y_{d,m}, t_{w_{d,m}}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\pi}_{1:T}) \right) d\boldsymbol{\theta}_d \quad (1.2)$$

In this model, $t_{w_{d,m}}$ is the observed annotation (or tag $t \in \{1, \dots, T\}$) associated with the word $w_{d,m}$ in the d^{th} document. Note that $\boldsymbol{\beta}_{1:K}$ and $\boldsymbol{\pi}_{1:T}$ are individually not multinomials.

It is interesting to note that if we partition a document with both word level annotations and document level metadata, then upto this point, LDA can model the document level metadata separately from TagLDA which can model the main document content with word level annotations. Due to the modular nature of the topic models, we can combine these two models into a more holistic model of documents containing both document level metadata and the main content annotated at the word level. This leads to an extension which we discuss briefly next.

In the context of nomenclature of the family of LDA models discussed in this thesis, the letter 'M' stands for discrete Multinomial distribution, the letter 'E' stands for distributions represented in an exponential form and the letter 'G' (forthcoming) represents Gaussian distributions over real-valued domains.

METag²LDA: Our new model (dubbed) METag²LDA (see Figure 1.7e) is a combination of both the LDA model and the TagLDA model and is described more thoroughly in Chapter 4. A similar model dubbed MMLDA in [Ramage et al., 2009b] (see Figure 1.7c) is also constructed using the same principles but cannot handle word level annotations. Its effectiveness for the multi-document summarization problem is investigated in Chapter 5. The marginal probability of a corpus \mathcal{D} under the METag²LDA model is given by:

$$p(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\pi}_{1:T}, \boldsymbol{\rho}_{1:K}) = \prod_{d=1}^D \int p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \left(\left[\prod_{m=1}^{M_d} \sum_{y_{d,m}=1}^K p(y_{d,m}|\boldsymbol{\theta}_d) p(w_{d,m}|y_{d,m}, t_{w_{d,m}}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\pi}_{1:T}) \right] \left[\prod_{n=1}^{N_d} \sum_{z_{d,n}=1}^K p(z_{d,n}|\boldsymbol{\theta}_d) p(w_{d,n}|z_{d,n}, \boldsymbol{\rho}_{1:K}) \right] \right) d\boldsymbol{\theta}_d \quad (1.3)$$

Here $\rho_{1:K}$ is another set of multinomial distributions which are the topics over the vocabulary of the document metadata. The observations for the metadata in document d are the variables named $w_{d,n}$ and the topic indicator for each such observation is $z_{d,n}$. Each of $y_{d,m}$ and $z_{d,n}$ has a 1-of- K representation denoting the number of possible number of topics as its range.

Corr-MMGLDA: The Correspondence-MMGLDA model combines the MMLDA model [Ramage et al., 2009b] (which is just a special case of the GM-LDA model in [Blei and Jordan, 2003]) and the Correspondence-LDA model [Blei and Jordan, 2003] to handle domain knowledge in the form of multimedia—in our case, videos with low level features in both the discrete (e.g. the variables $w_{d,h}$) and real (e.g. the variables $w_{d,o}$) valued domains. Figure 1.7i shows our proposed model which is explained more thoroughly in Chapter 6. It has been often observed that using an asymmetric prior α for the topic proportion variables improves on the topic quality by shifting the mass of more common observations to be aggregated on a few topics while leaving the other ones for discriminatory topic analysis [Wallach et al., 2009]. Although this is a welcome feature of the models, however, this also can easily lead to singularities in the precision matrices Λ_k governing the covariance of the real valued observations for some topics k . The marginal probability of a corpus \mathcal{D} for this model is given by:

$$p(\mathcal{D}|\alpha, \beta_{1:K}, \mu_{1:K}, \Lambda_{1:K}, \rho_{1:K}) = \left\{ \prod_{d=1}^D \int p(\theta_d|\alpha) \left(\prod_{m=1}^{M_d} \sum_{y_{d,m}=1}^{O_d} \sum_{z_{y_{d,m}}=1}^K p(y_{d,m}|O_d)p(w_{d,m}|z_{y_{d,m}}, \beta_{1:K}) \right) \right. \\ \left. \left[\prod_{h=1}^{H_d} \sum_{z_{d,h}=1}^K p(z_{d,h}|\theta_d)p(w_{d,h}|z_{d,h}, \rho_{1:K}) \right] \right. \\ \left. \left[\prod_{o=1}^{O_d} \sum_{z_{d,o}=1}^K p(z_{d,o}|\theta_d)p(w_{d,o}|z_{d,o}, \mu_{1:K}, \Lambda_{1:K}) \right] \right) d\theta_d \Bigg\} p(\mu, \Lambda) d(\mu, \Lambda) \quad (1.4)$$

where

$$p(\mu, \Lambda) = \prod_{k=1}^K p(\mu_k, \Lambda_k) = \prod_{k=1}^K p(\mu_k|\Lambda_k)p(\Lambda_k) = \prod_{k=1}^K \mathcal{N}(\mu_k|\mathbf{m}_0, (\kappa_0\Lambda_k)^{-1})\mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0) \quad (1.5)$$

is the prior over μ, λ .

The Corr-MMGLDA model (and a similar model dubbed MMGLDA) are explored more thoroughly in Chapter 6.

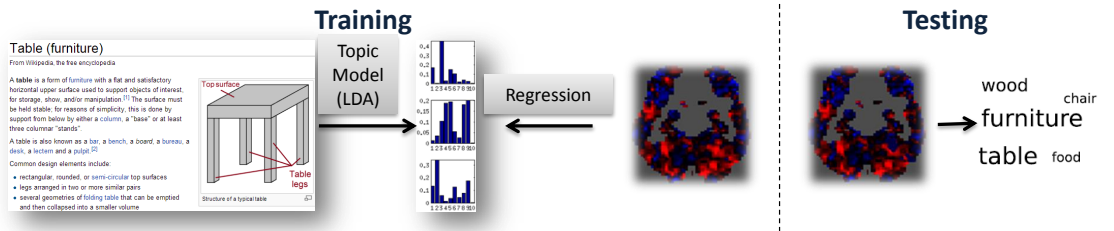


Figure 1.8: Mapping topics on Wikipedia articles on common visual objects to fMRI patterns. Unseen fMRI patterns are then used to predict words that appear to be semantically related through latent topics.

Recently, in the field of neurosciences, LDA has been used in conjunction with regression modeling to map fMRI patterns of the human brain scans to actual textual concepts. Figure 1.8 shows an example of such an experiment [Pereira et al., 2011]. A predetermined set of sixty simple concepts are chosen for

which there are both a Wikipedia entry as well as fMRI images available. The collection of Wikipedia articles are partitioned into sixty topics by running LDA. Regression has then been used to map the latent space of topics to the mean fMRI images. This way an unseen fMRI image can be efficiently mapped to a topic using the set of learnt regression weights. Although intuitive, the two models—LDA and regression act independently of each other and there is no direct correspondence between the pattern of fMRI data for the concepts to those in the text. The models shown in Figures 1.7i and 1.7h address the missing link and opens up further extensions for a supervised scenario as in [Blei and McAuliffe, 2007].



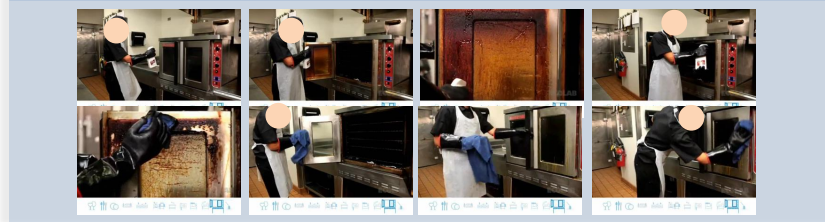
Figure 1.9: Changes in the BOLD (Blood Oxygen Level Dependent) patterns in a small area at the back of the visual cortex of the human brain (the region inside the green ellipses shown in the flattened brain scans) when the same subject is shown different movies. Reproduced here with permission from the authors in [Nishimoto et al., 2011]

A similar regression experiment was carried out in [Nishimoto et al., 2011] to map the responses of the Blood Oxygen Level Dependent (BOLD) signals in the early visual cortex area of the human brain to a set of movies. Figure 1.9 shows the changing patterns of the BOLD signals in response to changing scenes and nature of subjects and objects. These two examples clearly point out that there is some translation mechanism which translates visual patterns to text through some unknown target function which is usually captured as patterns in the BOLD or fMRI signals. Although a subject of conjecture, we believe that it can be safe to hypothesize that this mechanism of translation is different for different subjects and it is this difference which gives rise to different lingual descriptions of the same phenomenon.

Figure 1.10 shows some examples of practical applicability of the multi-modal topic models with regards to summarizing test videos into keyword summaries. Here we have shown the output of the MMLDA model (see Figure 1.7c) using only codebooks created through quantization of low-level action features constructed out of 3D gradients. Figure 1.10a shows two videos on the “Cleaning an appliance” event and the keywords predicted using words from some learned vocabulary constructed out of the textual summaries that corresponded with the videos in the training set. The keywords are tagged with semantic annotation using NLP tools as a post processing step. Figure 1.10b shows similar outputs on two videos on the “Working on a metal crafts project” event. Both of these examples represent hard examples for topic model based prediction as the scenes are cluttered with various objects which have high topical correlations. It is amazing at how well the human annotators capture the main event of the videos and express them in very short sentences (see “Human summary” lines in Figure 1.10). Such summaries can be looked upon as the generation of the most compact forms of information need for which the videos are most relevant. Generating such information needs can have wide spread applications in the searchability of videos for which there are little to no text available and also in a commercial setting to provide for relevant content for adword bidding.

In summary, we have thus far demonstrated the practical applicability of topic models which incorporate more domain knowledge—from probabilistic browsing of unstructured corpora of annotated

Cleaning an appliance



ClipID 448502 – machine/SUBJ-HUMAN man/SUBJ-HUMAN cleans/VERB clean/VERB cleaning/VERB water/OBJ microwave/OBJ espresso/NOUN refrigerator/OBJ shows/VERB person/SUBJ-HUMAN scrubbing/VERB wiping/VERB coffee/OBJ oven/OBJ

(Human Summary - a man cleans oven.)

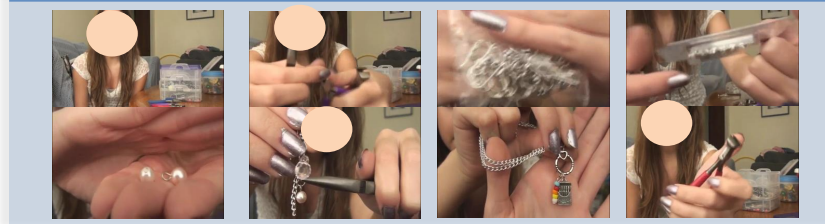


ClipID 243982 – cleaning/VERB cleans/VERB man/SUBJ-HUMAN items/NOUN food/OBJ refrigerator/OBJ woman/SUBJ-HUMAN guy/OBJ stove/OBJ brush/OBJ blender/OBJ microwave/OBJ evaporator/NOUN bucket/NOUN removing/VERB

(Human Summary - a guy dancing while cleaning kitchen appliances grill/stove top)

(a) Keyword predictions for two videos from the event: Cleaning an appliance

Working on a metal crafts project



ClipID 243220 – metal/OBJ man/SUBJ-HUMAN wire/NOUN shows/VERB make/VERB pliers/OTHER chain/OBJ earring/OBJ attaching/VERB meal/OTHER girl/SUBJ-HUMAN earrings/OBJ cutters/OBJ hooks/OBJ crimping/VERB

(Human Summary - a girl shows how to make earrings)



ClipID 641127 – metal/OBJ man/SUBJ-HUMAN rods/OBJ hammer/VERB copper/NOUN piece/OBJ making/VERB hot/ADJ hammering/VERB carving/VERB person/SUBJ-HUMAN silver/OBJ heater/OBJ soldering/VERB bending/VERB

(Human Summary – one guy making metal bracelet outdoors)

(b) Keyword predictions for two videos from the event: Working on a metal crafts project

Figure 1.10: Some examples of keyword predictions for test videos using MMLDA (Figure 1.7c) with action features

documents to generating captions in the form of lingual summaries of a video. In Chapter 2 we only *survey the preliminaries* and touch upon some important background materials used in this thesis.

1.2 Contributions of this thesis

Up to this point we have given an overview of what problems we are trying to address. We next discuss briefly the contributions of the following chapters in this thesis and the story they seam together from one end of the spectrum to the other.

1.2.1 Chapters 1 and 2

In this thesis, we have primarily focused on three objectives all of which summarizes contents of a data collection in different modalities in an unsupervised way. The first chapter starts by looking at the significance of the summarization problem in the context of the problems pursued in the later chapters. We then review, in the second chapter, some preliminaries that lay at the heart of the theory and practice on which our models stand. Towards the end of the second chapter we briefly discuss a previously state-of-the-art topic model—Latent Dirichlet Allocation (LDA) [Blei et al., 2003] and its implementation based on two different paradigms—one deterministic and which is followed throughout this thesis and another non-deterministic. We highlight some similarities within these two perspectives from the theoretical point of view and some dissimilarities from a parallelized implementation point of view. The original LDA model has been tremendously successful over previous models like pLSA [Hofmann, 1999] both for its effectiveness in finding topics and for its better generalization power. We extend LDA to overcome its basic limitation of its inability to incorporate multi-faceted domain knowledge present within a document itself and address these issues in the subsequent chapters.

1.2.2 Chapter 3

In Chapter 3, we describe a generative probabilistic topic model for text summarization that aims at extracting a small subset of sentences from the corpus with respect to some given query. Our initial hypothesis assumes that in addition to a bag of words, a document can also be viewed in a different manner. Words in a sentence always carry syntactic and semantic information and often such information (such as, the grammatical and semantic role (GSR) of a word like subject, object, noun and verb concepts etc.) is carried *across* adjacent sentences to enhance coherence in different parts of a document. However, alongside the principle statistical view of the main document as an *exchangeable* set of words, we also treat the corpora as a fixed set of exchangeable random variables each representing a sentence sampled from a discrete distribution over the GSR transitions. We then define a topic model which models documents by factoring in the GSR transitions for coherence and for a particular query, we rank sentences by a product of thematical salience and coherence through GSR transitions [Das and Srihari, 2009]. Although the model can directly select candidate summary sentences through the topic inference process, the principle shortcoming of this model is its fixed index of sentences which leads to generalization problems (same as in pLSA [Hofmann, 1999]) and the consideration of very coarse coherence triplets which disregard the surface form of the words across the sentences.

1.2.3 Chapter 4

The problem of lack of generalization in the models developed in Chapter 3 has been addressed in Chapters 4 and 5. There has been ongoing work in topic modeling of documents with lexical tags dubbed tag-topic models (see TagLDA [Zhu et al., 2006]) where words and part-of-speech tags typically reflect a single perspective, namely document content. To include this additional word level annotation per-

spective, we proposed new models [Das et al., 2011] which are primarily novel in: (i) the consideration of two different tag perspectives—a document level tag perspective that is relevant to the document as a whole and a word level annotation perspective pertaining to each word in the document; and (ii) the conditioning of latent topics with word level tag classes and labeling latent topics with images or videos in case of multimedia documents. These new models significantly improve on TagLDA depending on the variance of the signals in the document level perspective. We show the relevance of the models on multiple datasets including answering the question as to why the multimedia captions in a Wikipedia page may not only be relevant to the content in general but also the manually labeled categories.

1.2.4 Chapter 5

The field of automatic summary generation is increasingly gaining traction and there is a steady rise in demand of the summarization algorithms that is applicable to a wide variety of genres of text and other kinds of data as well (e.g. video). Producing summaries in a human readable form is very attractive particularly for very large datasets. However, for systems to achieve a level of human ingenuity on the task of summarization is very hard even for small sets of newswire documents. Experiments on human extractive summarization [Genest et al., 2009] show that even the best content-selection mechanism (e.g., a human summarizer) that is limited to pasting together sentences cannot achieve the same quality as fully manual summaries.

Recent extensions of LDA-based models that use more structure in the representation of documents have been proposed for generating more coherent and less redundant summaries, such as those in [Haghighi and Vanderwende, 2009]. These models use the collection and document-specific distributions of documents to be summarized in order to distinguish between the general and specific topics in documents. This amounts to identifying topic signature terms at multiple granularities in a corpus driven manner. Since many of these signature terms happen to be Named Entities (NE), it is often useful to use structured prediction methods, such as Conditional Random Fields, to identify them and influence the topic modeling process instead.

Our tag topic models developed in Chapter 4 has been found to be very useful in this respect. The finer aspects of the categories concerning {who, when, date, location} in the guided summarization task² naturally asks for highlighting the text with NE classes at the word level as well as using rhetorical parsing of texts [Marcu, 1999]. Our experiments in Chapter 5 show that important spans from Rhetorical Structure trees (RS-trees) together with sentence likelihood scores from models which are fit to the corpus as well as those from other local models of parts-of-speech importance, show state-of-the-art newswire multi-document summarization performance. We also obtain *bulleted list* summaries by using RS-tree construction [Das and Srihari, 2011, Das and Srihari, 2013].

1.2.5 Chapter 6

Finally in Chapter 6, by using topic models to summarize videos through text, we show that it is possible to outperform summaries generated through the labeling of videos using state-of-the-art object recognition [Das et al., 2013a, Das et al., 2013b] techniques from computer vision. For summarizing a video in terms of keywords and natural language generated henceforth, this approach removes expensive and manually laborious frame-by-frame bounding box annotation of videos required for training a large

²<http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html>

array of object detectors. The problem of summarizing videos through text is very important because summaries express information need which is paramount to any video search problem particularly when there is no accompanying text. The topic models that we use for this purpose are novel in handling both text and video features particularly when the video features present themselves in both discrete and real valued domain.

Predicted keyword summaries for videos with cluttered scenes is an extremely challenging task. For a scene like a kitchen or a metal crafts workshop, the abundance of topically relevant concepts causes the predicted keyword summary to lose relevancy to ground truth human summaries as measured by recall on unigram overlap even when more and more keywords are included. To combat this degradation of relevancy we select shorter sentences from the training set that are also topically relevant. Together with generated sentences using manual templates created out of the classes of concepts used for object detection, the final summaries significantly improve upon recall than just pulling in more keywords according to their prediction importance.

Chapter 2

Introductory Concepts

“Before turning to those moral and mental aspects of the matter which present the greatest difficulties, let the inquirer begin by mastering more elementary problems”

...

“I have already explained to you that what is out of the common is usually a guide rather than a hindrance. In solving a problem of this sort, the grand thing is to be able to reason backward. That is a very useful accomplishment, and a very easy one, but people do not practise it much. In the everyday affairs of life it is more useful to reason forward, and so the other comes to be neglected. There are fifty who can reason synthetically for one who can reason analytically. Let me see if I can make it clearer. Most people, if you describe a train of events to them, will tell you what the result would be. They can put those events together in their minds, and argue from them that something will come to pass. There are few people, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were which led up to that result. This power is what I mean when I talk of reasoning backward, or analytically.” – Sherlock Holmes: A Study in Scarlet, Chapters 2 and 7

In this Chapter, we begin with a preliminary introduction on exponential family distributions and gradually move towards general algorithms to obtain parameter estimates for models with distributions from the exponential family. We will also touch upon deterministic ways of finding approximate lower bounds to intractable integrals by exploiting pair-wise conjugate classes of exponential distributions. These integrals arise while calculating the likelihood functions. Further, in the light of some very recent work, a theoretical bound on the amount of data needed to learn the parameters of a topic model is mentioned in Section [2.7.2](#).

2.1 Exponential Family Distributions

The use of probability distributions which belong to the exponential family for random variables in graphical models has its foundation in the principles of maximum entropy [[Berger et al., 1996](#)]. Denote

$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ to be N independently and identically distributed i.e. i.i.d observations (for definition of i.i.d see Section 2.2.1) over which we compute empirical expectations of some set of functions $\hat{\boldsymbol{\mu}}_\alpha$ as follows:

$$\hat{\boldsymbol{\mu}}_\alpha = \frac{1}{N} \sum_{n=1}^N \Upsilon_\alpha(\mathbf{x}_n), \forall \alpha \in \mathbb{I} \quad (2.1)$$

where each α in some set \mathbb{I} indexes a function Υ_α of the sample \mathbf{X} only, with $\boldsymbol{\Upsilon} : \mathbf{X} \rightarrow \tilde{\mathbb{R}}$. The dimensionality of $\tilde{\mathbb{R}}$ depends on the dimensionality of \mathbf{X} and some r^{th} order moments corresponding to Υ_α . For example, in the case of scalar x , if we set $\Upsilon_1(x) = x$ and $\Upsilon_2(x) = x^2$, then these correspond to empirical versions of the first and second order moments of the random variable \mathbf{X} and thus $\tilde{\mathbb{R}} = \mathbb{R}^2$. For the same example, if \mathbf{X} was B dimensional, then $\tilde{\mathbb{R}} = \mathbb{R}^B \times \mathbb{R}^{B \times B}$.

Our goal is to infer a full probability distribution over the random variable \mathbf{X} , based on the $|\mathbb{I}|$ -dimensional vector of empirical expectations $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_\alpha, \alpha \in \mathbb{I})$ given some samples from the distribution over \mathbf{X} . This probability distribution as a density p can be continuous or discrete. In general, a probability distribution is symbolized with an uppercase alphabet and is defined as $P(x) = \int_{-\infty}^x p(x) dx$ in the continuous case and similarly in the discrete case where p is taken to be the probability density function (pdf) (or probability mass function (pmf)). We will use this distinction loosely in language and will often make of the lowercase symbols like p to refer to its distribution as well.

A given distribution P is consistent with the data if $E_{P_\theta}[\Upsilon_\alpha(\mathbf{X})] = \hat{\boldsymbol{\mu}}_\alpha, \forall \alpha \in \mathbb{I}$, i.e. the expectations $E_{P_\theta}[\Upsilon_\alpha(\mathbf{X})]$ under the distribution P are matched to the expectations under the empirical distribution. Here, the expression E_{P_θ} means expectation w.r.t. the probability distribution over \mathbf{X} i.e. $E_{p(\mathbf{X}|\theta)}$. Shannon's entropy [Cover and Thomas, 2006], expressed as $H(P) = - \int p(\mathbf{X}) \log p(\mathbf{X}) d\mathbf{X}$, is used to choose a P from among a family of \mathcal{P} 's which are consistent with the observations. The principle of *maximum entropy* is used to choose, from among the distributions consistent with the data, the distribution P^* whose Shannon entropy is maximal. Formally, letting \mathcal{P} be the set of all probability distributions over the random variable \mathbf{X} , the maximum entropy solution P^* is given by the solution to the constrained optimization problem:

$$P^* = \arg \max_{P \in \mathcal{P}} H(P), \quad \text{subject to } E_{P(\mathbf{X}|\theta)}[\Upsilon_\alpha(\mathbf{X})] = \hat{\boldsymbol{\mu}}_\alpha, \forall \alpha \in \mathbb{I} \quad (2.2)$$

This simply means that we are choosing the distribution with maximal uncertainty, as measured by its entropy, while remaining faithful to certain statistics of the observed samples. Assuming that problem 2.2 is feasible, it can be shown using calculus of variations that the optimal probability density solution p^* takes the form

$$p^*(\mathbf{X}|\theta) \propto \exp \left\{ \sum_{\alpha \in \mathbb{I}} \theta_\alpha \Upsilon_\alpha(\mathbf{X}) \right\} \quad (2.3)$$

where θ represents a parameterization of the distribution in an exponential family form. The constant of proportionality is dictated by the values of the Lagrange multipliers for the constraints which turns out to be the marginal distribution over θ i.e. integration of $p(\mathbf{X}, \theta) = \exp \left\{ \sum_{\alpha \in \mathbb{I}} \theta_\alpha \Upsilon_\alpha(\mathbf{X}) \right\}$ over \mathbf{X} . The functional form of the joint distribution over (\mathbf{X}, θ) in Equ. 2.3 belongs to the class of exponential family of probability distributions parameterized by θ and defined to be a set of distributions of the form:

$$p(\mathbf{X}|\theta) = h(\mathbf{X})g(\theta) \exp \left\{ \theta^T \boldsymbol{\Upsilon}(\mathbf{X}) \right\} \quad (2.4)$$

where \mathbf{X} may be a scalar or a vector and can be discrete or continuous. The random variables $\boldsymbol{\theta}$ are called the natural parameters of the distribution over the random variable \mathbf{X} . The realization of \mathbf{X} as a sample is \mathbf{x} and $\Upsilon(\mathbf{X})$ is some function of \mathbf{X} independent of $\boldsymbol{\theta}$. The function $g(\boldsymbol{\theta})$ ensures that the distribution is normalized and therefore satisfies

$$g(\boldsymbol{\theta})^{-1} = \int h(\mathbf{X}) \exp \left\{ \boldsymbol{\theta}^T \Upsilon(\mathbf{X}) \right\} d\mathbf{X} \quad (2.5)$$

where the integration can be replaced by a summation if \mathbf{X} is a discrete variable. Of course, $g(\boldsymbol{\theta}) > 0$ and must be bounded from above. Equations 2.4 and 2.5 are often written in the following alternate form:

$$p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N | \boldsymbol{\theta}) = \exp \left\{ \boldsymbol{\theta}^T \Upsilon(\mathbf{X}) - A(\boldsymbol{\theta}) \right\},$$

where $A(\boldsymbol{\theta}) = \ln \int \exp \left\{ \boldsymbol{\theta}^T \Upsilon(\mathbf{X}) \right\} d\mathbf{X}$ (2.6)

Here $\Upsilon = \{\Upsilon_\alpha, \alpha \in \mathbb{I}\}$ is a collection of functions $\{\Upsilon_\alpha\}$, known either as potential functions or sufficient statistics (see Section 2.2.2). The index set is denoted by \mathbb{I} with $d = |\mathbb{I}|$ components to be specified, so that Υ can be viewed as a vector-valued mapping from some sample space \mathcal{X} to some real valued set $\tilde{\mathbb{R}}$. For a given vector of sufficient statistics Υ , we denote $\boldsymbol{\theta} = \{\theta_\alpha, \alpha \in \mathbb{I}\}$ to be an associated vector of canonical or exponential parameters.

The parameters typically dealt with in practical problems are just points in (usually) a very high dimensional space which is at least as large as the dimensionality of the observations. Finding this point given the set of samples at hand thus reduces to searching a set of size $2^{\aleph_0} = \aleph_1$ where the set is that of the non-denumerable real numbers. This search though is intuitively only restricted to the neighborhoods of the subsets of the total yet unseen input sample space \mathcal{X} .

The quantity $A(\boldsymbol{\theta})$ in Equ. 2.6 or $-\log g(\boldsymbol{\theta})$ in Equ. 2.4 is called the log partition function or the cumulant function which makes sure that $p(\mathbf{x}|\boldsymbol{\theta})$ in Equ. 2.6 is properly normalized in accordance with probability measure theory [Feller, 1968]. With the set of potentials Υ fixed, each parameter vector $\boldsymbol{\theta}$ indexes a particular member $P_\boldsymbol{\theta}$ of the family \mathcal{P} . The canonical parameters $\boldsymbol{\theta}$ of interest belong to the set $\Omega = \{\boldsymbol{\theta} \in \tilde{\mathbb{R}} \mid A(\boldsymbol{\theta}) < +\infty\}$. It is shown in [Wainwright and Jordan, 2008] that for exponential family models, A is a convex function of $\boldsymbol{\theta}$, which in turn implies that Ω must be a convex set. For examples on a wide of range of popular models with parameters in the exponential family see [Wainwright and Jordan, 2008].

For an example concerning discrete distributions, let us consider a discrete distribution for a single observation \mathbf{x} which takes the form

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k} = \exp \left\{ \sum_{k=1}^K x_k \ln \mu_k \right\} \quad (2.7)$$

where $\mathbf{x} = (x_1, \dots, x_K)^T$ is the K dimensional binary observation vector with only one of the x_k s being one and the rest zeros. We can write this as $p(\mathbf{x}|\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^T \mathbf{x})$ where $\theta_k = \ln \mu_k$. Comparing with Equ. 2.4 we have:

$$\Upsilon(\mathbf{x}) = \Upsilon_1(\mathbf{x}) = \mathbf{x}; \quad h(\mathbf{x}) = 1; \quad g(\boldsymbol{\theta}) = 1 \quad (2.8)$$

The parameters θ_k are not independent because the parameters μ_k are subject to the constraint $\sum_{k=1}^K \mu_k = 1$ and also $0 \leq \mu_k \leq 1$. We now write an alternate version of $p(\mathbf{x}|\boldsymbol{\mu})$ as $p(\mathbf{x}|\boldsymbol{\theta})$ where

we remove the simplex constraint of $\boldsymbol{\mu}$ by expressing $\boldsymbol{\theta}$ in terms of only the $K - 1$ independent parameters of $\boldsymbol{\mu}$.

$$\begin{aligned} \exp \left\{ \sum_{k=1}^K x_k \ln \mu_k \right\} &= \exp \left\{ \sum_{k=1}^{K-1} x_k \ln \mu_k + \left(1 - \sum_{k=1}^{K-1} x_k \right) \ln \left(1 - \sum_{k=1}^{K-1} \mu_k \right) \right\} \\ &= \exp \left\{ \sum_{k=1}^{K-1} x_k \ln \left(\frac{\mu_k}{1 - \sum_{j=1}^{K-1} \mu_j} \right) + \ln \left(1 - \sum_{k=1}^{K-1} \mu_k \right) \right\} \end{aligned} \quad (2.9)$$

Let us denote $\ln \left(\frac{\mu_k}{1 - \sum_{j=1}^{K-1} \mu_j} \right) = \theta_k$ where $\sum_{j=1}^{K-1} \exp \theta_j = 1$. To express μ_k in terms of θ_k we do the following:

$$\begin{aligned} \mu_K + \sum_{j=1}^{K-1} \mu_j - \mu_K &= \left(\sum_{j=1}^{K-1} \exp \theta_j \right) \mu_K \\ \Rightarrow 1 - \mu_K &= \left(\sum_{j=1}^{K-1} \exp \theta_j \right) \mu_K \\ \Rightarrow \mu_K &= \frac{1}{1 + \left(\sum_{j=1}^{K-1} \exp \theta_j \right)} \Rightarrow \mu_k = \frac{\exp \theta_k}{1 + \left(\sum_{j=1}^{K-1} \exp \theta_j \right)} \end{aligned} \quad (2.10)$$

This is called the softmax function which is very commonly used in many supervised learning scenarios where the goal is to transform a real valued output into a probability space. The discrete distribution takes the form:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \left(1 + \sum_{j=1}^{K-1} \exp \theta_j \right)^{-1} \exp(\boldsymbol{\theta}^T \mathbf{x}) \quad (2.11)$$

with

$$\Upsilon_1(\mathbf{x}) = \mathbf{x}, \quad h(\mathbf{x}) = 1 \quad \text{and} \quad g(\boldsymbol{\theta}) = \left(1 + \sum_{j=1}^{K-1} \exp \theta_j \right)^{-1} \quad (2.12)$$

The expression in Equ. 2.4 also holds for any application of a smooth function \mathbf{w} over $\boldsymbol{\theta}$ and can thus be expressed in a more generic form as:

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta}) \exp \left(\sum_{i=1}^K w_i(\boldsymbol{\theta}) \Upsilon_i(\mathbf{x}) \right). \quad (2.13)$$

Here $h(\mathbf{x}) > 0$ and $\Upsilon_i(\mathbf{x})$ are real-valued functions of the observation \mathbf{x} , $g(\boldsymbol{\theta}) \geq 0$ and $w_1(\boldsymbol{\theta}), w_2(\boldsymbol{\theta}), \dots, w_K(\boldsymbol{\theta})$ are real valued functions of the parameter vector $\boldsymbol{\theta}$ (they cannot depend on \mathbf{x}). The set $\{\boldsymbol{\theta}\} \in \Theta$ is the natural parameter space for the family of exponential distributions and $\{w_i(\boldsymbol{\theta})\}$ is a subset of that space. Many common probability distributions belong to the exponential family—continuous distributions such as normal, gamma, beta as well as discrete distributions such as Poisson, binomial, negative binomial, multinomial etc. Further, for distributions belonging to the exponential families, we have the following theorem from [Casella and Berger, 2001] which implies that the log partition function $g(\boldsymbol{\theta})$ is smooth and convex in terms of $\boldsymbol{\theta}$.

Theorem 2.1.1. If \mathbf{X} is a random variable with pdf or pmf of the form $p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right)$ which is a more generic version of Equ. 2.4, then

$$E\left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{X})\right) = -\frac{\partial}{\partial \theta_j} \log g(\boldsymbol{\theta}) \quad (2.14)$$

$$\text{Var}\left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{X})\right) = -\frac{\partial^2}{\partial \theta_j^2} \log g(\boldsymbol{\theta}) - E\left(\sum_{i=1}^K \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} \boldsymbol{\Upsilon}_i(\mathbf{X})\right) \quad (2.15)$$

Proof.

$$\begin{aligned} 0 &= \frac{\partial}{\partial \boldsymbol{\theta}} \int h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} \\ &= \int h(\mathbf{x})g'(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} + \int h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) \left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} \\ &= \int h(\mathbf{x}) \left[\frac{\partial \log g(\boldsymbol{\theta})}{\partial \theta_j}\right] g(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} \\ &\quad + \int h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) \left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} \\ &= \frac{\partial \log g(\boldsymbol{\theta})}{\partial \theta_j} + E\left[\left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right)\right] \end{aligned} \quad (2.16)$$

Hence Equ. 2.14 holds

$$\begin{aligned} 0 &= \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \int h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} \\ &= \int h(\mathbf{x})g''(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} + \int h(\mathbf{x})g'(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) \left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} \\ &\quad + \int h(\mathbf{x})g'(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) \left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} \\ &\quad + \int h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) \left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right)^2 d\mathbf{x} \\ &\quad + \int h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) \left(\sum_{i=1}^K \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} \boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} \\ &= \int h(\mathbf{x}) \left[\frac{\partial^2}{\partial \theta_j^2} \log g(\boldsymbol{\theta})\right] \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} + \int h(\mathbf{x}) \left[\frac{g'(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\right]^2 g(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^K w_i(\boldsymbol{\theta})\boldsymbol{\Upsilon}_i(\mathbf{x})\right) d\mathbf{x} \\ &\quad + 2\left(\frac{\partial}{\partial \theta_j} \log g(\boldsymbol{\theta})\right) E\left[\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right] + E\left[\left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right)^2\right] + E\left[\left(\sum_{i=1}^K \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} \boldsymbol{\Upsilon}_i(\mathbf{x})\right)\right] \\ &= \frac{\partial^2}{\partial \theta_j^2} \log \boldsymbol{\theta} + \left[\frac{\partial}{\partial \theta_j} \log \boldsymbol{\theta}\right]^2 - 2E\left[\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right] E\left[\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \boldsymbol{\Upsilon}_i(\mathbf{x})\right] \end{aligned}$$

$$\begin{aligned}
& + E \left[\left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \Upsilon_i(\mathbf{x}) \right)^2 \right] + E \left[\left(\sum_{i=1}^K \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} \Upsilon_i(\mathbf{x}) \right) \right] \\
& = \frac{\partial^2}{\partial \theta_j^2} \log \boldsymbol{\theta} + \text{Var} \left(\sum_{i=1}^K \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \Upsilon_i(\mathbf{x}) \right) + E \left[\left(\sum_{i=1}^K \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} \Upsilon_i(\mathbf{x}) \right) \right] \tag{2.17}
\end{aligned}$$

Hence Equ. 2.15 holds.

Note: If we consider a sample to consist of N i.i.d samples \mathbf{x}_n , then $\Upsilon_i(\mathbf{x})$ in the above theorem is replaced by $\sum_{n=1}^N \Upsilon_i(\mathbf{x}_n)$ due to the multiplication of independent exponential terms. \square

In general, for an exponential family, the set $\{\mathbf{x}\}$ of values for which $p(\mathbf{x}|\boldsymbol{\theta}) > 0$ cannot depend on $\boldsymbol{\theta}$. This is easily noted by observing the limits of the integration over the pdf to compute the normalization constant. This constraint is included into the definition of the exponential family distribution by an indicator function. The indicator function of a set A , most often denoted by $I_A(x)$ or $\delta(x, A)$ is the function:

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{2.18}$$

This indicator function can be incorporated into the definition of $h(\mathbf{x})$ in Equ. 2.13 or Equ. 2.4. Since exponential is always positive, it can be observed that for any $\boldsymbol{\theta} \in \Theta$ and $g(\boldsymbol{\theta}) > 0$, we have $\{\mathbf{x} : p(\mathbf{x}|\boldsymbol{\theta}) > 0\} = \{\mathbf{x} : h(\mathbf{x}) > 0\}$ and this set does not depend on $\boldsymbol{\theta}$. So a pdf like $p(x|\theta) = \theta^{-1} \exp(1 - (x/\theta))$, $0 < \theta < x < \infty$ is not in the exponential family since the indicator function here is expressed as $I_{[\theta, \infty]}(x)$.

The METag²LDA (see Fig. 1.7e) model can be represented using a 6-tuple representation for the hidden and observed variables, $(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}, \mathbf{w}_M, \mathbf{t}_T, \mathbf{w}_N)$, as follows:

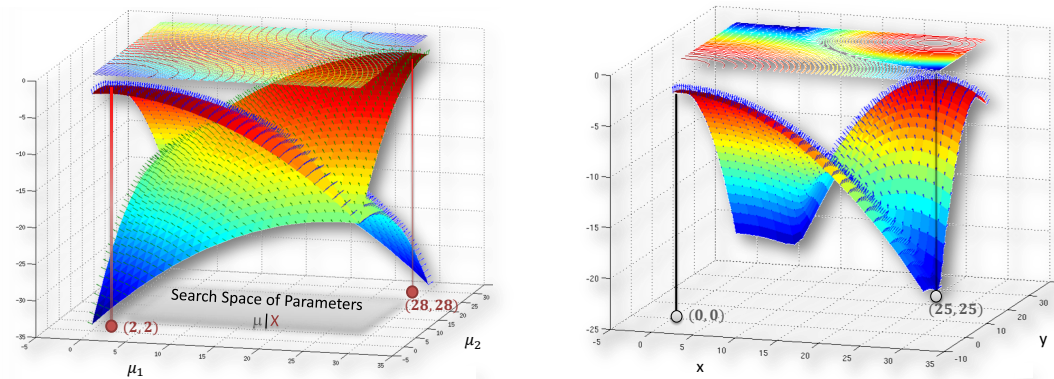
$$\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{\alpha}) p(\mathbf{y}|\boldsymbol{\theta}) p(\mathbf{z}|\boldsymbol{\theta}) p(\mathbf{w}_M|\mathbf{y}_M, \mathbf{t}_T, \boldsymbol{\beta}, \boldsymbol{\pi}) p(\mathbf{w}_N|\mathbf{z}_N, \boldsymbol{\rho}) \propto \exp \left(\sum_{k=1}^K \alpha_k \log \theta_k + \sum_{k=1}^K \log \theta_k [I_{\{k\}}(\mathbf{y}) + I_{\{k\}}(\mathbf{z})] \right. \\
\left. + \sum_{k=1}^K \sum_{t=1}^T \sum_{j=1}^V \log \beta_{k,j} \log \pi_{t,j} I_{\{k\}}(\mathbf{y}) I_{\{(j,t)\}}(\mathbf{w}_M) + \sum_{k=1}^K \sum_{j=1}^{corrV} \log \rho_{k,j} I_{\{k\}}(\mathbf{z}) I_{\{(j)\}}(\mathbf{w}_N) \right) \tag{2.19}
\end{aligned}$$

The model thus belongs to the exponential family with parameter vector $\Theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\rho})$, with an associated density of the form shown in Equ. 2.19.

The way the parameters are inferred from the data depends upon the choice of inference algorithms (see Sections 2.6 for a brief overview of gradient based approximation algorithms and Section 2.7.6 for a brief overview of sampling based algorithms). It is common knowledge that the performance of gradient based iterative algorithms may depend crucially on how the problem is formulated. Proper attention needs to be put on the scaling of the variables being optimized. In unconstrained optimization, a function $f(\boldsymbol{\theta})$ is said to be poorly scaled if changes to the variable $\boldsymbol{\theta}$ in a certain direction produce much larger variations in the value of f than do changes to $\boldsymbol{\theta}$ in another direction. This causes a problem for gradient descent (or ascent) type of algorithms which are very sensitive to scaling in the absence of a Newton step involving the Hessian [Nocedal and Wright, 2006] or even fixed point iterations where the slope of $f(\boldsymbol{\theta})$ is too large in absolute value near the fixed point of $f(\boldsymbol{\theta})$ [Conte and Boor, 1980]. The problem is alleviated by the use of fixed regularizers or (conjugate) priors over the variables being optimized.

2.2 Maximum Likelihood, Sufficient Statistics and Conjugate Priors

In this section we discuss the problem of estimating parameters of a model. The most common form of estimation is point estimation where a parameter $\theta \in \Theta$ is treated as a point in some high dimensional space and we need to find this point based on the observations at hand so that θ becomes a representative of the sample space with high probability i.e. we want our model to find the true encoding of the knowledge about a population given some samples from it. We estimate the *value* of this encoding using some *estimator* and this is captured in the parameters of a model which we are using to explain the occurrence of the samples. For example, the sample mean is an unbiased estimator of the population mean and the value of the estimate improves as the size of the sample grows larger. When sampling is from a population described by a pdf (probability density function) or pmf (probability mass function) $p(\mathbf{x}|\theta)$, knowledge of θ yields knowledge of the entire population.



(a) Graph of a two Gaussian-mixture model ELBO over μ given \mathbf{x} (b) Graph of a two Gaussian-mixture model ELBO given true μ over samples from the input space \mathcal{X}

Figure 2.1: Graphs of the objective function of a two component Gaussian-mixture model over a set of mean parameters and over a set of sample points. The surface normals are shown as tiny blue arrows

Let us illustrate what we just said using a concrete example of a two component Gaussian mixture model. Figure 2.1 shows two graphs of a function of real-valued data and real-valued parameters, called the likelihood function, which is defined in Equ. 2.20.

$$\mathcal{L}(\mathbf{X}, \mu_1, \mu_2, \sigma_1, \sigma_2) = \log \left(0.8 \times \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2} \|\mathbf{x} - \mu_1\|_2^2} + 0.2 \times \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{1}{2\sigma_2^2} \|\mathbf{x} - \mu_2\|_2^2} \right) \quad (2.20)$$

In this visualization, we hold four parameters fixed—the mixing proportions with values of 0.8 and 0.2 and the variances or the scaling parameters σ_1 and σ_2 . Figure 2.1a shows the graph of the likelihood function as two sample points (2, 2) and (28, 28) are held fixed and we explore the space of location parameters μ_1 and μ_2 plotting the value of $\mathcal{L}(\mathbf{X}, \mu_1, \mu_2, \sigma_1, \sigma_2)$ at each possible pair. Clearly the likelihood values are very high around the two sample points but taper off as we move away. Intuitively, the location parameters which we are searching for are really sample points but we do not know which ones. This means that the likelihood function merely serves to explain how close is a sample point to *the* sample point which is representative of *all other* sample points surrounding it. The latter point, in case of real valued observations, is a mean parameter of the distribution of all such samples. There can be

more than one mean parameters in the case of a multi-modal or mixture distribution.

Alternatively, fixing the parameters of the mixture model to be known, we plot the graph of the likelihood function from Equ. 2.20 but varying only the sample points. Figure 2.1b shows the graph of the two component mixture density and contours on top show the regions of high (red) to low (blue) probabilities. From this perspective, we are trying to optimize the likelihood function in Equ. 2.20 as a function the sample points alone but w.r.t the parameters. The parameters can be loosely thought of *true representative sample points* in case of real-valued data, but not so for discrete samples. Mean parameters for discrete samples have a more geometric interpretation in terms of polytopes where each corner of the polytope is a possible configuration of a sample point.

Given these two perspectives, the importance of a *training dataset* becomes very evident. Section 2.3 briefly discusses the impact of the in-sample error for a classifier on its generalization performance when subject to classification of out-of-sample points. We now introduce some basic definitions which are used throughout this chapter.

A **point estimator** is *any function* $\Upsilon(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ of a sample $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$. In statistics literature, a **statistic** is also *any function* of the data which is *not* a function of the parameter. Formally, if $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be a random sample of size N from a population and $T(\mathbf{x}_1, \dots, \mathbf{x}_N)$ be a function whose domain is the sample space \mathcal{X} of $(\mathbf{X}_1, \dots, \mathbf{X}_N)$, then the random variable $\Upsilon = T(\mathbf{X}_1, \dots, \mathbf{X}_N)$ is called a statistic. The probability distribution of a statistic Υ is called the sampling distribution of Υ . Any statistic is thus a point estimator. This broad definition does not mention the range of the statistic $\Upsilon(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ coinciding with that of the parameter θ . Usually this is the case but not always [Berger, 1985]. Thus an *estimator* is a **function** of the sample, while an *estimate* is the **realized value** of the estimator.

There are several methods of finding estimators: i) Method of moments ii) Maximum Likelihood iii) Bayes Estimators and iv) Expectation Maximization Algorithm. We touch upon the last three briefly as they are used in the course of this thesis for some model or the other. The method of moments is less frequently used and is found by equating the first k sample moments to the corresponding k population moments and solving the resulting system of simultaneous equations.

2.2.1 Maximum Likelihood

The method of Maximum Likelihood is one of the most popular methods for deriving estimators. If $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ are i.i.d sample from a population with pdf or pmf $p(\mathbf{x}|\theta)$, the likelihood function is defined by:

$$\mathcal{L}(\mathbf{X}|\theta) = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N|\theta_1, \theta_2, \dots, \theta_K) = \prod_{n=1}^N p(\mathbf{x}_n|\theta_1, \theta_2, \dots, \theta_K) \quad (2.21)$$

The likelihood function may not necessarily be a probability density function (pdf) or probability mass function (pmf) although p is a valid pdf or pmf.

Definition 2.2.1. *Two random variables \mathbf{X}_1 and \mathbf{X}_2 are **identically distributed** iff for every set $\mathcal{Q} \in \mathcal{B}$, where \mathcal{B} is the Sigma algebra corresponding to the sample space S , $p(\mathbf{X}_1 \in \mathcal{Q}) = p(\mathbf{X}_2 \in \mathcal{Q})$.*

Definition 2.2.2. *The random variables $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ are called a random sample of length N from a population $p(\mathbf{X})$ if $\{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ are mutually independent and the marginal pdf or pmf of each \mathbf{X}_i is the same $p(\mathbf{X})$. This set of random variables is then referred to as a set of **independent and identically***

distributed random variables. Mutual independence of random variables mean that the outcome of one random variable has no relationship with or not dependent on the outcome of another random variable.

Consider a sample corresponding to the random variable \mathbf{X} and denote $\theta^*(\mathbf{X})$ to be a parameter value at which $\mathcal{L}(\mathbf{X}|\theta)$ attains its maximum as a function of θ with \mathbf{X} held fixed. This $\theta^*(\mathbf{X})$ is a Maximum Likelihood Estimator (MLE) of the parameter θ based on a sample \mathbf{X} . The ML estimate is the parameter point θ^* for which the observed sample \mathbf{X} is most likely. The likelihood function is an important **statistic** that is used to *summarize* the data in a statistical sense [Hastie et al., 2009].

If the likelihood function is differentiable (in θ_i), the possible candidates for the MLE are the values of $\theta_1, \theta_2, \dots, \theta_K$ that solve

$$\frac{\partial}{\partial \theta_i} \mathcal{L}(\mathbf{X}|\theta) = 0, \quad i = 1, \dots, k \quad (2.22)$$

The solutions to Equ. 2.22 are only *possible candidates* for the MLE since the first derivative being 0 is only a necessary condition for a maximum not a sufficient one. Furthermore, the zeros of the first derivative locate only extreme points in the interior of the domain of the function. If the extrema occur on the boundary of the parameter space, the first derivative may not be zero and the condition of $\frac{\partial^2 \mathcal{L}(\mathbf{X}|\theta)}{\partial^2 \theta} < 0$ will not hold. This means that no estimate of such a parameter can be found through MLE.

Maximum Likelihood and Maximum Entropy are convex duals i.e. the set \mathcal{P} of distributions which has maximum entropy but are consistent with the data and the set \mathcal{Q} of distributions from the exponential family (often referred to as Gibbs distributions) contains only a single point which solves both problems. This means that the following are equivalent:

- $p^* = \arg \max_{p \in \mathcal{P}} H(p)$ which solves Maximum Entropy and
- $p^* = \arg \max_{p \in \mathcal{Q}} \sum_n \log p(\mathbf{x}_n)$ which solves Maximum Likelihood have a single point in common i.e.
- $p^* \in \mathcal{P} \cap \mathcal{Q}$ and anyone of these properties uniquely defines p^* .

The outline of the proof is as follows:

Proof. The objective function of Maximum Entropy, if we consider discrete distributions and first order constraints, is

$$\mathcal{L} = \arg \max_{p \in \mathcal{P}} - \sum_{\mathbf{X}} p(\mathbf{X}) \log p(\mathbf{X}) \quad (2.23)$$

subject to $\sum_{\mathbf{X} \in \mathcal{X}} p(\mathbf{X}) = 1$ and $\hat{E}[f_j(\mathbf{X})] = \sum_{\mathcal{X}} f_j(\mathbf{X}) p(\mathbf{X})$ where $\hat{E}[f_j(\mathbf{X})]$ is the empirical expectation given only the samples $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} = \mathbf{X}$. To find the optimal p^* , we minimize $\sum_{\mathcal{X}} p(\mathbf{X}) \log p(\mathbf{X})$ subject to the same constraints. This yields the following steps for the necessary conditions to be satisfied:

$$\mathcal{L} = \arg \min_{p \in \mathcal{P}} \sum_{\mathbf{X} \in \mathcal{X}} p(\mathbf{X}) \log p(\mathbf{X}) + \sum_j \lambda_j \left[\hat{E}[f_j(\mathbf{X})] - \sum_{\mathbf{X} \in \mathcal{X}} f_j(\mathbf{X}) p(\mathbf{X}) \right] + \gamma \left(\sum_{\mathbf{X} \in \mathcal{X}} p(\mathbf{X}) - 1 \right) \quad (2.24)$$

where the λ_j s and γ are Lagrange multipliers.

$$\frac{\partial \mathcal{L}}{\partial p(\mathbf{X})} = \log p(\mathbf{X}) + 1 + \sum_j \lambda_j [-f_j(\mathbf{X})] + \gamma$$

$$\Rightarrow p^*(\mathbf{X}) = \frac{\exp\left\{\sum_j \lambda_j f_j(\mathbf{X})\right\}}{\mathcal{Z}} \quad (2.25)$$

where $\mathcal{Z} = \exp(1 + \gamma) = \sum_{\mathbf{X} \in \mathcal{X}} \exp\left\{\sum_j \lambda_j f_j(\mathbf{X})\right\}$. Plugging the value of $p^*(\mathbf{X})$ into \mathcal{L} in Equ. 2.24, we have:

$$\begin{aligned} \mathcal{L} &= \sum_{\mathbf{X} \in \mathcal{X}} p^*(\mathbf{X}) \left[\sum_j \lambda_j f_j(\mathbf{X}) - \log \mathcal{Z} \right] + \sum_j \lambda_j \left[\hat{E}[f_j(\mathbf{X})] - \sum_{\mathbf{X} \in \mathcal{X}} f_j(\mathbf{X}) p^*(\mathbf{X}) \right] + \gamma \left(\sum_{\mathbf{X} \in \mathcal{X}} p^*(\mathbf{X}) - 1 \right) \\ &= \sum_j \lambda_j \left[\hat{E}[f_j(\mathbf{X})] \right] - \log \mathcal{Z} + (\gamma \times 0) \\ &= \frac{1}{N} \sum_{n=1}^N \underbrace{\left[\underbrace{\sum_j \lambda_j f_j(\mathbf{x}_n)}_{\log p^*(\mathbf{x}_n)} \right] - \log \mathcal{Z}}_{\log \prod_{n=1}^N p^*(\mathbf{x}_n)} \end{aligned} \quad (2.26)$$

Thus the Maximum Entropy solution $p^*(\mathbf{X})$ solves the Maximum Likelihood problem as well. Incidentally p^* belongs to Gaussian distribution if we include second order constraints into the objective of Maximum Entropy. \square

2.2.2 Sufficient Statistics

Any statistic $\Upsilon(\mathbf{X})$, which is just a function of the data, represents a way to summarize the data i.e. defines a form of data reduction. This form of data reduction through sufficient statistics can be thought of a partition of the sample space \mathcal{X} . If the set $\mathcal{U} = \{\mathbf{u} : \Upsilon(\mathbf{X}) = \mathbf{u} \text{ for some } \mathbf{X} \in \mathcal{X}\}$ be the image of \mathcal{X} under $\Upsilon(\mathbf{X})$, then $\Upsilon(\mathbf{X})$ partitions the sample space into sets $\{\mathcal{S}_{\mathbf{u}}\}$, $\mathbf{u} \in \mathcal{U}$ defined by $\mathcal{S}_{\mathbf{u}} = \{\mathbf{X} : \Upsilon(\mathbf{X}) = \mathbf{u}\}$. For example, if $\Upsilon(\mathbf{X}) = \mathbf{u}$ represents the sum of a sample $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, then $\mathcal{S}_{\mathbf{u}}$ is the set of all data points $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ that amounts to the *same* sum \mathbf{u} . A statistic is said to be sufficient for a parameter θ if it obeys the ‘‘Sufficiency principle’’ [Casella and Berger, 2001].

Sufficiency principle: If $\Upsilon(\mathbf{X})$ is a sufficient statistics for a parameter θ , then any inference about θ should depend on the sample \mathbf{X} through the value $\Upsilon(\mathbf{X})$. Thus if \mathbf{x} and \mathbf{y} are two sample values such that $\Upsilon(\mathbf{x}) = \Upsilon(\mathbf{y})$ then inference about θ should be the same upon the observation of either \mathbf{x} or \mathbf{y} .

Formally, a statistic $\Upsilon(\mathbf{X})$ is said to be a **sufficient statistic** for θ if the conditional distribution of the sample \mathbf{X} given the value of $\Upsilon(\mathbf{X})$ does not depend on θ . The implication of this is that if $p(\mathbf{X}|\theta)$ is the joint pdf or pmf of \mathbf{X} and $q(\mathbf{U}|\theta)$ is the pdf or pmf of $\Upsilon(\mathbf{X})$, then $\Upsilon(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{X} \in \mathcal{X}$, the ratio $p(\mathbf{X}|\theta)/q(\mathbf{U}|\theta)$ is a constant i.e. does not depend on θ . For example if $\mathbf{X} = \{x_1, \dots, x_N\}$ are N i.i.d. Bernoulli random variables with parameter θ , then $\Upsilon(\mathbf{X}) = \sum_{n=1}^N x_n = u$, which just counts the number of 1s, is a sufficient statistic for θ because $p(\mathbf{x}|\theta)/q(u|\theta) = 1/\binom{N}{\sum x_n}$.

Factorization theorem: To find a sufficient statistic by simple inspection of the pdf or pmf of the sample \mathbf{X} , the factorization theorem provides for a very important tool. The theorem states that if $p(\mathbf{X}|\theta)$ is the pdf or pmf of a sample \mathbf{X} , then a statistic $\Upsilon(\mathbf{X})$ is sufficient for θ iff there exists functions $f(\Upsilon(\mathbf{X})|\theta)$ and $h(\mathbf{x})$ such that $\forall \mathbf{X}$ and $\forall \theta$, $p(\mathbf{X}|\theta) = h(\mathbf{X})f(\Upsilon(\mathbf{X}) = \mathbf{u}|\theta)$

Theorem 2.2.1. Sufficient statistics for exponential families: If $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ are N i.i.d. observations from a pdf or pmf $p(\mathbf{x}|\boldsymbol{\theta})$ that belongs to an exponential family of distributions with functional forms: $p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta}) \exp\left(\sum_{k=1}^K w_k(\boldsymbol{\theta})\Upsilon_k(\mathbf{x})\right)$, then $\boldsymbol{\Upsilon}(\mathbf{X}) = \left(\sum_{n=1}^N [\Upsilon_1(\mathbf{x}_n), \dots, \Upsilon_K(\mathbf{x}_n)]\right)$ is a sufficient statistic for the K -dimensional parameter $\boldsymbol{\theta}$.

Proof. We can write the joint pdf or pmf of $p(\mathbf{X}|\boldsymbol{\theta})$ as

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{n=1}^N \left\{ h(\mathbf{x}_n)g(\boldsymbol{\theta}) \exp\left(\sum_{k=1}^K w_k(\boldsymbol{\theta})\Upsilon_k(\mathbf{x}_n)\right) \right\} = \underbrace{g(\boldsymbol{\theta})^N \exp\left(\sum_{k=1}^K w_k(\boldsymbol{\theta}) \left[\sum_{n=1}^N \Upsilon_k(\mathbf{x}_n)\right]\right)}_{f(\boldsymbol{\Upsilon}(\mathbf{x})|\boldsymbol{\theta})} \underbrace{\prod_{n=1}^N h(\mathbf{x}_n)}_{h(\mathbf{x})} \quad (2.27)$$

Hence, By the Factorization theorem, $\left(\sum_{n=1}^N \{\Upsilon_1(\mathbf{x}_n), \dots, \Upsilon_K(\mathbf{x}_n)\}\right)$ is a sufficient statistic for $\boldsymbol{\theta}$. \square

Let us now consider the problem of estimating the parameter vector $\boldsymbol{\theta}$ in the general exponential family distribution given in Equ. 2.5. Taking the gradient of both sides of Equ. 2.5 w.r.t. $\boldsymbol{\theta}$, we have:

$$\begin{aligned} \nabla g(\boldsymbol{\theta}) \int h(\mathbf{X}) \exp\left\{\boldsymbol{\theta}^T \boldsymbol{\Upsilon}(\mathbf{X})\right\} d\mathbf{X} + g(\boldsymbol{\theta}) \int h(\mathbf{X}) \exp\left\{\boldsymbol{\theta}^T \boldsymbol{\Upsilon}(\mathbf{X})\right\} \boldsymbol{\Upsilon}(\mathbf{X}) d\mathbf{X} \\ \implies -\frac{1}{g(\boldsymbol{\theta})} \nabla g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \int h(\mathbf{X}) \exp\left\{\boldsymbol{\theta}^T \boldsymbol{\Upsilon}(\mathbf{X})\right\} \boldsymbol{\Upsilon}(\mathbf{X}) d\mathbf{X} \\ \implies -\nabla \ln g(\boldsymbol{\theta}) = E_{p(\mathbf{x}|\boldsymbol{\theta})}[\boldsymbol{\Upsilon}(\mathbf{X})] \end{aligned} \quad (2.28)$$

This can also be derived by setting $w_i(\boldsymbol{\theta}) = \theta_i$ in Equ. 2.13 and then using Equ. 2.14.

Example: Let us consider a set of independent and identically distributed data denoted by $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, for which the likelihood function is given by

$$p(\mathbf{X}|\boldsymbol{\theta}) = \left(\prod_{n=1}^N h(\mathbf{x}_n)\right) g(\boldsymbol{\theta})^N \exp\left\{\sum_{n=1}^N \boldsymbol{\theta}^T \boldsymbol{\Upsilon}(\mathbf{x}_n)\right\} \quad (2.29)$$

Taking the derivative of $\ln p(\mathbf{X}|\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ and setting it to 0, we obtain the following Maximum Likelihood (ML) estimate of $\boldsymbol{\theta}$:

$$-\nabla \ln g(\boldsymbol{\theta}_{ML}) = \frac{1}{N} \sum_{n=1}^N \boldsymbol{\Upsilon}(\mathbf{x}_n) \quad (2.30)$$

The solution of the MLE depends on the data only through $\sum_n \boldsymbol{\Upsilon}(\mathbf{x}_n)$ i.e the sufficient statistic of the distribution corresponding to the pdf $p(\mathbf{X}|\boldsymbol{\theta}) = h(\mathbf{X})g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\theta}^T \boldsymbol{\Upsilon}(\mathbf{X})\right\}$ as given in Equ. 2.4. The importance of the likelihood function as a tool for data reduction is highlighted by the **likelihood principle**.

Definition 2.2.3. The likelihood principle states that if \mathbf{x} and \mathbf{y} are two sample points such that $\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}) = C(\mathbf{x}, \mathbf{y})\mathcal{L}(\mathbf{y}|\boldsymbol{\theta}) \forall \boldsymbol{\theta}$ and some constant $C(\mathbf{x}, \mathbf{y})$ then the conclusions drawn from \mathbf{x} and \mathbf{y} should be identical [Casella and Berger, 2001].

The constant $C(\mathbf{x}, \mathbf{y})$ may be different for different pairs of (\mathbf{x}, \mathbf{y}) but is independent of $\boldsymbol{\theta}$. The implication of this principle is that if two sample points have only proportional likelihoods, they contain

equivalent information about θ .

The function $\Upsilon(\mathbf{X})$ is a sufficient statistic *under the sufficiency principle* and the **value** of $\Upsilon(\mathbf{X})$ is the *set* of all likelihood functions proportional to $\mathcal{L}(\mathbf{X}|\theta)$ *under the likelihood principle*.

2.2.3 Conjugate Priors

Prior distributions over parameters provide us with a mathematically robust formulation to incorporate uncertainty over point estimates of parameters. For a specific dataset \mathbf{X} , we can only have one specific optimal configuration of the parameters but such a configuration is not flexible enough to handle the variances in another set of observations \mathbf{X}' sampled from $p(\mathbf{X})$. In other words, priors usually help improve the generalization power of a model.

In general, for a given probability distribution $p(\mathbf{X}|\theta)$, we seek a prior $p(\theta)$ which is conjugate to the likelihood function so that the posterior distribution has the same functional form as the prior. Using conjugate priors simplifies approximate inference for exponential family models and at the same time provides simple intuitive explanations for updates of the posterior distributions over hidden variables. For any member of the exponential family $p(\mathbf{X}|\theta) = h(\mathbf{X})g(\theta) \exp\{\theta^T \Upsilon(\mathbf{X})\}$, there exists a conjugate prior of the form:

$$p(\theta|\nu, \chi) = h(\chi, \nu)g(\theta)^\nu \exp\{\theta^T \chi\} \quad (2.31)$$

where $h(\chi, \nu)$ is a normalization coefficient, and $g(\theta)$ is the same log partition or cumulant function that appears in Equ. 2.4. If we multiply Equ. 2.29 i.e. $p(\mathbf{X}|\theta) = \left(\prod_{n=1}^N h(\mathbf{x}_n)\right) g(\theta)^N \exp\left\{\sum_{n=1}^N \theta^T (\Upsilon(\mathbf{x}_n))\right\}$ with Equ. 2.31, we obtain:

$$p(\theta|\mathbf{x}, \chi, \nu) = \mathcal{Z}(\chi, \nu)g(\theta)^{\nu+N} \exp\left\{\sum_{n=1}^N \theta^T (\Upsilon(\mathbf{x}_n) + \chi)\right\} \quad (2.32)$$

$$\text{where, } \mathcal{Z}(\chi, \nu)^{-1} = \int g(\theta)^{\nu+N} \exp\left\{\sum_{n=1}^N \theta^T (\Upsilon(\mathbf{x}_n) + \chi)\right\} d\theta$$

The posterior distribution over θ thus is again in exponential form as shown in Equ. 2.32. The parameter ν can be interpreted as the effective *count* of pseudo-observations in the prior and the parameter (vector) χ represents the *values* of the pseudo-observations with each component being mapped to the corresponding component of the sufficient statistic (vector) $\Upsilon(\mathbf{X})$. Examples of multinomial-Dirichlet conjugacy in the basic topic model (LDA) are shown in [Griffiths and Steyvers, 2004] in the context of Gibbs sampling and in [Blei et al., 2003] in the context of a full Bayesian treatment of the topic multinomials. A full Bayesian treatment of some parameters simply mean that we do not want a specific point estimate of the parameter given the data but rather allow for some variance in its estimation (i.e. posit a posterior probability distribution) given some prior distributions with “hyparameters” over the parameters of the model. The uncertainty arises out of the deviation of the moments of the test samples w.r.t those obtained from the training dataset. Additionally, section 2.4 explains why finding the posterior is so important from a risk minimization point of view. However, it is worthwhile to mention that we pay a price by introducing a prior of our choice in that the posterior means of the parameter estimators, i.e. the Bayesian point estimators of θ , are not unbiased estimators unless all of the data collapses on the true mean parameter [Casella and Berger, 2001].

2.2.4 Asymptotics and MLE

The branch of asymptotics in statistics deal with various properties of estimators in the limit of infinite data. An estimator should have the property that it converges to the correct value as sample size becomes infinite. An estimator possessing this property is said to be *consistent*, although, technically speaking, consistency is a property of a sequence of estimators rather than a single one.

A sequence of estimators $\Upsilon_n = \Upsilon_n(\mathbf{x}_1, \dots, \mathbf{x}_n)$ is a consistent sequence of estimators of the parameter θ , if for every $\epsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n \rightarrow \infty} p_{\theta}(|\Upsilon_n - \theta| < \epsilon) = 1 \quad \text{or equivalently} \quad \lim_{n \rightarrow \infty} p_{\theta}(|\Upsilon_n - \theta| \geq \epsilon) = 0 \quad (2.33)$$

This means that a consistent sequence of estimators converges in probability to the parameter θ it is estimating. The sequence of sample means $\Upsilon(\mathbf{X}) = \bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ is a consistent sequence of estimators. The following theorem from [Casella and Berger, 2001] on Maximum Likelihood Estimators also shows that they are consistent.

Theorem 2.2.2. Consistency of Maximum Likelihood Estimators (MLEs): *Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be i.i.d $p(\mathbf{x}|\theta)$ and let $\mathcal{L}(\theta) \equiv \mathcal{L}(\mathbf{X}|\theta) = \prod_{i=1}^n p(\mathbf{x}_i|\theta)$ be the likelihood function. Let $\hat{\theta}$ denote the MLE of θ . Also let $\tau(\theta)$ be a continuous function of θ . Under certain regularity conditions on $p(\mathbf{X}|\theta)$ and hence $\mathcal{L}(\theta)$, for every $\epsilon > 0$ and every $\theta \in \Theta$, $\lim_{n \rightarrow \infty} p_{\theta}(|\tau(\hat{\theta}) - \tau(\theta)| \geq \epsilon) = 0$ i.e. $\tau(\hat{\theta})$ is a consistent estimator of $\tau(\theta)$ since $\tau(\hat{\theta})$ converges to $\tau(\theta)$ in probability.*

Proof. See [Stuart et al., 1999]. For regularity conditions on $p(\mathbf{x}|\theta)$, see below. \square

Regularity conditions:

- [i] The samples $\mathbf{x}_1, \dots, \mathbf{x}_n$ be i.i.d $p(\mathbf{x}|\theta)$
- [ii] The parameter θ is identifiable i.e. if $\theta \neq \theta'$ then $p(\mathbf{x}|\theta) \neq p(\mathbf{x}|\theta')$
- [iii] The densities $p(\mathbf{X}|\theta)$ have common support and are differentiable in θ
- [iv] The parameter space Ω contains an open set ω of which the true parameter value θ^{true} is an interior point

Another very important concept for sequence of estimators with regards to asymptotics is that of *Efficiency*. While the property of consistency is concerned with the asymptotic accuracy of an estimator, efficiency is more concerned with the asymptotic variance of an estimator.

We now state an important theorem from [Casella and Berger, 2001] about the asymptotic efficiency of MLEs. Before stating the theorem, we first formally define what it means for a sequence of estimators to be *asymptotically efficient*. A sequence of estimators Υ_n for a parameter $\tau(\theta)$ is asymptotically efficient if $\sqrt{n}[\Upsilon_n - \tau(\theta)] \rightarrow \mathcal{N}(0, \nu(\theta))$ in distribution where $\nu(\theta) = [\tau'(\theta)]^2 / E_{p(\mathbf{x}|\theta)} \left(\left(\frac{\partial}{\partial \theta} \log p(\mathbf{x}|\theta) \right)^2 \right)$. This means that the asymptotic variance of Υ_n achieves Cramér-Rao lower bound which is a lower bound on the variance of the best unbiased estimator of θ and is defined as follows:

Definition 2.2.4. Cramér Rao-Inequality: *Let $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be a sample with pdf $p(\mathbf{x}|\theta)$ and let $\Upsilon(\mathbf{X}) = \Upsilon(\mathbf{x}_1, \dots, \mathbf{x}_N)$ be any estimator satisfying*

$$\frac{\partial}{\partial \theta} E_{p(\mathbf{x}|\theta)} (\Upsilon(\mathbf{X})) = \int_{\mathcal{X}} \Upsilon(\mathbf{X}) p(\mathbf{X}|\theta) d\mathbf{X} \quad \text{and} \quad Var_{p(\mathbf{x}|\theta)} (\Upsilon(\mathbf{X})) < \infty, \quad (2.34)$$

then,

$$\text{Var}_{p(\mathbf{X}|\theta)}(\Upsilon(\mathbf{X})) \geq \frac{(E_{p(\mathbf{X}|\theta)}[\Upsilon(\mathbf{X})])^2}{E_{p(\mathbf{X}|\theta)}\left[\left(\frac{\partial}{\partial\theta}\log p(\mathbf{X}|\theta)\right)^2\right]} \quad (2.35)$$

The quantity $E_{p(\mathbf{X}|\theta)}\left(\frac{\partial}{\partial\theta}\log p(\mathbf{X}|\theta)\right)^2$ is called the *information number* or the Fisher information of the sample. This number gives a bound on the variance of the best unbiased estimator of θ . As the number gets larger, we have more information about θ and the bound on the variance of the best unbiased estimator becomes smaller. Given an estimator Υ_n based on a sample size n , the finite-sample variance $\text{Var}(\Upsilon_n)$ is first calculated and then the limit $\lim_{k \rightarrow \infty} k_n \text{Var}(\Upsilon_n)$ is evaluated with k_n being some normalizing constant.

Theorem 2.2.3. Asymptotic Efficiency of Maximum Likelihood Estimators (MLEs): Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be i.i.d $p(\mathbf{X}|\theta)$ and let $\mathcal{L}(\theta) \equiv \mathcal{L}(\mathbf{X}|\theta) = \prod_{i=1}^n p(\mathbf{x}_i|\theta)$ be the likelihood function. Let $\hat{\theta}$ denote the MLE of θ . Also let $\tau(\theta)$ be a continuous function of θ . Under certain regularity conditions on $p(\mathbf{x}|\theta)$ and hence $\mathcal{L}(\theta)$, $\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \rightarrow \mathcal{N}(0, \nu(\theta))$ where $\nu(\theta)$ is the Cramér-Rao lower bound i.e. $\tau(\hat{\theta})$ is a consistent and efficient estimator of $\tau(\theta)$.

Proof. See [Casella and Berger, 2001] □

The implication of theorem 2.2.3 is that a larger training set (hypothetically in the limit of infinite data) almost always make any statistical learning algorithm better estimate the model parameters if the assumptions of the modeling process very closely resemble the true generation process.

2.3 How much training data is necessary?

In the supervised learning scenario, the problem setup seeks to find the function $f : \mathcal{X} \rightarrow \mathcal{T}$ which maps an input space \mathcal{X} to some output space \mathcal{T} . The range of the function f can be discrete (often binary) as in the case of classification or continuous as in the case of regression. However, we are never told what the function f is, instead, we just observe some sample $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \in \mathcal{X}$ from \mathbf{X} and the realization $t_i \in \mathcal{T}$ of the function f for each individual datum \mathbf{x}_i . A learning algorithm is a function that operates on \mathbf{X} to produce an mapping $\mathbf{X} \rightarrow \mathbf{T}$. The goal of the learning algorithm is to choose a function g from a set of candidate functions referred to as the hypotheses set \mathcal{H} such that g approximates f as best as possible. Note that the cardinality of \mathcal{H} can be infinite, for e.g. all possible straight lines in \mathbb{R}^2 . The main concern of the learning algorithm is to find g so that it performs optimally on samples not in \mathbf{X} but assumed to be generated from \mathcal{X} under some unknown probability distribution on \mathcal{X} i.e. the learning algorithm must be able to bound the out-of-sample error $E_{out}(g)$ based on the in-sample error $E_{in}(g)$. We can express this formally as:

$$p(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2Me^{-2\epsilon^2 N} \quad (2.36)$$

for any $\epsilon > 0$. What Equ. 2.36 says is that the probability of a “bad event” $|E_{in}(g) - E_{out}(g)| > \epsilon$ happening must get lower as we observe more realizations of \mathbf{X} through the function $f : \mathcal{X} \rightarrow \mathcal{T}$. Inequality 2.36 is known as Hoeffding’s inequality [Abu-Mostafa et al., 2012] and is a function of the number of candidate hypotheses M and the sample (i.e. dataset) size N . Hoeffding’s inequality can be expressed in terms of a *generalization bound* such that the probability of the “good event” $|E_{in}(g) -$

$E_{out}(g) \leq \epsilon$ happening is at least $1 - \delta$ where δ is should be as small as possible. Identifying this with Equ. 2.36, we observe that $\delta = 2Me^{-2\epsilon^2 N}$ from which we obtain:

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}} \quad (2.37)$$

where δ is a tolerance level that replaces ϵ from Equ. 2.36. Thus the generalization bound for a learning algorithm based only on the in-sample error E_{in} and a second term which depends on M which is the size of the set of candidate hypotheses \mathcal{H} .

To make the generalization bound meaningful due to the problem that the cardinality of \mathcal{H} being infinite, a notion of a *growth function* is introduced for \mathcal{H} on the sample points $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ as $m_{\mathcal{H}}(N) = \max_{\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \dots, \mathbf{x}_N)| = \max_{\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \in \mathcal{X}} |\{h(\mathbf{x}_1), \dots, h(\mathbf{x}_N) \mid h \in \mathcal{H}\}|$ where $|\cdot|$ is the cardinality of the corresponding set.

To compute $m_{\mathcal{H}}(N)$, we consider all possible choices of N points $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ from \mathcal{X} and select the one which results in the most number of dichotomies. It is easy to see that $m_{\mathcal{H}}(N) \leq 2^N$. The notion of a **break point** is then used to polynomially bound the growth function [Abu-Mostafa et al., 2012]. A break point for a hypothesis set \mathcal{H} is the minimum dataset size that cannot be shattered by \mathcal{H} . Shattering means that \mathcal{H} is capable of producing all possible dichotomies (binary labels) on the dataset.

The Vapnik Chervonenkis (VC) dimension is the order of the polynomial bound on $m_{\mathcal{H}}(N)$ [Abu-Mostafa et al., 2012] expressed as:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i} \leq N^{d_{VC}} + 1 \quad (2.38)$$

If we replace M in Equ. 2.37 by $m_{\mathcal{H}}(N)$, we obtain a bound of the form:

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2m_{\mathcal{H}}(N)}{\delta}} \quad (2.39)$$

For any finite d_{VC} , the error bar for the generalization error will converge to zero at a speed determined by d_{VC} —smaller the d_{VC} faster the convergence. Further, if $d_{VC} \geq N$ then there exists a dataset \mathcal{D} of size N such that \mathcal{H} shatters \mathcal{D} [Abu-Mostafa et al., 2012] and the following theorem constitutes one of the most important results in statistical learning theory.

Theorem 2.3.1. *VC generalization bound.*

For any tolerance $\delta > 0$,

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}} \quad (2.40)$$

with probability $1 - \delta$

Sample complexity: The sample complexity of a learning model denotes the number of training examples N needed to achieve a certain generalization performance and can be obtained using the VC bound.

The generalization error is bounded by $\sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$ and making $\sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}} \leq \epsilon$, we obtain:

$$\begin{aligned} N &\geq \frac{8}{\epsilon^2} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta} \\ &= \frac{8}{\epsilon^2} \ln \frac{4((2N)^{d_{VC}} + 1)}{\delta} \end{aligned} \quad (2.41)$$

The error tolerance ϵ determines the allowed generalization error and the confidence parameter δ determines how often is the error tolerance violated. However, the caveat in using the VC bound in its original form is that it often gives gross over-estimate for the value of N in the order of thousands; for e.g. with $d_{VC} = 5$, we need $N = 50,000$. In practical scenarios, the constant of proportionality should be set closer to 10 [Abu-Mostafa et al., 2012].

The polynomial in d_{VC} in Equ. 2.40 also suggests that it is imperative to choose a model with a lower VC dimension than a higher one. However, in general a rigorous evaluation of the VC dimension for complex models is skipped in favor of the effective number of parameters of the model [Abu-Mostafa et al., 2012]—the more complex model will need more training samples to provide for better generalization performance.

Problems arise when we cannot measure the dichotomization due to the absence of any labeling of the training dataset in the unsupervised scenario and we cannot readily apply VC analysis in this case. If we denote $p(\mathbf{X}, \mathbf{t})$ to be the distribution which governs the true relationship between the input \mathbf{X} and the target class \mathbf{t} , then, neither do we know p nor we know \mathbf{t} . We are only presented with a set of unlabeled examples $\mathbf{x}_1, \dots, \mathbf{x}_n$ drawn i.i.d. from the true p and we need to approximate it with some $\hat{p}(\mathbf{X})$ with some parameters that explain the observations as best as possible. A standard approach is to use EM algorithm to optimize the empirical likelihood of the incomplete data— $\hat{E} \log p(\mathbf{X}|\mathbf{Z}, \theta)$ where $\hat{E}f(\mathbf{X}) = \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i)$ denotes averaging over some function f of the training data. The posterior over the discrete random variable \mathbf{Z} accounts for the explanation of the true class label t_n for the observation \mathbf{x}_n as best as possible.

The discrepancy between p^* and $p_{\widehat{\theta}_{EM}}$ depends on the distribution over \mathbf{t} whereas learning depends only on the distribution over \mathbf{X} . For most unsupervised probabilistic models, we optimize a non-convex objective owing to tractability requirements and hence the locally optimal $\widehat{\theta}_{EM}$ is different from the globally optimal $\widehat{\theta}$ which can be obtained using unlimited computational resources to process the dataset \mathbf{X} . Further $\widehat{\theta}$ can be different from the true parameters of p^* due to noise in the data.

It has been reported in [Liang and Klein, 2008] that local optima issues which typically plague EM can be somewhat alleviated by increasing the number of training examples which results in less noise in the aggregate sufficient statistics. This is intuitive from the point of view of asymptotic efficiency of MLEs however rigorous analyses of generalization errors in unsupervised models is an extremely difficult problem. Some earlier efforts on such analyses on simpler models like Principle Component Analysis and K-Means have been reported in [Hansen and Larsen, 1996].

As an alternative, the “Meta model” in [Liang and Klein, 2008] is used for analyzing EM. They look at predictions made by the model and study how these predictions change over time instead of treating parameters as the primary object of study. A similar study in the context of topic models has been performed in [Chang et al., 2009]. On the other hand, if we are just interested in clustering accuracy, then, for a topic model like LDA [Blei et al., 2003], we can set the number of topic multinomials to be some parameter K (e.g. fifty topics for the Twenty Newsgroup [Lang, 1995] dataset with 20 classes) and use a clustering quality metric like *purity* or *normalized mutual information* [Manning et al., 2008] as a meta model to analyze EM. We also use a similar criteria for evaluating topic models from a summarization perspective using the widely used automatic ROUGE [Lin and Hovy, 2003] evaluation for problems on multi-document text summarization as well as summarizing videos to text (see Chapters 5 and 6).

A very recent but mostly theoretical research on the amount of data needed to learn the parameters of a topic model is mentioned in Section 2.7.2.

2.4 Bayes Estimator and its Relation to Posterior

The main idea behind building several models to describe the causal phenomenon of the same observed dataset \mathbf{X} is to explore several different families of the parameters θ which explain the statistical moments (mainly the first and second order) of the observations as best as possible. A decision is then taken to choose one these models from based on some value of a loss function.

The parameter θ is thought to be an unknown, but fixed, quantity. A random sample $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is drawn from a population indexed by θ and based on the sample, knowledge about the value of θ is obtained. In the Bayesian approach θ is considered to be a random quantity whose variation can be described by a prior probability distribution imposed over it. A sample is then taken from a population indexed by θ and the prior distribution is updated with information obtained from the statistics of this sample. If we denote the prior distribution by $p(\theta)$ and the sampling distribution by $p(\mathbf{X}|\theta)$, then the posterior distribution i.e. the conditional distribution of θ given the sample, \mathbf{X} , is

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{A(\mathbf{X})} \quad (2.42)$$

where, $A(\mathbf{X})$ is the marginal distribution of \mathbf{X} i.e. the normalizer used to make p a valid probability distribution and is defined as:

$$A(\mathbf{X}) = \int p(\mathbf{X}|\theta)p(\theta)d\theta \quad (2.43)$$

Notice that the posterior distribution is a conditional distribution, conditional upon observing the sample. The posterior distribution can now be used to make statements about θ which is still a random quantity. For example the mean of the posterior distribution can be used as a point estimate for θ .

Loss Function Optimality: Generally point estimators are based on their mean squared error performance which is defined as follows:

Definition 2.4.1. *The mean squared error (MSE) of an estimator $\Upsilon \equiv \Upsilon(\mathbf{X})$ of a parameter θ is defined by the function $E_{p(\mathbf{X}|\theta)} [(\Upsilon - \theta)^2]$*

The mean squared error is analytically tractable and can be decomposed into terms that represent the bias which measures the accuracy and variance which measures the precision of the estimator as shown in Equ. 2.44.

$$E_{p(\mathbf{X}|\theta)} [(\Upsilon - \theta)^2] = Var_{p(\mathbf{X}|\theta)}(\Upsilon) + (E_{p(\mathbf{X}|\theta)}[\Upsilon] - \theta)^2 = Var_{p(\mathbf{X}|\theta)}(\Upsilon) + (Bias_{p(\mathbf{X}|\theta)}(\Upsilon))^2 \quad (2.44)$$

The bias of a point estimator Υ of a parameter θ is the difference between the expected value of Υ and θ . Needless to mention that a good estimator should have both a low bias as well as a low variance. However, mean squared error is only a special case of an error function or loss function. After a set of samples $\mathbf{X} = \mathbf{x}$ has been observed, where $\mathbf{X} \sim p(\mathbf{X}|\theta)$, $\theta \in \Theta$, a decision, Δ , is made regarding θ . The set of allowable decisions form the action space \mathcal{A} .

The quality of an estimator is measured by its risk function in a loss function formulation—for an estimator $\Upsilon(\mathbf{X})$ of θ , the risk function can be written as:

$$R(\theta, \Upsilon) = E_{p(\mathbf{X}|\theta)}[L(\theta, \Upsilon(\mathbf{X}))] \quad (2.45)$$

where the expectation is over \mathbf{X} . Thus the risk function is the average loss that is incurred if the estimator $\Upsilon(\mathbf{X})$ is used given a particular θ . Since the true value of θ is unknown, our goal is to use an estimator that has a small value of $R(\theta, \Upsilon)$ for all values of θ . Different estimators are just compared based on the values of some predefined risk functions. For example, if we build a topic model with K_1 topics and another with K_2 topics, then we can make a decision on which model to choose based on how high the relative log likelihoods of the observations in an held-out test set are under the respective models. For squared error loss, the risk function is the Mean Squared Error (MSE). The MSE of an estimator is just $E_{p(\mathbf{X}|\theta)}[L(\theta, \Upsilon(\mathbf{X}))] = R(\theta, \Upsilon)$ if $L(\Upsilon(\mathbf{X}) = \mathbf{u}, \theta) = \|\mathbf{u} - \theta\|_2$.

The problem of loss function optimality can also be defined through a Bayesian approach where there typically is a prior distribution $p(\theta)$. In a Bayesian approach this prior distribution is used to compute an average risk, L_{Bayes} , known as the *Bayes risk* defined as:

$$L_{Bayes} = \int_{\theta \in \Theta} R(\theta, \Upsilon) p(\theta) d\theta \quad (2.46)$$

Averaging the risk function gives us a number for assessing the the performance of an estimator w.r.t a given loss function. An estimator that yields the smallest value of Bayes risk is called the Bayes rule w.r.t. a prior $p(\theta)$ and is often denoted as $\Delta^{p(\theta)}$ [Casella and Berger, 2001].

For $\mathbf{X} \sim p(\mathbf{X}|\theta)$ and $\theta \sim p(\theta)$, the Bayes risk of choosing a decision rule Δ can be written as

$$\int_{\theta \in \Theta} R(\theta, \Delta) p(\theta) d\theta = \int_{\theta \in \Theta} \left(\int_{\mathcal{X}} L(\theta, \Delta(\mathbf{X})) p(\mathbf{X}|\theta) d\mathbf{X} \right) p(\theta) d\theta \quad (2.47)$$

Now if we write $p(\mathbf{X}|\theta)p(\theta) = p(\theta|\mathbf{X})p(\mathbf{X})$, where $p(\theta|\mathbf{X})$ is the posterior distribution of θ and $p(\mathbf{X})$ is the marginal distribution of \mathbf{X} , we can write the Bayes risk as

$$\int_{\theta \in \Theta} R(\theta, \Delta) p(\theta) d\theta = \int_{\mathcal{X}} \left(\int_{\theta \in \Theta} L(\theta, \Delta(\mathbf{X})) p(\theta|\mathbf{X}) d\theta \right) p(\mathbf{X}) d\mathbf{X} \quad (2.48)$$

The quantity $\int_{\theta \in \Theta} L(\theta, \Delta(\mathbf{X})) p(\theta|\mathbf{X}) d\theta$ in Equ. 2.48 is the expected value of the loss function with respect to the posterior distribution over θ and is called the *posterior expected loss*. It is a function of \mathbf{X} only and not a function of θ . Thus for each random sample \mathbf{X} , if we choose the action $\Delta(\mathbf{X})$ to minimize the posterior expected loss, we will minimize Bayes risk. Hence for a given observation \mathbf{x} , the Bayes rule minimizes posterior expected loss.

The loss function optimality also is very closely related to the classical hypothesis testing problem which focuses on two allowable actions—given a null hypothesis \mathbb{H}_0 , we have to make a choice between “accepting \mathbb{H}_0 ” or “rejecting \mathbb{H}_0 .” If the first choice corresponds to taking an action a_0 and the latter corresponds to taking an action a_1 , then the action space in hypothesis testing is the two point set $\mathcal{A} = \{a_0, a_1\}$. The decision rule $\delta(\mathbf{X})$ in this case is the hypothesis test which is a function of the sample space \mathcal{X} and takes only two values, a_0 and a_1 . The set of sample values $\{\mathbf{x} : \delta(\mathbf{X}) = a_0\}$ is the acceptance region w.r.t the hypothesis test while the set of sample values $\{\mathbf{x} : \delta(\mathbf{X}) = a_1\}$ is the rejection region w.r.t the hypothesis test.

The loss function in this case can be written as:

$$L(\theta, a_0) = \begin{cases} 0, & \text{if } \theta \in \Theta_0 \\ c_2, & \text{if } \theta \in \widehat{\Theta}_0 \end{cases} \quad L(\theta, a_1) = \begin{cases} c_1, & \text{if } \theta \in \Theta_0 \\ 0, & \text{if } \theta \in \widehat{\Theta}_0 \end{cases} \quad (2.49)$$

where Θ_0 is the parameter space of the null hypothesis and $\widehat{\Theta}_0$ is the parameter space of the alternative hypothesis. In this definition of the loss function, c_1 is the cost of a Type 1 error i.e. the error of falsely rejecting \mathbb{H}_0 and c_2 is the cost of a Type 2 error i.e. the error of falsely accepting \mathbb{H}_0 . The ratio c_2/c_1 is more a relevant quantity to evaluate rather than individual errors. The risk function associated with the loss function expressions in Equ. 2.49 can similarly be written as:

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) = \begin{cases} 0 \times p(\mathbf{X}|\boldsymbol{\theta})(\boldsymbol{\delta}(\mathbf{X}) = a_0) + c_1 \times p(\mathbf{X}|\boldsymbol{\theta})(\boldsymbol{\delta}(\mathbf{X}) = a_1), & \text{if } \boldsymbol{\theta} \in \boldsymbol{\theta}_0 \\ c_2 \times p(\mathbf{X}|\boldsymbol{\theta})(\boldsymbol{\delta}(\mathbf{X}) = a_0) + 0 \times p(\mathbf{X}|\boldsymbol{\theta})(\boldsymbol{\delta}(\mathbf{X}) = a_1), & \text{if } \boldsymbol{\theta} \in \widehat{\boldsymbol{\theta}}_0 \end{cases} \quad (2.50)$$

2.5 Bayesian vs. Frequentist

Building upon the discussion on loss function optimality in the previous section, we highlight a few key differences between two classical schools of machine learning—the Bayesian perspective vis-a-vis the frequentist perspective. The differences can be summarized as follows:

Bayesian perspective:

1. It is a *conditional perspective*: The inferences are made conditional on the current data
2. Usefulness: It is natural to use Bayesian modeling for a project which involves one or more domain experts
3. It has an *optimistic viewpoint*: Given the current dataset, make the best possible use of sophisticated inference tools to provide for a best fit of the validation data to the model.

Frequentist perspective:

1. It has an *unconditional perspective*: The inferential methods are supposed to produce good answers in repeated use
2. Usefulness: It is natural to use the frequentist perspective in the setting of writing software that will be used by many people with many data sets
3. It has a *pessimistic viewpoint*: protect ourselves against bad decisions by averaging out the randomness about the sample space given that the inference procedure is based on a simplification of reality

Decision theoretic perspective:

1. Define a family of probability models for the input sample \mathbf{x} , indexed by a “parameter” $\boldsymbol{\theta}$
2. Define a “procedure” $\Delta(\mathbf{X})$ that operates on the data to produce a decision $H(\mathbf{X})$; $H(\mathbf{X})$ is often called a hypothesis belonging to a (infinite in case of real valued parameters) hypothesis space \mathcal{H}
3. Define a loss function: $L(\Delta(\mathbf{X}), \boldsymbol{\theta})$

Using the notations in Section 2.4, the procedure $\Delta(\mathbf{X})$ amounts to computing a statistic $\Upsilon(\mathbf{X})$ through $H(\mathbf{x})$ which best estimates the parameters $\boldsymbol{\theta}$ of a model. The goal, then, is to use the loss function to compare procedures and hence computing the sufficient statistics, but both of its arguments are unknown. The input space \mathcal{X} from which \mathbf{X} is drawn is definitely unknown and $\boldsymbol{\theta}$ is also unknown since we do

not also know about the exact distribution that generated the data. The question then arises: “how can we optimize over the loss function to choose the right hypothesis $H(\mathbf{x})$ to compute Υ ?” These two unknowns give rise to two perspectives—either we start with computing $\Upsilon(\mathbf{X})$ or we start with θ .

From the frequentist perspective, we fix θ and take expectation over \mathbf{X} in terms of $\Upsilon(\mathbf{X})$ w.r.t. a particular θ . The randomness about \mathbf{X} thus goes away. In other words we have the following:

$$R(\theta, \Upsilon) = E_{p(\mathbf{X}|\theta)}[L(\Upsilon(\mathbf{X}), \theta)|\theta] \quad (2.51)$$

Note here that $R(\theta)$ is still not a single number since it is dependent on a particular distribution from among a family of distributions indexed by θ .

From the Bayesian perspective, it is fine to put a distribution *over* θ and integrate out the randomness over θ given \mathbf{X} . This gives rise to a single number $R'(\mathbf{X})$ which is called Bayesian Risk which is conditioned on \mathbf{X} . Thus we can optimize over $\Upsilon(\mathbf{X})$ having defined $R'(\mathbf{X})$ as:

$$R'(\mathbf{X}) = E_{p(\theta|\mathbf{X})}[L(\Upsilon(\mathbf{X}), \theta|\mathbf{X})] \quad (2.52)$$

It is interesting to note that we can plug-in $R(\theta, \Upsilon)$ from Equ. 2.51 to further refine $R'(\mathbf{X})$ and vice versa for $\Upsilon(\mathbf{X})$ and keep iterating. Doing this yields exactly the same conclusion provided the integrals exist in their respective domains [Fubini, 1958].

2.6 Expectation Maximization (EM) and variational Bayesian EM (VBEM)

The Expectation Maximization (EM) [Dempster et al., 1977] machinery is an algorithm rather than a direct point estimator that is based on the idea of reducing the difficulty of a likelihood maximization, usually those involving hidden state variables, with a sequence of easier maximizations whose limit yields the answer to the first problem.

Let us consider a scenario where $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ are the hidden variables and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)$ are the observed data, making (Z, X) the complete data. The densities $p(\cdot|\theta)$ of \mathbf{X} and $p(\cdot|\theta)$ of (\mathbf{Z}, \mathbf{X}) have the relationship

$$p(\mathbf{X}|\theta) = \int p(\mathbf{Z}, \mathbf{X}|\theta)d\mathbf{Z} \quad (2.53)$$

with sums replacing integrals in the discrete case. As for the likelihoods, $\mathcal{L}(\mathbf{X}|\theta) = \prod_{\mathbf{x} \in \mathbf{X}} p(\mathbf{x}|\theta)$ is the incomplete-data likelihood and $\mathcal{L}(\mathbf{Z}, \mathbf{X}|\theta) = \prod_{(\mathbf{z}, \mathbf{x}) \in \{(\mathbf{z}, \mathbf{x})\}} p(\mathbf{z}, \mathbf{x}|\theta)$ is the complete data likelihood. In almost all practical problems of interest, the likelihood $\mathcal{L}(\mathbf{X}|\theta) = \int_{\mathbf{Z}} \mathcal{L}(\mathbf{Z}, \mathbf{X}|\theta)d\mathbf{Z} = \int_{\mathbf{Z}} \mathcal{L}(\mathbf{X}|\mathbf{Z}, \theta)p(\mathbf{Z}|\theta)d\mathbf{Z}$ is difficult to work with in the presence of hidden indicator variables due to exponential configurations of the state space over which the true posteriors over the indicators need to be searched given the input samples. The EM algorithm allows us to maximize the incomplete data log-likelihood $\mathcal{L}(\mathbf{X}|\theta)$ by working only with the complete data log-likelihood $\mathcal{L}(\mathbf{X}, \mathbf{Z}|\theta)$.

Let us now give an example of the EM algorithm through a discussion on a hypothetical scenario.

Example 2.6.1. Suppose we observe the number of queries submitted to a search engine pertaining to flu that is represented by the random variables $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$. The x_i s are mutually independent. Suppose $x_i \sim \text{Poisson}(\beta\tau_i)$ where the underlying rate of query submission is a function of an overall

effect β and an additional factor τ_i which can be the number of over the counter cold medications distributed in area i . We do not observe τ_i but obtain information on it through the random variables $\mathbf{Z} = \{z_1, z_2, \dots, z_N\}$ with $z_i \sim \text{Poisson}(\tau_i)$. The z_i s are also mutually independent.

Discussion: The joint probability mass function (pmf) can be written as:

$$f((z_1, x_1), (z_2, x_2), \dots, (z_N, x_N) | \beta, \tau_{z_1}, \tau_{z_2}, \dots, \tau_{z_N}) = \prod_{n=1}^N \frac{e^{-\beta\tau_n} (\beta\tau_n)^{x_n}}{x_n!} \frac{e^{-\tau_n} (\tau_n)^{z_n}}{z_n!} \quad (2.54)$$

The likelihood can be obtained by differentiation yielding:

$$\hat{\beta} = \frac{\sum_{n=1}^N x_n}{\sum_{n=1}^N z_n} \quad \text{and} \quad \hat{\tau}_n = \frac{z_n + x_n}{1 + \hat{\beta}} \quad (2.55)$$

where we use the maximum likelihood estimates:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{n=1}^N x_n}{\sum_{n=1}^N \tau_n} \quad \text{and} \quad \hat{\tau}_n = \frac{z_n + x_n}{1 + \beta} \\ \text{with } \sum_{n=1}^N \tau_n &= \frac{\sum_{n=1}^N x_n + \sum_{n=1}^N z_n}{1 + \beta} \end{aligned} \quad (2.56)$$

Now suppose that the value of z_1 is missing. The question is then how well can the model parameters β and τ be estimated? Of course we can ignore x_i and proceed with the usual maximum likelihood solution on $n-1$ data points but that will only make our estimate poorer. Additionally this latter approach cannot be used when all z_i s have missing values.

In this scenario, we will want to maximize the incomplete-data likelihood which can be written as

$$\sum_{z_1=0}^{\infty} f((z_1, x_1), (z_2, x_2), \dots, (z_N, x_N) | \beta, \tau_1, \tau_2, \dots, \tau_N) \quad (2.57)$$

where the incomplete-data is $(x_1, (z_2, x_2), \dots, (z_N, x_N))$. Let us now write down the incomplete-data likelihood:

$$\mathcal{L} = \left[\prod_{n=1}^N \frac{e^{-\beta\tau_i} (\beta\tau_i)^{x_n}}{x_n!} \right] \left[\prod_{n=2}^N \frac{e^{-\tau_n} (\tau_n)^{z_n}}{z_n!} \right] \quad (2.58)$$

As before, taking derivatives leads to the maximum likelihood estimates:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{n=1}^N x_n}{\sum_{n=1}^N \hat{\tau}_n}, \quad x_1 = \hat{\tau}_1 \hat{\beta}, \\ z_j + x_j &= \hat{\tau}_j (\hat{\beta} + 1), \quad j = 2, 3, \dots, N \end{aligned} \quad (2.59)$$

which now can be solved using the EM algorithm.

The EM algorithm in its general form: The EM algorithm allows us to maximize $\mathcal{L}(\mathbf{X} | \theta)$ by working only with $\mathcal{L}(\mathbf{Z}, \mathbf{X} | \theta)$ and the conditional distribution of \mathbf{Z} given \mathbf{X} and θ . This conditional turns out to be the posterior distribution of \mathbf{Z} given \mathbf{X} and θ in case of exact EM which is denoted by $q(\mathbf{Z} | \theta, \mathbf{X}) =$

$\frac{\mathcal{L}(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})}{\mathcal{L}(\mathbf{X}|\boldsymbol{\theta})}$ and we have:

$$\log \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) = \log \mathcal{L}(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta}) - \log q(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{X}) \quad (2.60)$$

As \mathbf{Z} is missing data, we replace the right hand side of Equ. 2.60 with its expectation under $q(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{X})$ creating the new identity

$$\log \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) = E_q[\log \mathcal{L}(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})] - E_q[\log q(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{X})] \quad (2.61)$$

where the expression $-E_q[\log q(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{X})]$ is simply the entropy of the q distribution over the posterior for \mathbf{Z} . For the algorithm to proceed, we select an initial value $\boldsymbol{\theta}^{(0)}$, we create a sequence $\boldsymbol{\theta}^{(t)}$ according to

$$\boldsymbol{\theta}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} E_{q(\mathbf{Z}|\boldsymbol{\theta}^{(t)}, \mathbf{X})}[\log \mathcal{L}(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})] \quad (2.62)$$

The E-step of the algorithm computes the expected log likelihood and the M-step finds its maximum w.r.t the parameters. This sequence of maximums $\{\hat{\boldsymbol{\theta}}^{(t)}\}$ satisfy:

$$\mathcal{L}(\mathbf{X}|\hat{\boldsymbol{\theta}}^{(t+1)}) \geq \mathcal{L}(\mathbf{X}|\hat{\boldsymbol{\theta}}^{(t)}) \quad (2.63)$$

with the equality holding iff $E_{q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})}[\log \mathcal{L}(\mathbf{Z}, \mathbf{X}|\hat{\boldsymbol{\theta}}^{(t+1)})] = E_{q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})}[\log \mathcal{L}(\mathbf{Z}, \mathbf{X}|\hat{\boldsymbol{\theta}}^{(t)})]$. To validate this we proceed as follows. At time step t , we have:

$$\log \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) = E_{q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})}[\log \mathcal{L}(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})] - E_{q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})}[\log q(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{X})] \quad (2.64)$$

The next iterate $\hat{\boldsymbol{\theta}}^{(t+1)}$ is obtained by maximizing the new complete-data log likelihood. Thus for any $\boldsymbol{\theta}$, $E_{q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})}[\log \mathcal{L}(\mathbf{Z}, \mathbf{X}|\hat{\boldsymbol{\theta}}^{(t+1)})] \geq E_{q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})}[\log \mathcal{L}(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})]$.

Now, if p and q are densities, since logarithm is a concave function, from Jensen's inequality we have

$$\begin{aligned} \int \log \left(\frac{p(x)}{q(x)} \right) q(x) dx &\leq \log \int \left(\frac{p(x)}{q(x)} \right) q(x) dx = \log \int p(x) dx = 0 \\ \implies \int \log[p(x)]q(x) dx &\leq \int \log[q(x)]q(x) dx \end{aligned} \quad (2.65)$$

We thus have:

$$E_{q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})}[\log \mathcal{L}(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{X})] = \int \log[q(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{X})]q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X}) \leq \int \log[q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})]q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X}) \quad (2.66)$$

$$\therefore \int \log[q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t+1)}, \mathbf{X})]q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X}) \leq \int \log[q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})]q(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{(t)}, \mathbf{X})$$

This implies that the second term in the right hand side of Equ. 2.64 is always decreasing. The sequence of iterates $\{\hat{\boldsymbol{\theta}}^{(t)}\}$ thus monotonically improves the log likelihood under the application of the EM algorithm. \square

Returning to the problem scenario at hand, let us denote $(\mathbf{z}, \mathbf{x}) = ((z_1, x_1), (z_2, x_2), \dots, (z_N, x_N))$

to be the complete data and $(\mathbf{z}_{(-1)}, \mathbf{x}) = (x_1, (z_2, x_2), \dots, (z_N, x_N))$ to be the incomplete data.

$$\begin{aligned}
E_{q(z_1|\mathbf{x},\theta)}[\log \mathcal{L}(\mathbf{z}, \mathbf{x}|\mathbf{z}_{(-1)}, \mathbf{x}, \Theta)] &= \sum_{z_1=0}^{\infty} \log \left(\prod_{n=1}^N \frac{e^{-\beta\tau_n} (\beta\tau_n)^{x_n}}{x_n!} \frac{e^{-\tau_n} (\tau_n)^{z_n}}{z_n!} \right) \frac{e^{-\tau_1^{(t)}} (\tau_1^{(t)})^{z_1}}{z_1!} \\
&= \sum_{n=1}^N [-\beta\tau_n + x_n(\log \beta + \log \tau_n) - \log x_n!] + \sum_{n=2}^N [-\tau_n + z_n \log \tau_n - \log z_n!] \\
&\quad + \sum_{z_1=0}^{\infty} [-\tau_1 + z_1 \log \tau_1 - \log z_1!] \frac{e^{-\tau_1^{(t)}} (\tau_1^{(t)})^{z_1}}{z_1!} \\
&= \left(\sum_{n=1}^N [-\beta\tau_n + x_n(\log \beta + \log \tau_n)] + \sum_{n=2}^N [\tau_n z_n \log \tau_n] \right. \\
&\quad \left. + \sum_{z_1=0}^{\infty} [-\tau_1 + z_1 \log \tau_1] \frac{e^{-\tau_1^{(t)}} (\tau_1^{(t)})^{z_1}}{z_1!} \right) \\
&\quad - \left(\sum_{n=1}^N \log x_n! + \sum_{n=2}^N \log z_n! + \sum_{z_1=0}^{\infty} [\log z_1!] \frac{e^{-\tau_1^{(t)}} (\tau_1^{(t)})^{z_1}}{z_1!} \right)
\end{aligned} \tag{2.67}$$

where in the last equality we have grouped together terms involving β and τ_n and terms that do not involve these parameters. To maximize w.r.t β and τ_n , we have to consider only the terms in the first parenthesis. Next we note that $-\tau_1 + \log \tau_1 \sum_{z_1=0}^{\infty} z_1 \frac{e^{-\tau_1^{(t)}} (\tau_1^{(t)})^{z_1}}{z_1!} = -\tau_1 + \tau_1^{(t)} \log \tau_1$ since $\sum_{z_1=0}^{\infty} z_1 \frac{e^{-\tau_1^{(t)}} (\tau_1^{(t)})^{z_1}}{z_1!}$ is the mean for the Poisson distribution for Z_1 . When we substitute this into Equ. 2.67 we see that the *expected* complete-data likelihood is the same as the original complete-data likelihood with z_1 replaced by $\tau_1^{(t)}$. Thus, following Eqs. 2.55, in the t^{th} step, the ML estimates are given by:

$$\begin{aligned}
\hat{\beta}^{(t+1)} &= \frac{\sum_{n=1}^N x_n}{\tau_1^{(t)} + \sum_{n=2}^N z_n}, & \hat{\tau}_1^{(t+1)} &= \frac{\hat{\tau}_1^{(t)} + x_1}{1 + \hat{\beta}^{(t+1)}} \\
\hat{\tau}_j^{(t+1)} &= \frac{z_j + x_j}{1 + \hat{\beta}^{(t+1)}} & \forall j &= 2, 3, \dots, N
\end{aligned} \tag{2.68}$$

The properties of the EM algorithm assures us that the sequence $(\hat{\beta}^{(t)}, \hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, \dots, \hat{\tau}_N^{(t)})$ converges to the incomplete-data MLE as $t \rightarrow \infty$.

We next state an important theorem that emphasizes why an algorithm like EM is useful in the course of topical analysis of unstructured data.

De Finetti's theorem: Let \mathbf{X} be observed variables and further assume that the variables are exchangeable. By DeFinetti's theorem, there should be an underlying parameter θ that gives rise to the observations. This means that the joint distribution of an infinitely exchangeable sequence of random variables is as if a random parameter were drawn from some distribution and then the random variables in question are independent and identically distributed, conditioned on that parameter. For a proof of the theorem see [Feller, 1968]. The elegance of this theorem is that it justifies the principle for hierarchical Bayesian modeling as well as the use of the EM machinery to find a local optimum for the hidden state space variables.

De Finetti's theorem actually explains how a basic topic model like LDA [Blei et al., 2003] has a representation that intuitively gives rise to the latent space or topic within the document collection. In LDA, we assume that words are generated by topics (fixed conditional distributions) and that those topics

are infinitely exchangeable within a document. By De Finetti’s theorem, the probability of a sequence of words and topics must therefore have the form:

$$p(\mathbf{z}_d, \mathbf{w}_d) = \int p(\boldsymbol{\theta}_d) \prod_{n=1}^{N_d} p(z_{d,n}, w_{d,n} | \boldsymbol{\theta}_d) d\boldsymbol{\theta}_d \quad (2.69)$$

where $\boldsymbol{\theta}_d$ is the random parameter of a multinomial over topics defined for each document d ; $w_{d,n}$ is the observed word in position n in document d and the latent (hidden) variable $z_{d,n}$ is the topic indicator for $w_{d,n}$.

Note: If an identically distributed sequence is independent, then the sequence is exchangeable; however, the converse is false—there exist exchangeable random variables that are statistically dependent, for example the Polya urn model which is just the opposite of sampling without replacement exhibits “the rich getting richer” phenomenon.

We next mention the basics of the techniques underlying parameter estimations of all models used in this thesis. The crux of these techniques revolve around the central theme that the original problem of finding the marginal likelihood of the data given the parameters by integrating out the uncertainties over the hidden variables (and parameters in case of priors) is intractable and thus a new functional is formulated which acts as a lower bound on the original problem. This introduces expectations of the complete data log likelihood which are to be taken under the distributions over the hidden variables and/or parameters (in the case of priors) and the estimates of the lower bound are iteratively refined using the EM algorithm and its variational variants [Beal, 2003, Wainwright and Jordan, 2008].

2.6.1 Finding a lower bound to the log likelihood

The log likelihood function which is just a statistic of the data can be written as:

$$\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) = \ln p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{n=1}^N \ln \int p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\theta}) d\mathbf{z}_n \quad (2.70)$$

where \mathbf{z}_n are the hidden variables corresponding to the \mathbf{x}_n s. Previously, in absence of hidden variables, we have sought a maximum likelihood setting of $\boldsymbol{\theta}$ such that

$$\boldsymbol{\theta}_{ML} \equiv \arg \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \quad (2.71)$$

In the case of incomplete data, the hidden variables often encode the latent state of the observations and give us an idea as to what factor may be responsible for generating the datum by inspecting other observations generated from the same factor. The problem with the hidden variables is that if there are many such variables, the integral (or sum) over those variables may be intractable—a classic example of which can be found in the seminal work of Blei and Jordan on probabilistic topic models [Blei et al., 2003]. This implies that computing the log partition function $A(\boldsymbol{\theta})$ is intractable.

To avoid the problem of intractability, we often optimize a lower bound on the likelihood function instead. The hallmark of variational methods in performing this kind of optimization is the consideration of a simpler dual representation such as treating hidden variables as marginally independent. These variables in the dual formulation are each endowed with their own distributions with “free” parameters (c.f. the distribution q in Equ. 2.72 with variational parameters ϕ) and the approximation to the actual

posterior probability of these variables are obtained through computing expectations over these “free” variational distributions subject to probability measure constraints. Any valid mean parameter specifies a lower bound on the log partition function.

Note that variational methods do not automatically induce any approximation, however, for a large class of practical models, exact computation of the marginal (i.e. the integral in Equ. 2.70) is not feasible in polynomial time. Finding the lower bound often exploits Jensen’s inequality for convex (and hence log concave) functions and is obtained in the following way:

$$\begin{aligned}
\mathcal{L}(\mathbf{X}|\theta) &= \sum_{n=1}^N \ln \int p(\mathbf{x}_n, \mathbf{z}_n|\theta) d\mathbf{z}_n \\
&= \sum_{n=1}^N \ln \int q(\mathbf{z}_n|\phi_n) \frac{p(\mathbf{x}_n, \mathbf{z}_n|\theta)}{q(\mathbf{z}_n|\phi_n)} d\mathbf{z}_n \\
&\geq \sum_{n=1}^N \int q(\mathbf{z}_n|\phi_n) \ln \frac{p(\mathbf{x}_n, \mathbf{z}_n|\theta)}{q(\mathbf{z}_n|\phi_n)} d\mathbf{z}_n \\
&= \sum_{n=1}^N \left\{ \int q(\mathbf{z}_n|\phi_n) \ln p(\mathbf{x}_n, \mathbf{z}_n|\theta) - \int q(\mathbf{z}_n|\phi_n) \ln q(\mathbf{z}_n|\phi_n) \right\} \\
&= \mathcal{L}(q_{\mathbf{z}}, \theta)
\end{aligned} \tag{2.72}$$

In all of our subsequent developments of newer topic models in this thesis, we will seek mathematical expressions for the lower bound of the form $\mathcal{L}(q_{\mathbf{z}}, \theta)$ or $\mathcal{L}(q_{\mathbf{z}}, q_{\theta}, \theta)$ and optimize it. The additional q_{θ} in \mathcal{L} is used when **priors** over parameters are imposed in a more general hierarchical setting.

A class of methods called Mean Field optimization (see Section 2.6.6) is used which “break” the original model structure and uses tractable free distributions over both the hidden variables and model parameters with priors. Breaking the original graphical model structure means simply to remove the edges between random variables that cause coupling under the D-separation criterion [Bishop, 2006, Shachter, 1998] due to head-to-head arrows on a set of observed variables (see Fig. 2.3).

The second term together with the negative sign in Equ. 2.72, is the entropy of the q distribution over the hidden variables \mathbf{z} . What the lower bound $\mathcal{L}(q_{\mathbf{z}}, \theta)$ then means is that the variational distribution q tries to balance the two competing goals: assign values to the hidden variables \mathbf{z} that have high probability under $p(\mathbf{z}, \mathbf{x})$ (the first term) and at the same time entertain a large number of distinct assignments (the entropy term). The implications of this for the case of LDA is mentioned in Section 2.7.1.

2.6.2 EM for *Exact* Unconstrained Optimization

The Expectation-Maximization (EM) algorithm [Dempster et al., 1977] alternates between an E step, which infers posterior distributions over hidden variables given a current parameter setting, and an M step, which maximizes the data log likelihood with respect to θ given the statistics gathered from the E step. Such a set of updates can be derived using the lower bound: at each iteration, the E step maximizes $F(q_{\mathbf{z}}, \theta)$ with respect to each of the $q_{z_n} \equiv q(z_n|\theta)$ distributions, and the M step does so with respect to θ . Using a superscript (t) to denote iteration number, starting from some initial parameters $\theta^{(0)}$, the update equations are:

$$\mathbf{E}\text{-Step} : \quad q_{z_n}^{(t+1)} \leftarrow \arg \max_{q_{z_n}} F(q_{\mathbf{z}}, \theta^{(t)}) \quad \forall n \in \{1, \dots, N\} \tag{2.73}$$

$$\mathbf{M}\text{-Step} : \quad q_{\boldsymbol{\theta}}^{(t+1)} \leftarrow \arg \max_{q_{\mathbf{z}_n}} F(q_{\mathbf{z}}^{(t+1)}, \boldsymbol{\theta}) \quad (2.74)$$

For the EM algorithm in the exact case, the maximum over $q_{\mathbf{z}_n}$ can be obtained by setting

$$q_{\mathbf{z}_n}^{(t+1)} = p(\mathbf{z}_n | \mathbf{x}_n, \boldsymbol{\theta}^{(t)}), \forall n \in \{1, \dots, N\} \quad (2.75)$$

at which point the bound becomes an inequality. Thus the exact case means that the distribution $q_{\mathbf{z}_n}(\mathbf{z}_n)$ over the hidden variables \mathbf{z}_n is the exact posterior distribution over \mathbf{z}_n given the model parameters $\boldsymbol{\theta}$ and the observations \mathbf{X} .

Proof. We have $\int q_{\mathbf{z}_n} d\mathbf{z}_n = 1, \forall i$. The constraints on $q_{\mathbf{z}_n}$ can be imposed through N Lagrange multipliers $\{\lambda_n\}_{n=1}^N$ forming the new functional:

$$\hat{\mathcal{L}}(q_{\mathbf{z}}, \boldsymbol{\theta}) = \mathcal{L}(q_{\mathbf{z}}, \boldsymbol{\theta}) + \sum_n \lambda_n \left(\int q_{\mathbf{z}_n} d\mathbf{z}_n - 1 \right) \quad (2.76)$$

Differentiating Equ. 2.76 w.r.t. $q_{\mathbf{z}_n}$, we have:

$$\begin{aligned} \frac{\partial}{\partial q_{\mathbf{z}_n}} \hat{\mathcal{L}}(q_{\mathbf{z}}, \boldsymbol{\theta}^{(t)}) &= \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}^{(t)}) - \ln q_{\mathbf{z}_n} - 1 + \lambda_n \\ \implies q_{\mathbf{z}_n}^{(t+1)} &= \exp(-1 + \lambda_n) p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}^{(t)}) \\ &= p(\mathbf{z}_n | \mathbf{x}_n, \boldsymbol{\theta}^{(t)}), \forall n \end{aligned} \quad (2.77)$$

where each λ_n is related to the normalization constant: $\lambda_n = 1 - \ln \int p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}^{(t)}) d\mathbf{z}_n, \forall n$ \square

The optimal parameters are obtained in the M-Step by setting the derivatives of Equ. 2.78 w.r.t $\boldsymbol{\theta}$ to zero.

$$\mathbf{M}\text{ Step}: \quad \boldsymbol{\theta}^{(t+1)} \leftarrow \arg \max_{\boldsymbol{\theta}} \sum_n \int p(\mathbf{z}_n | \mathbf{x}_n, \boldsymbol{\theta}^{(t)}) \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}) d\mathbf{z}_n \quad (2.78)$$

Note that the optimization is over the second $\boldsymbol{\theta}$ in the integrand while holding $p(\mathbf{z}_n | \mathbf{x}_n, \boldsymbol{\theta}^{(t)})$ fixed. It is interesting to note that the Lagrange multipliers which are ascribed to the nodes \mathbf{z}_n (see Equ. 2.76) are nothing but messages passed to them in order to evaluate the corresponding local marginal distributions in a Belief Propagation framework [Zeng et al., 2011, Wainwright and Jordan, 2008].

Essentially $\mathcal{L}(q_{\mathbf{z}}, \boldsymbol{\theta})$ acts as a functional which lower bounds $\mathcal{L}(\boldsymbol{\theta})$ for any $q(\mathbf{z} | \boldsymbol{\theta})$, attaining equality after each E step. Here we have expressed the E step as obtaining the **full distribution** over the hidden variables for each data point. However, in general, the M step may require only a few statistics of the hidden variables and so only these need be computed in the E step. This is the case for topic models as well.

2.6.3 EM for *Approximate* Constrained Optimization

In many real life modeling scenarios and datasets, interaction between multiple hidden variables need to be explained which can result in intractable posterior distributions. Such situations easily arise in cases where the graph is partitioned into cliques. In the variational approach we can constrain the posterior distributions to be of a particular tractable form, for example factorized over the variables $\mathbf{z}_n = \{\mathbf{z}_{n,j}\}_{j=1}^{|\mathbf{z}_n|}$ where $|\mathbf{z}_n|$ is the number of variables in a partition of the graph to which \mathbf{z}_n belongs. Here we have assumed N such partitions. This notation is a more general case to handle structured partitioning.

The size of the set depends on the size of the partition and in most cases, it is a singleton with $\{\mathbf{z}_{n,j}\} = \mathbf{z}_n$. Using calculus of variations we can still optimize the functional $\mathcal{L}(q_{\mathbf{z}}, \boldsymbol{\theta})$ as a functional of constrained distributions $q_{\mathbf{z}_n}$ which are n independent distributions over each of the \mathbf{z}_n s. The M step, which optimizes $\boldsymbol{\theta}$, is conceptually identical to that described in the previous section, except that it is based on sufficient statistics calculated with respect to the **constrained posterior** $q_{\mathbf{z}_n} \equiv q(\mathbf{z}_n|\phi_n)$ for some variational parameter ϕ_n for each n *instead of the exact posterior* $q(\mathbf{z}_n|\boldsymbol{\theta})$. We can write the lower bound for the exact posterior $\mathcal{L}(q_{\mathbf{z}}, \boldsymbol{\theta})$ over $\boldsymbol{\theta}$ as

$$\mathcal{L}(q_{\mathbf{z}}|\boldsymbol{\theta}) = \sum_n \int q(\mathbf{z}_n|\boldsymbol{\theta}) \ln \frac{p(\mathbf{z}_n, \mathbf{x}_n|\boldsymbol{\theta})}{q(\mathbf{z}_n|\boldsymbol{\theta})} d\mathbf{z}_n \quad (2.79)$$

$$= \sum_n \int q(\mathbf{z}_n|\boldsymbol{\theta}) \ln p(\mathbf{x}_n|\boldsymbol{\theta}) d\mathbf{z}_n + \sum_n \int q(\mathbf{z}_n|\boldsymbol{\theta}) \ln \frac{p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta})}{q(\mathbf{z}_n|\boldsymbol{\theta})} d\mathbf{z}_n \quad (2.80)$$

$$= \sum_n \ln p(\mathbf{x}_n|\boldsymbol{\theta}) - \sum_n \int q(\mathbf{z}_n|\boldsymbol{\theta}) \ln \frac{q(\mathbf{z}_n|\boldsymbol{\theta})}{p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta})} d\mathbf{z}_n \quad (2.81)$$

Thus in the **E step**, maximizing $\mathcal{L}(q_{\mathbf{z}}|\boldsymbol{\theta})$ w.r.t. $q_{\mathbf{z}_n}(\mathbf{z}_n)$ is equivalent to minimizing

$$\int q_{\mathbf{z}_n}(\mathbf{z}_n) \ln \frac{q_{\mathbf{z}_n}(\mathbf{z}_n)}{p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta})} d\mathbf{z}_n \equiv KL[q_{\mathbf{z}_n}||p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta})] \geq 0 \quad (2.82)$$

which is the Kullback-Leibler divergence between the variational distribution $q_{\mathbf{z}_n}(\mathbf{z}_n)$ and the exact posterior $p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\theta})$ over the hidden variables. The E step does not generally result in the bound becoming an equality, unless of course the exact posterior lies in the family of constrained posteriors $q(\mathbf{z}|\phi)$ [Beal, 2003].

KL is an asymmetric measure of the difference between two probability distributions P and Q . Specifically, the Kullback-Leibler divergence of Q from P is a measure of the information lost when Q is used to approximate P : KL measures the expected number of extra bits required to code samples from P when using a code based on Q , rather than using a code based on P . Typically P represents the “true” distribution of data, observations, or a precisely calculated theoretical distribution. The measure Q typically represents a theory, model, description, or approximation of P .

The KL divergence in the variational Bayesian setting is minimized globally over all terms in the approximation. The consequence of this fact is very clearly elucidated in [Bishop, 2006] where the factorized variational approximation tends to approximate the posterior by distributions which are too compact. Suppose we have a bi-modal mixture of Gaussians with $p(\mathbf{Z})$ following a bi-modal distribution. We try to fit a single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$. Using the variational approximation framework to minimize $KL(Q||P)$ the mode of the unimodal Q distribution is identified with one of the modes of the P distribution. Naive minimization of $KL(P||Q)$ on the other hand will tend to average across all the modes leading to very poor predictive distributions. These two forms of divergences belong to the so called α family of divergences defined by $D_\alpha(P||Q) = \frac{4}{1-\alpha^2} (1 - \int p(z)^{(1+\alpha)/2} q(z)^{(1-\alpha)/2} dz)$ where $-\infty < \alpha < \infty$ and $\alpha \in R$. The divergence $KL(Q||P)$ corresponds to the limit $\alpha \rightarrow -1$. For $\alpha \leq -1$, $D_\alpha(P||Q)$ is “zero forcing” which means that $q(z)$ will seek modes of $p(z)$ which have the largest mass. In practical applications, true posterior distributions are multi-modal with most of the posterior probability mass (or density) concentrated in some number of small regions in parameter space.

The M step in this approximate case is based on the current variational posterior over hidden vari-

ables:

$$\mathbf{M} \text{ Step: } \boldsymbol{\theta}^{(t+1)} \leftarrow \arg \max_{\boldsymbol{\theta}} \sum_n \int q_{\mathbf{z}_n}^{(t+1)}(\mathbf{z}_n) \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}) d\mathbf{z}_n \quad (2.83)$$

As before, one can choose $q_{\mathbf{z}_n}(\mathbf{z}_n)$ to be in a particular parametrized family:

$$q_{\mathbf{z}_n} = q(\mathbf{z}_n | \boldsymbol{\lambda}_n) \quad (2.84)$$

where $\boldsymbol{\lambda}_n = \{\lambda_{n_1}, \dots, \lambda_{n_K}\}$ are K variational parameters for each observation. If we constrain each $q_{\mathbf{z}_n}(\mathbf{z}_n | \boldsymbol{\lambda}_n)$ to have easily computable moments (e.g. a Multinomial or a Gaussian), and especially if $\ln p(\mathbf{z}_n | \mathbf{x}_n, \boldsymbol{\theta})$ is polynomial in \mathbf{z}_n , then we can compute the KL divergence up to a constant and can take its derivatives with respect to the set of variational parameters $\boldsymbol{\lambda}_n$ of each $q_{\mathbf{z}_n}$ distribution to perform the constrained E step.

The E step of the variational EM algorithm therefore consists of an **inner-loop** in which each of the $q(\mathbf{z}_n | \boldsymbol{\lambda}_n)$ is optimized by taking derivatives with respect to each $\lambda_{n,k}$, for $k = 1, \dots, K$.

2.6.4 EM for Maximum-A-Posteriori Learning and its Connection with VBEM

In Maximum-A-Posteriori (MAP) learning the parameter optimization includes prior information about the parameters of $p(\boldsymbol{\theta})$ and the M step seeks to find

$$\boldsymbol{\theta}_{MAP} \equiv \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) p(\mathbf{x} | \boldsymbol{\theta}) \quad (2.85)$$

given this prior. In the case of an exact E step, the M step is simply augmented to:

$$\mathbf{M} \text{ Step: } \boldsymbol{\theta}^{(t+1)} \leftarrow \arg \max_{\boldsymbol{\theta}} [\ln p(\boldsymbol{\theta}) + \sum_n \int p(\mathbf{z}_n | \mathbf{y}_n, \boldsymbol{\theta}^{(t)}) \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}) d\mathbf{z}_n] \quad (2.86)$$

In the case of a constrained approximate E step, the M step is given by

$$\mathbf{M} \text{ Step: } \boldsymbol{\theta}^{(t+1)} \leftarrow \arg \max_{\boldsymbol{\theta}} [\ln p(\boldsymbol{\theta}) + \sum_n \int q(\mathbf{z}_n)^{(t+1)} \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}) d\mathbf{z}_n] \quad (2.87)$$

Examples of this form of learning is presented in Chapter 4 in the context of parameter regularization within the Tag²LDA class of models and later in Chapter 6 in the context of regularizing Gaussian parameters with conjugate priors.

The same variational treatment can also be used to approximate the integrals required for Bayesian learning involving priors over parameters. The basic idea is to approximate the distribution over both hidden variables and parameters with simpler distributions, usually one which assumes that the hidden states and parameters are independent given the data.

There are two main goals in approximate Bayesian learning. The first is approximating the marginal likelihood of the data $p(\mathcal{D} | m) = \prod_{n=1}^N p(\mathbf{x}_n | \boldsymbol{\theta}_m)$ in order to perform model comparison over a family of models $m \in \mathbb{M}$ where a model m is nothing but a set of parameters $\boldsymbol{\theta}_m$ usually represented as probability distributions. The second is approximating the posterior distribution over the parameters of a model $p(\boldsymbol{\theta} | \mathbf{X}, m)$, which can then be used for prediction. This Bayesian prediction is a weighted average of the individual predictions, with weights proportional to the posterior probability of each model.

2.6.5 EM for Bayesian Learning using Variational distributions

In the Bayesian learning of models with hidden variables with variational distributions, we approximate the true posterior over the hidden variables by distributions from some tractable family such that the computation of the lower bound to the marginal log likelihood of the data remains feasible. Consider a mixture model with parameters θ with some prior over them as $p(\theta)$ giving rise to observations \mathbf{x} with the latent components of the mixture being identified by the hidden (indicator) variables \mathbf{z} . A lower bound on the model log marginal likelihood can be written as:

$$\ln p(\mathbf{x}|\theta) \geq \mathcal{L}(q(\mathbf{z}), q(\theta)) \equiv \int q(\mathbf{z})q(\theta) \ln \frac{p(\mathbf{z}, \mathbf{x}|\theta)p(\theta)}{q(\mathbf{z})q(\theta)} d\theta \quad (2.88)$$

The objective in Equ. 2.88 can be optimized using iterative techniques such as the fixed point iteration scheme where we find the optimal settings over the approximate posteriors in the VBE step and then using those posteriors as constant optimize the parameters in the VBM step during the $(t)^{th}$ iteration. We thus have the following updates for the VBE and VBM steps:

VBE Step:

$$q^{(t+1)}(\mathbf{z}) = \frac{1}{\mathcal{Z}(\mathbf{z})} \exp \left[\int q^{(t)}(\theta) \ln p(\mathbf{z}, \mathbf{x}|\theta) d\theta \right] \quad (2.89)$$

where $\mathcal{Z}(\mathbf{z})$ is the normalizer to make $q^{(t+1)}(\mathbf{z})$ a probability distribution and is defined over hidden variables \mathbf{z} only. Further $q^{(t+1)}(\mathbf{z}) = \prod_{n=1}^N q^{(t+1)}(\mathbf{z}_n)$. The proof is as follows:

$$\begin{aligned} \frac{\partial}{\partial q(\mathbf{z})} \mathcal{L}(q(\mathbf{z}), q(\theta)) &= \int q(\theta) \left[\frac{\partial}{\partial q(\mathbf{z})} \int q(\mathbf{z}) \ln \frac{p(\mathbf{z}, \mathbf{x}|\theta)}{q(\mathbf{z})q(\theta)} d\mathbf{z} \right] d\theta \\ &= \int q(\theta) [p(\mathbf{z}, \mathbf{x}|\theta) - \ln q(\mathbf{z}) - 1] d\theta \\ &= 0 \end{aligned} \quad (2.90)$$

This implies that:

$$\ln q^{(t+1)}(\mathbf{z}) = \int q^{(t)}(\theta) p(\mathbf{z}, \mathbf{x}|\theta) d\theta - \ln \mathcal{Z}^{(t+1)}(\mathbf{z}) \quad (2.91)$$

where $\mathcal{Z}^{(t+1)}(\mathbf{z})$ is the normalizer to enforce that $q^{(t+1)}(\mathbf{z})$ is a probability distribution using constraints on the free q distributions with Lagrange Multipliers. The update in Equ. 2.91 is true for every n^{th} data point whence $\ln q^{(t+1)}(\mathbf{z}) = \sum_{n=1}^N \ln q^{(t+1)}(\mathbf{z}_n)$ and $\mathcal{Z}^{(t+1)}(\mathbf{z}) = \prod_{n=1}^N \mathcal{Z}^{(t+1)}(\mathbf{z}_n)$ holds true assuming i.i.d. property of the random variables.

Thus there is a unique stationary point for each $q(\mathbf{z}_n)$ for a given $q_\theta(\theta)$. We now turn to the expression for parameter updates in the VBM step.

VBM Step:

$$q^{(t+1)}(\theta) = \frac{1}{\mathcal{Z}(\theta)} \exp \left[\int q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{z}, \mathbf{x}|\theta) d\mathbf{z} \right] p(\theta) \quad (2.92)$$

The proof is as follows:

$$\begin{aligned}
\frac{\partial}{\partial q(\boldsymbol{\theta})} \mathcal{L}(q(\mathbf{z}), q(\boldsymbol{\theta})) &= \frac{\partial}{\partial q(\boldsymbol{\theta})} \int q(\boldsymbol{\theta}) \left[\int q(\mathbf{z}) \ln p(\mathbf{z}, \mathbf{x}|\boldsymbol{\theta}) d\mathbf{z} + \ln p(\boldsymbol{\theta}) - \ln q(\boldsymbol{\theta}) \right] d\boldsymbol{\theta} \\
&= \int q(\mathbf{z}) \ln p(\mathbf{z}, \mathbf{x}|\boldsymbol{\theta}) d\mathbf{z} + \ln p(\boldsymbol{\theta}) - q(\boldsymbol{\theta}) + c \\
&= 0
\end{aligned} \tag{2.93}$$

where c is a constant independent of $\boldsymbol{\theta}$. Setting the derivative of Equ. 2.93 to zero, we have:

$$\ln q^{(t+1)}(\boldsymbol{\theta}) = \ln p(\boldsymbol{\theta}) + \int q^{(t+1)}(\mathbf{z}) \ln p(\mathbf{z}, \mathbf{x}|\boldsymbol{\theta}) d\mathbf{z} - \ln \mathcal{Z}^{(t+1)}(\boldsymbol{\theta}) \tag{2.94}$$

where $\mathcal{Z}^{(t+1)}(\boldsymbol{\theta})$ is the normalisation constant attributed to the Lagrange multipliers which have been omitted in the derivations above. Thus for a given $q(\mathbf{z})$, there is a unique stationary point for $q(\boldsymbol{\theta})$.

2.6.6 Mean Parameters

In Section 2.1, we have seen that any exponential family member $p_\theta \equiv p(\mathbf{X}|\theta)$ can be represented by its vector of canonical parameters $\theta \in \Theta$. Additionally, any exponential family has an alternative parameterization in terms of a vector of *mean parameters*. Moreover, statistical computations such as marginalization and maximum likelihood estimation, can be understood as transforming from one parameterization to the other [Wainwright and Jordan, 2008].

The mean parameter $\boldsymbol{\mu}_\alpha$ **associated** with a sufficient statistic Υ_α corresponding to a **given density** p_θ and an index set \mathbb{I} , is defined by the expectation:

$$\boldsymbol{\mu}_\alpha = E_p[\Upsilon_\alpha(\mathbf{X})] = \int \Upsilon_\alpha(\mathbf{X}) p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{X}, \forall \alpha \in \mathbb{I}. \tag{2.95}$$

We thus define a vector of mean parameters $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_d)$, one for each of the $|\mathbb{I}| = d$ sufficient statistics Υ_α , with respect to an arbitrary density $p(\mathbf{X}|\boldsymbol{\theta})$. Formally, the set defined as:

$$\mathcal{M} = \{\boldsymbol{\mu} \in \tilde{\mathbb{R}} \mid \exists p \text{ s.t. } E_p[\Upsilon_\alpha(\mathbf{X})] = \boldsymbol{\mu}_\alpha, \forall \alpha \in \mathbb{I}\} \tag{2.96}$$

corresponds to *all* realizable mean parameters and the dimensionality of $\tilde{\mathbb{R}}$ depends on the dimensionality of the observations \mathbf{X} and the moments computed w.r.t. it. This definition does not restrict the density p to be associated with the exponential family corresponding to the sufficient statistics Υ . Equation 2.95 says that there is a single $\boldsymbol{\mu}_\alpha$ corresponding to a given p and Equ. 2.96 says that as p is varied, we obtain a set of $\boldsymbol{\mu}_\alpha$ s.

Example 2.6.2. *Using the canonical parameterization, for a Gaussian Markov Random field over observations with dimensionality B , i.e. $\mathbf{X} \in \mathbb{R}^B$, the mean parameters are the second-order moment matrix $\Sigma = E[\mathbf{X}\mathbf{X}^T] \in \mathbb{R}^{B \times B}$, and the mean vector $\boldsymbol{\mu} = E[\mathbf{X}] \in \mathbb{R}^B$. The mean parameter set in this case is $\mathcal{M} = \{(\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^B \times S_+^B \mid \Sigma - \boldsymbol{\mu}\boldsymbol{\mu}^T \succeq 0\}$ where S_+^B denotes the set of $B \times B$ symmetric positive semidefinite matrices.*

Example 2.6.3. *In the case of **discrete random variables**, the set \mathcal{M} is convex. Particularly, for any random vector $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ such that the associated state space \mathcal{X} is finite, we have the following representation for the mean parameter set \mathcal{M} given that the underlying probability distribution belongs*

to the exponential family:

$$\begin{aligned} \mathcal{M} &= \left\{ \boldsymbol{\mu} \in \tilde{\mathbb{R}} \mid \boldsymbol{\mu} = \sum_{\mathbf{X} \in \mathcal{X}} \boldsymbol{\Upsilon}(\mathbf{X}) p(\mathbf{X}|\boldsymbol{\theta}) \quad \text{for some } p(\mathbf{X}|\boldsymbol{\theta}) > 0 \text{ and } \sum_{\mathbf{X} \in \mathcal{X}} p(\mathbf{X}|\boldsymbol{\theta}) = 1 \right\} \\ &= \text{conv} \{ \boldsymbol{\Upsilon}(\mathbf{X}), \mathbf{X} \in \mathcal{X} \} \end{aligned} \quad (2.97)$$

where $\text{conv}(S)$ is the convex hull of a set S which is the smallest set that contains all its convex combinations and by definition $\text{conv}(S)$ is a convex set. When the state space \mathcal{X}^N is finite then \mathcal{M} is called the **convex polytope** [Wainwright and Jordan, 2008, Sontag and Jaakkola, 2007]. and the geometric representation of a convex polytope is that of a convex polyhedron. Additionally an element of a convex set is an extreme point if it cannot be expressed as a convex combination of two distinct elements of the set. If we consider a graph with just three variables x_1, x_2 and x_3 , each independently drawn from *Bernoulli*(0.5), i.e. $p(x_1, x_2, x_3) = \prod_{n=1}^3 \theta^{x_n} (1 - \theta)^{1-x_n}$ with $\theta = 0.5$, then $\boldsymbol{\Upsilon}(\mathbf{x}) = \{x_1, x_2, x_3\}$ and $\mathcal{M} = \text{conv}\{(0, 0, 0), (0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0), (1, 1, 1)\}$ i.e. the cube $[0, 1]^3$. In this case, $\boldsymbol{\mu}(\boldsymbol{\theta}) = (1/2, 1/2, 1/2)$ is obtained by a forward mapping (see Section 2.6.8) from the canonical parameters, $\boldsymbol{\theta}$, of the distribution over \mathbf{X} and $\boldsymbol{\Upsilon}(\mathbf{x})$ consists of all the extreme points where the joint probability distribution puts its most mass on. The mean parameter of the joint probability distribution, $\boldsymbol{\mu}(\boldsymbol{\theta})$ in this case, can also be visualized as the point of intersection of the normals to the faces of the unit cube pointing inwards from the mid point of each of the six faces. A simple but more detailed illustration is provided at the end of Section 2.6.9.

Intuition behind Equ. 2.97: Our goal is to understand the phenomenon behind the observations. However, neither we know what the actual form of $\boldsymbol{\theta}$ is nor we know the exact extremum for such a function—we only observe samples \mathbf{X} drawn from the distribution of \mathbf{X} albeit with inherent noise. We assume a generative process for this \mathbf{X} which gives rise to a particular graphical model structure with underlying assumptions and parameters. If we had known the true causal distribution with parameter, say $\boldsymbol{\theta}^*$, then using Equ. 2.97 we compute a convex hull of $\boldsymbol{\Upsilon}(\mathbf{X})$ by taking each possible configuration of the discrete random variable \mathbf{X} . The extreme points $\boldsymbol{\mu}_e$ gives rise to the intersection of the boundaries of this convex hull and the problem thus becomes an optimization problem over the space bounded by this convex hull such that $\boldsymbol{\theta}^* = \hat{\boldsymbol{\mu}}$ where $\hat{\boldsymbol{\mu}} \in \text{conv}\{\boldsymbol{\Upsilon}(\mathbf{X})\}$.

We thus have the following at our disposal:

- It is easy to obtain sufficient statistics $\boldsymbol{\Upsilon}(\mathbf{X})$ if $p(\mathbf{X}|\boldsymbol{\theta})$ is in exponential family by using Factorization theorem (Theorem 2.2.1).
- If the posterior distributions over the hidden variables (and parameters) also belong to exponential family then expectations of these sufficient statistics w.r.t the posterior can be computed (Theorem 2.1.1).
- The use of exponential family distributions is validated by Maximum Entropy principle (Section 2.1).

These operations lead us to obtain an understanding of population parameter $\boldsymbol{\theta}^*$ through optimizing over the set of \mathcal{M} in a way which is computationally tractable.

2.6.7 Significance of Mean Parameters on Inference Problems

A fundamental class of inference problems in exponential family models is the computation of the forward mapping: the mapping from the canonical parameters $\boldsymbol{\theta} \in \Theta$ to the mean parameters $\boldsymbol{\mu} \in \mathcal{M}$. The backward mapping from mean parameters to canonical parameters also is very significant. In particular, suppose that we are given a set of samples $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, drawn independently from an exponential

family member $p(\mathbf{X}|\boldsymbol{\theta})$, where the parameter $\boldsymbol{\theta}$ is unknown. The principle of maximum likelihood dictates that to obtain the estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, one needs to maximize the likelihood function of the data given by $\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) = \prod_{n=1}^N p(\mathbf{x}_n|\boldsymbol{\theta})$ in its logarithmic and rescaled version as:

$$\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}_n|\boldsymbol{\theta}) = \hat{\boldsymbol{\mu}}^T \boldsymbol{\theta} - A(\boldsymbol{\theta}) \quad (2.98)$$

where $\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^N E[\boldsymbol{\Upsilon}(\mathbf{x}_n)]$ is the vector of **empirical mean parameters** defined by the data \mathbf{X} . The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is chosen to achieve the maximum of this objective function. Computing $\hat{\boldsymbol{\theta}}$ is a challenging problem since the **objective function involves** the log partition function A which is the normalizer used to obtain valid probability distributions and computed through marginalization over observations.

Among several techniques, the EM algorithm is a practical tool to find maximum likelihood estimates of estimators involving incomplete data i.e. observations with hidden state space variables. Under suitable conditions, the maximum likelihood estimate is *unique*, and specified by the stationarity condition $E_{p(\mathbf{X}|\hat{\boldsymbol{\theta}})}[\boldsymbol{\Upsilon}(\mathbf{X})] = \hat{\boldsymbol{\mu}}$. Finding the unique solution to this equation is equivalent to computing the backward mapping $\boldsymbol{\mu} \in \mathcal{M} \rightarrow \boldsymbol{\theta} \in \Theta$: from mean parameters to canonical parameters.

In general, computing this inverse mapping is also computationally intensive particularly while inferring states on datasets with missing values.

Properties of $A(\boldsymbol{\theta})$: The most important property of A is its convexity for exponential family distributions. Under suitable conditions, the derivatives of the function A and its conjugate dual A^* define a one-to-one and surjective mapping between the canonical and mean parameters [Wainwright and Jordan, 2008]. The conjugate dual function A^* of the function A is defined as:

$$A^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \Theta} \left\{ \boldsymbol{\mu}^T \boldsymbol{\theta} - A(\boldsymbol{\theta}) \right\} \quad (2.99)$$

Here $\boldsymbol{\mu}$ is a fixed vector of *dual* variables of the same dimension as $\boldsymbol{\theta}$ that are computable from the dataset at hand. The conjugate function of any function $f(\mathbf{x})$ is the function $f^*(\mathbf{y})$ defined as $f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathbf{X}} \{ \mathbf{y}^T \mathbf{x} - f(\mathbf{x}) \}$ where, the supremum can be obtained by maximizing $\mathbf{y}^T \mathbf{x} - f(\mathbf{x})$ over \mathbf{X} .

2.6.8 What does Forward Mapping of Canonical to Mean Parameters mean?

Forward mapping of the canonical parameters $\boldsymbol{\theta} \in \Theta$, which define a distribution $p(\boldsymbol{\theta})$, to the mean parameters $\boldsymbol{\mu}$ essentially tries to determine that for which mean parameter vectors $\boldsymbol{\mu} \in \mathcal{M}$ do there exist a vector $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\mu}) \in \Theta$ such that $E_{p(\mathbf{X}|\boldsymbol{\theta})}[\boldsymbol{\Upsilon}(\mathbf{X})] = \boldsymbol{\mu}$.

The mapping is defined in terms of the log partition function $A(\boldsymbol{\theta})$ where $-\ln g(\boldsymbol{\theta}) = A(\boldsymbol{\theta})$ in which $g(\boldsymbol{\theta})$ is defined in Equ. 2.4 and that $\nabla A(\boldsymbol{\theta}) = E_{p(\mathbf{X}|\boldsymbol{\theta})}[\boldsymbol{\Upsilon}(\mathbf{X})]$. To answer this theoretical question, it asks the following two important questions:

- a) when does ∇A define a one-to-one mapping?
- b) when does the image of Θ under the mapping ∇A i.e. the set $\nabla A(\Theta)$, fully cover the set \mathcal{M} ?

where \mathcal{M} is defined as before in Equ. 2.96 in the following way: $\mathcal{M} = \{ \boldsymbol{\mu} \in \tilde{\mathbb{R}} \mid \exists p \text{ s. t. } E_{p(\mathbf{X}|\boldsymbol{\theta})}[\boldsymbol{\Upsilon}_\alpha(\mathbf{X})] = \boldsymbol{\mu}_\alpha, \forall \alpha \in \mathcal{I} \}$.

The answer to the first question depends on whether or not the exponential family is minimal [Wainwright and Jordan, 2008].

Minimal Exponential Family: A minimal exponential family is an exponential family of distributions where there is a unique parameter vector θ associated with each corresponding distribution i.e. the components of sufficient statistics ensemble, Υ_α , form a basis.

The answer to the second question which concerns itself with the image of $\nabla A(\Theta)$ is simply the interior \mathcal{M}^i of the constraint set of realizable mean parameters \mathcal{M} [Wainwright and Jordan, 2008]. This fact is quite significant in that it means that (disregarding boundary points) all mean parameters \mathcal{M} that are realizable by *some distribution* can be realized by *a member of the exponential family*. So we need only bother ourselves with distributions from the exponential family to be ascribed to the random variables in the graphical model or by artificially representing interactions between random variables through some exponential function and normalizing the interactions to induce a valid probability measure. Proposition 2.6.1 and Theorem 2.6.1 both of which are stated and proved in [Wainwright and Jordan, 2008] also reinforces the answer to the second question.

Proposition 2.6.1. *The gradient mapping $\nabla A : \Theta \rightarrow \mathcal{M}$ is one-to-one if and only if the exponential representation is minimal.*

Theorem 2.6.1. *In a minimal exponential family, the gradient map ∇A is onto the interior of \mathcal{M} , denoted by \mathcal{M}^i . Consequently, for each $\mu \in \mathcal{M}^i$, there exists some $\theta = \theta(\mu) \in \Theta$ such that $E_{p(\mathbf{X}|\theta)}[\Upsilon(\mathbf{x})] = \mu$.*

Theorem 2.6.1 is important since it guarantees that for minimal exponential families, each mean parameter $\mu \in \mathcal{M}^i$ is uniquely realized by some density $p(\mathbf{X}|\theta(\mu))$ in the exponential family. However, a typical exponential family $\{p(\mathbf{X}|\theta)|\theta \in \Theta\}$ describes only a strict subset of all possible densities. In this case, there must exist at least some other density p which is not a member of an exponential family that also realizes μ . However, what differentiates an exponential distribution $p(\mathbf{X}|\theta(\mu))$ is that, among the set of all distributions that realize μ , it has the maximum entropy.

2.6.9 Conjugate Duality

The notion of *conjugate dual* functions plays a very important role in the field of convex analysis [Boyd and Vandenberghe, 2004, Nocedal and Wright, 2006]. As mentioned in Equ. 2.99, the conjugate dual function $A^*(\mu)$ for the log partition function $A(\theta)$ expressed as a function of an extremum of $\theta \in \Theta$ is defined as follows:

$$A^*(\mu) = \sup_{\theta \in \Theta} \left\{ \mu^T \theta - A(\theta) \right\} \quad (2.100)$$

The dual variables μ has the same dimensionality as θ and has a natural interpretation in terms of mean parameters. For example, finding the parameters of a model by maximizing the log likelihood statistic is sensible only when the vector μ belongs to the set \mathcal{M} such as the vector of empirical moments $\hat{\mu} = \frac{1}{N} \sum_{n=1}^N \Upsilon(\mathbf{x}_n)$ induced by a dataset \mathbf{X} consisting of N sample points. Also the conjugate dual function is very closely related to Shannon's entropy as elucidated by Theorem 2.6.2 (see [Wainwright and Jordan, 2008] for proof).

Theorem 2.6.2.

a) *For any $\mu \in \mathcal{M}^i$, the interior of \mathcal{M} , denote by $\theta(\mu)$ the unique canonical parameter satisfying the dual matching condition $E_{p(\mathbf{X}|\theta(\mu))}[\Upsilon(\mathbf{X})] = \nabla A(\theta(\mu)) = \mu$. The conjugate dual function A^**

takes the form:

$$A^*(\boldsymbol{\mu}) = \begin{cases} -H(p(\mathbf{X}|\boldsymbol{\theta}(\boldsymbol{\mu}))) & \text{if } \boldsymbol{\mu} \in \mathcal{M}^i & \{\mathcal{M}^i = \text{interior}(\mathcal{M})\} \\ +\infty & \text{if } \boldsymbol{\mu} \notin \mathcal{M}^c & \{\mathcal{M}^c = \text{closure}(\mathcal{M})\} \end{cases} \quad (2.101)$$

For any boundary point $\boldsymbol{\mu} \in \mathcal{M}^c \setminus \mathcal{M}^i$, we have $A^*(\boldsymbol{\mu}) = \lim_{n \rightarrow +\infty} A^*(\boldsymbol{\mu}^{(n)})$ taken over any sequence $\{\boldsymbol{\mu}^{(n)}\} \in \mathcal{M}^i$ converging to $\boldsymbol{\mu}$.

b) In terms of this dual, the log partition function $A(\boldsymbol{\theta})$ has the variational representation in terms of an extremum of $\boldsymbol{\mu} \in \mathcal{M}$ as:

$$A(\boldsymbol{\theta}) = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} - A^*(\boldsymbol{\mu}) \right\} \quad (2.102)$$

c) For all $\boldsymbol{\theta} \in \Theta$, the supremum in Equ. 2.102 is attained uniquely at the vector $\boldsymbol{\mu} \in \mathcal{M}^i$ specified by the moment matching conditions

$$\boldsymbol{\mu} = \int_{\mathbf{X} \in \mathcal{X}} \Upsilon(\mathbf{X}) p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{X} = E_{p(\mathbf{X}|\boldsymbol{\theta})}[\Upsilon(\mathbf{X})] \quad (2.103)$$

The main result of Theorem 2.6.2 is that when $\boldsymbol{\mu} \in \mathcal{M}^i$, the interior of \mathcal{M} , the value of the dual function $A^*(\boldsymbol{\mu})$ is precisely the negative entropy of the exponential family distribution $p(\mathbf{X}|\boldsymbol{\theta}(\boldsymbol{\mu}))$, where $\boldsymbol{\theta}(\boldsymbol{\mu})$ is the unique vector of canonical parameters satisfying the relation $E_{p(\mathbf{X}|\boldsymbol{\theta}(\boldsymbol{\mu}))}[\Upsilon(\mathbf{X})] = \nabla A(\boldsymbol{\theta}(\boldsymbol{\mu})) = \boldsymbol{\mu}$.

The value of $-A^*(\boldsymbol{\mu})$ corresponds to the optimum of the maximum entropy problem, where $\boldsymbol{\mu}$ parameterizes the constraint set. The event $A^*(\boldsymbol{\mu}) = +\infty$ corresponds to infeasibility of the maximum entropy problem. Thus, the take home message is that it is sufficient to maximize over the set \mathcal{M} , as expressed in the variational representation Equ. 2.102. This fact implies that the structure of the set \mathcal{M} plays a critical role in determining the complexity of computing the log partition function.

The gradient mapping ∇A maps Θ in a one-to-one manner onto \mathcal{M}^i , whereas the inverse mapping from \mathcal{M}^i to Θ is given by the gradient ∇A^* of the dual function. This flow of mappings in between sets can be visualized as: $\boldsymbol{\mu} \rightarrow (\nabla A)^{-1}(\boldsymbol{\mu}) \rightarrow \boldsymbol{\theta}(\boldsymbol{\mu}) \rightarrow -H(p(\mathbf{X}|\boldsymbol{\theta}(\boldsymbol{\mu}))) \rightarrow A^*(\boldsymbol{\mu})$. In many models of interest, $A(\boldsymbol{\theta})$ is not feasible to compute because of the complexity of \mathcal{M} or the lack of any explicit form for $A^*(\boldsymbol{\mu})$. However, we can bound $A(\boldsymbol{\theta})$ using:

$$A(\boldsymbol{\theta}) \geq \left\{ \boldsymbol{\theta}^T \boldsymbol{\mu} - A^*(\boldsymbol{\mu}) \right\} \quad (2.104)$$

for any mean parameter $\boldsymbol{\mu} \in \mathcal{M}$. The tightness of this bound is measured by a Kullback-Leibler divergence expressed in terms of the dual representation of the parameters as:

$$\begin{aligned} KL(p(\mathbf{X}|\boldsymbol{\theta}(\boldsymbol{\mu}))||p(\mathbf{X}|\boldsymbol{\theta})) &= E_{p(\mathbf{X}|\boldsymbol{\theta}(\boldsymbol{\mu}))}[\log p(\mathbf{X}|\boldsymbol{\theta}(\boldsymbol{\mu})) - \log p(\mathbf{X}|\boldsymbol{\theta})] \\ &= \boldsymbol{\theta}(\boldsymbol{\mu})^T \boldsymbol{\mu} - A(\boldsymbol{\theta}(\boldsymbol{\mu})) - \boldsymbol{\theta}^T \boldsymbol{\mu} + A(\boldsymbol{\theta}) \\ &= A(\boldsymbol{\theta}) - \boldsymbol{\theta}^T \boldsymbol{\mu} + A^*(\boldsymbol{\mu}) \end{aligned} \quad (2.105)$$

Computing the dual value $A^*(\check{\boldsymbol{\mu}})$ at some point $\check{\boldsymbol{\mu}} \in \mathcal{M}^i$ requires computing the inverse mapping $(\nabla A)^{-1}(\check{\boldsymbol{\mu}})$. This is in itself a nontrivial problem, and then evaluating the entropy requires high-dimensional integration for general graphical models. These difficulties motivate the use of approx-

imations to \mathcal{M} and A^* . The Naive Mean Field procedure (see Section 2.6.11) induces the simplest form of approximations to restrict the structure of the constraint set \mathcal{M} so as to enforce tractability in computing moments. Figure 2.2 illustrates these ideas for a discrete three dimensional random variable $\mathbf{Z} = \{z_1, z_2, z_3\}$.

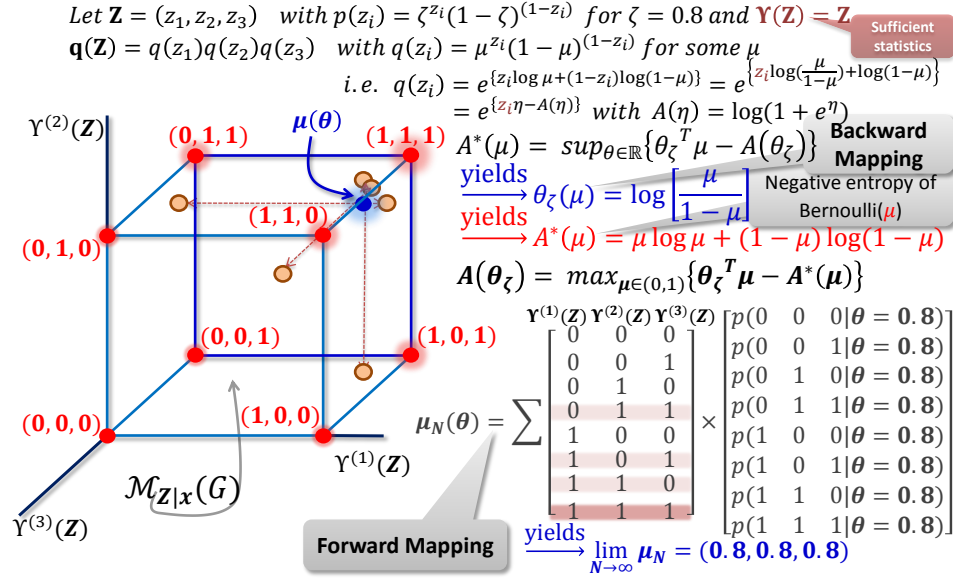


Figure 2.2: Simple illustration on forward mapping from canonical parameters to mean parameters and vice versa. We consider a distribution over the discrete random variable $\mathbf{Z} = \{z_1, z_2, z_3\}$ as a product of three independent Bernoulli distributions. The realizations of the binary random variables z_i form the corners of the marginal polytope (shown here without any constraint cutting planes which would typically arise out of the constraints on each of the $\mu_i \in [0, 1]$). Given a biased coin with probability of heads being 0.8, we are more likely to observe realizations of \mathbf{Z} which have more ones. The red shadow bubbles beneath each of the red nodes in the $[0, 1]$ cube on the left are indicative of this. The forward mapping is shown at the bottom right half of the illustration where the darker rows are indicative of more probable configurations of \mathbf{Z} . The value of the mean parameter μ will tend to the true mean with a value of $(0.8, 0.8, 0.8)$ of the generating distribution as we observe an infinite number of samples with more and more configurations of two or more ones. This problem of forward mapping to find the mean parameters from the observations generated from the distribution with canonical parameters is equivalent to the problem of finding $\theta_\zeta(\mu)$ through the backward mapping which in this illustrative case has a closed form solution. The Bernoulli nature of the $q(z_i)$ s is also verified by the form of the log partition function of μ which in this case is the negative of the entropy of the Bernoulli distribution with mean parameter μ . Note that if we observe all configurations of \mathbf{Z} only once then the forward mapping of μ yields $(0.768, 0.768, 0.768)$ for which the backward mapping causes $\theta_\zeta(\mu)$ to be greater than one. To avoid this possibility, Lagrange multipliers are used to constrain μ thereby making θ valid.

2.6.10 Mean Field and Tractable Families

Mean Field theory [Kadanoff, 2009, Parisi, 1988] in the context of physics and probability theory studies the behavior of large and complex stochastic models through a simpler model. Such models consider a large number of interacting variables. The effect of all the other variables on any given variable is approximated by a single averaged effect, thus reducing a many-body problem to a one-body problem.

The “mean” in “Mean Field” usually denotes some kind of averaging such as taking expectations under some probability distributions and the “field” refers to some interaction phenomenon between a group of entities (usually following Markov properties) from the perspective of physics. In general, a “field,” however, points to a somewhat more abstract concept in mathematics [Herstein, 1964].

Mean field theory allows us to impose a specific type of approximation to the exact variational principle as laid out in Equ. 2.102. As discussed in Section 3.7, there are two fundamental difficulties associated with the variational principle given by Equ. 2.99: the nature of the constraint set \mathcal{M} , and the lack of an explicit form for the dual function A^* . The mean field approach at its core lets us limit the optimization problem for finding θ to a subset of distributions for which both \mathcal{M} and A^* are relatively easy to characterize. Throughout this thesis, we refer to any such characterizations as “tractable.” The simplest choice is the family of product distributions, which gives rise to the *naive* mean field method.

Tractable families: Given a graphical model based on a graph G , mean field methods are based on the notion of a tractable subgraph, $\mathcal{M}_F(G)$, by which we mean a subgraph F of the graph G over which it is feasible to perform exact calculations. The simplest example of a tractable subgraph is the fully disconnected subgraph which contains all the vertices of G but none of the edges. When performing approximate inference for incomplete data problems involving hidden and observed variables, this subgraph reflects the causal dependencies of the “free” conjugate parametric distributions on the hidden variables with the notion of “free” referring to the fact that the only assumption being made is that of constructing a tractable subgraph (such as a product distribution of possible cliques or singleton nodes) while the exact parameterization of those distributions is not theoretically constrained in any way.

The exponential family defined by the sufficient statistic Υ and graph G is associated with the set $\mathcal{M}(G, \Upsilon)$ of all mean parameters realizable by any distribution, as previously defined in Equ. 2.96. For a given tractable subgraph F , mean field methods are based on optimizing over the subset of mean parameters which can be obtained by the subset of exponential family densities $\{p(\mathbf{X}|\theta), \theta \in \Theta(F)\}$ denoted by $\mathcal{M}_F(G, \Upsilon) = \{\mu \in \mathbb{R}^d | \mu = E_{p(\mathbf{X}|\theta)}[\Upsilon(\mathbf{X})] \text{ for some } \theta \in \Theta(F)\}$. The subgraph \mathcal{M}_F is thus an inner approximation to the set \mathcal{M} of realizable mean parameters [Wainwright and Jordan, 2008].

Figure 2.3 shows the factorization of the hidden variables in the original graph in Fig. 2.3a that is needed for approximate inference. Looking simply from the point of view of the factorization, one might need to approximate the posterior further than simply the hidden variable / parameter factorizations. One reason for this is that the parameter posterior may still be intractable despite the hidden variable/parameter factorization. We therefore need to assume some simpler space of parameter posteriors particularly those distributions with just a few sufficient statistics, such as the Multinomial, Gaussian or Dirichlet distributions.

A good variational approximation is the one that removes as few arcs as possible from the original graphical model representation (or the moralization of it) such that inference becomes tractable. Some edges may capture crucial dependencies between nodes and must be preserved, whereas other edges might induce a weak local correlation at the expense of a long-range correlation which can be ignored to first order. In general, the more edges we remove, the more we achieve tractability.

The advantage of the variational Bayesian procedure is that *any factorization of the posterior* yields a lower bound on the marginal likelihood. It is expected that the more complex the factorizations the more the compute time, however, more complex factorizations can also yield progressively tighter lower bounds.

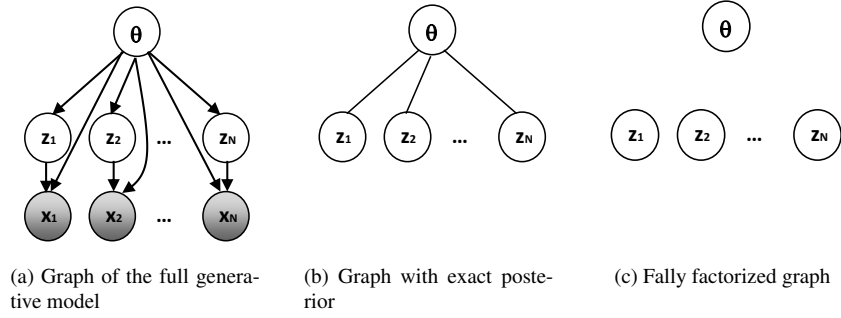


Figure 2.3: Graphical depiction of the hidden-variable / parameter factorization. Fig. 2.3a: The original generative model for N hidden and observed variables. Fig. 2.3b: The exact posterior graph for $p(\mathbf{Z}, \theta | \mathbf{X})$. The z_i and z_j pairs are not directly coupled, but interact through θ [Shachter, 1998]. By De Finetti’s theorem, the hidden variables are conditionally independent of one another, but only given the parameters. Fig. 2.3c: the posterior graph after the variational approximation between parameters and hidden variables where the arcs between parameters and hidden variables are removed. Due to this type of factorization the hidden variables become independent because of i.i.d assumption. A similar kind of “product” factorization is shown in Fig. 2.5

Finally, at this point it is worth repeating the comparison of EM for MAP estimation and variational Bayesian EM (see Section 2.6) from [Beal, 2003].

EM for MAP estimation of parameters	VBEM for models defined over parameters with priors
<p>Goal: Maximize $p(\theta \mathbf{X})$ w.r.t. θ</p> <p>E step: $q^{(t+1)}(\mathbf{Z}) = p(z \mathbf{X}, \theta^{(t)})$. Amounts to finding the <i>exact</i> posterior over the hidden variables given the data and the parameters of the model.</p> <p>M step: $\theta^{(t+1)} = \arg \max_{\theta} (\int q^{(t+1)}(\mathbf{Z}) \ln p(\mathbf{Z}, \mathbf{X} \theta) d\mathbf{Z} + \ln p(\theta))$. This amounts to finding point estimates for the parameters.</p>	<p>Goal: Bound from below $p(\mathbf{X} \theta)$</p> <p>E step: $q^{(t+1)}(\mathbf{Z}) = p(z \mathbf{X}, \int w(\theta^{(t)}) q^{(t)}(\theta) d\theta)$. This amounts to finding the <i>constrained</i> posterior (often with easily computable moments such as those in some tractable family) over the hidden variables given the data and the <i>expected</i> natural parameters with $w(\theta)$ being any well-behaved function of θ (see Theorem 2.1.1). Note that the expectation over $w(\theta)$ is performed only when the parameter θ is treated as a random variable with an appropriate prior.</p> <p>M step: $q^{(t+1)}(\theta) \propto \exp\{\int q^{(t+1)}(\mathbf{Z}) \ln p(\mathbf{Z}, \mathbf{X} \theta) d\mathbf{Z} + \ln p(\theta)\}$. This amounts to finding a distribution over parameters.</p>

Table 2.1: Key differences between EM and VBEM. The variable θ is the parameter of the model which we wish to find. The observations are denoted by \mathbf{X} and the hidden state variables are denoted by \mathbf{Z} .

The introduction of conjugate priors over parameters of an exponential family model (see Section 2.2.3) leads to a fuller Bayesian treatment of the underlying inference machinery. In the VBM step the functional form of the variational posterior $q(\boldsymbol{\theta})$ does not change during iterations of VBEM. The priors in a conjugate-exponential model allow us to treat the ensemble of parameters of $q(\boldsymbol{\theta})$ as random quantities each with its own uncertainty over the precision of its mean. This means that the VBM step in the conjugate-exponential setting replaces the many (possibly infinite) inference steps which compute the individual ML/MAP point estimates of each of the model parameters within the ensemble with a single step computing a weighted average.

2.6.11 Mean Field Procedure

The main problem of maximization over the set of realizable mean parameters is the structure of the set \mathcal{M} . The goal of Mean Field procedure is to induce a simpler structure on \mathcal{M} based on removal of certain edges in the original graphical model such that the number of configurations of the state space which needs to be explored while computing the log partition (i.e. cumulant) function becomes manageable (for e.g. from exponential to polynomial). Without such a procedure, we cannot impose a probability measure on the marginal log likelihood of the observations. Intuitively this means that in order to find the mean parameters of a model as best as possible, we are averaging the effect of the probabilistic interactions or correlations of all variables over an exponential state space w.r.t. the true probability distributions of the parameters with distributions that are Markov i.e. those distributions whose parameters are governed only by a very limited set of interactions between variables.

Generally, we are interested in approximating some target distribution $p(\mathbf{X}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$. Mean field methods generate lower bounds on the value $A(\boldsymbol{\theta})$ of the cumulant function, as well as approximations to the mean parameters $\boldsymbol{\mu} = E_{p(\mathbf{X}|\boldsymbol{\theta})}[\boldsymbol{\Upsilon}(\mathbf{X})]$ of this target distribution $p_{\boldsymbol{\theta}}$. The key property of any mean field method is the fact that *any valid mean parameter specifies a lower bound on the log partition function*.

Proposition 2.6.2. (Mean Field Lower Bound). *Any mean parameter $\boldsymbol{\mu} \in \mathcal{M}^i$ yields a lower bound on the cumulant function $A(\boldsymbol{\theta})$. Formally,*

$$A(\boldsymbol{\theta}) \geq \boldsymbol{\theta}^T \boldsymbol{\mu} - A^*(\boldsymbol{\mu}) \quad (2.106)$$

The equality holds if and only if $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ are dually coupled i.e., $\boldsymbol{\mu} = E_{p(\mathbf{X}|\boldsymbol{\theta})}[\boldsymbol{\Upsilon}(\mathbf{X})]$.

Proof. See [Wainwright and Jordan, 2008] □

Proposition 2.6.2 essentially means that the lower bound $A(\boldsymbol{\theta}) \geq \boldsymbol{\theta}^T \boldsymbol{\mu} + H(q)$ (with the entropy term $H(q) = -E_{q(\mathbf{X}|\boldsymbol{\theta})}[q(\mathbf{X}|\boldsymbol{\theta})]$) holds for any distribution $q(\mathbf{X}|\boldsymbol{\theta})$ satisfying the moment matching condition $E_q[\boldsymbol{\Upsilon}(\mathbf{X})] = \boldsymbol{\mu}$, the optimal q by Theorem 2.6.2 turns out to satisfy $q = p(\mathbf{X}|\boldsymbol{\theta}(\boldsymbol{\mu}))$ for which $H(q^*) = -A^*(\boldsymbol{\mu})$.

Due to the lack of an explicit form of the dual function A^* , it is intractable in general to compute the lower bound. The mean field approach overcomes this difficulty by restricting the choice of $\boldsymbol{\mu}$ to the tractable subset $\mathcal{M}_F(G)$, for which the dual function has an explicit form.

If $A_F^* = A^*|_{\mathcal{M}_F(G)}$ is the dual function restricted to the set $\mathcal{M}_F(G)$, then, provided that $\boldsymbol{\mu}$ belongs to $\mathcal{M}_F(G)$, the mean field method finds the best approximation, as measured in terms of the tightness

of the lower bound 2.106. More precisely, the best lower bound from within $\mathcal{M}_F(G)$ is given by

$$\max_{\boldsymbol{\mu} \in \mathcal{M}_F(G)} \boldsymbol{\mu}^T \boldsymbol{\theta} - A_F^*(\boldsymbol{\mu}) \quad (2.107)$$

The value of $\boldsymbol{\mu}$ is defined to be the mean field approximation of the true mean parameters. It is important to note that whether we consider the set $\mathcal{M}(G, \boldsymbol{\Upsilon}(\mathbf{X}))$ or $\mathcal{M}_F(G, \boldsymbol{\Upsilon}(\mathbf{X}))$ the extreme point $\boldsymbol{\mu}_e(\mathbf{X})$ is realized by the distribution that places all its mass on \mathbf{X} .

Naive Mean Field Algorithm: The naive mean field approach is the easiest to implement and characterized. It is based on choosing a factorized distribution

$$p(z_1, z_2, \dots, z_N | \boldsymbol{\theta}) = \prod_{s \in V_G} p(\mathbf{z}_s | \boldsymbol{\theta}) \quad (2.108)$$

as the tractable approximation where V_G is the vertex set of the graph G . The naive mean field updates are a particular set of recursions for finding a stationary point in $\mathcal{M}_F(G)$ of the resulting lower bound optimization problem. For a vast majority of problems that assume exchangeability of random variables, this approach involves applying fixed point iteration techniques in the E step and possibly some non-linear gradient ascent algorithms for parameter optimization in the M step. The fixed point iterations in the E step arise out of solving for the root of the maximum likelihood expressions involving coupled dependencies between variables.

Let us now consider the case where we have $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ as the observations and $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ being the hidden variables. Under this scenario, we consider distributions of the form $q(z_n | \boldsymbol{\phi}) = \prod_{n=1}^N q(\mathbf{z}_n | \boldsymbol{\phi}_n)$, where $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots, \boldsymbol{\phi}_N\}$ are variational parameters. Using this family of distributions, we simplify the likelihood bound using the chain rule:

$$\log p(\mathbf{X} | \boldsymbol{\theta}) \geq \log p(\mathbf{X} | \boldsymbol{\theta}) + \sum_{n=1}^N E_q[\log p(\mathbf{z}_n | \mathbf{Z}_{-n}, \mathbf{X}, \boldsymbol{\theta})] - \sum_{n=1}^N E_q[\log q(\mathbf{z}_n | \boldsymbol{\phi}_n)] \quad (2.109)$$

To optimize with respect to $\boldsymbol{\phi}_n$, we select the factors from Equ. 2.109 that depend on $\boldsymbol{\phi}_n$ to obtain:

$$f_n = E_q[\log p(z_n | \mathbf{Z}_{-n}, \mathbf{X}, \boldsymbol{\theta})] - E_q[\log q(\mathbf{z}_n | \boldsymbol{\phi}_n)] \quad (2.110)$$

Under the assumption that the variational distribution $q(z_n | \boldsymbol{\phi}_n)$ is in the exponential family, we have:

$$q(\mathbf{z}_n | \boldsymbol{\phi}_n) = h(\mathbf{z}_n) \exp\{\boldsymbol{\phi}_n^T \mathbf{z}_n - A(\boldsymbol{\phi}_n)\} \quad (2.111)$$

Equ. 2.110 thus simplifies as follows:

$$\begin{aligned} f_n &= E_q[\log p(z_n | \mathbf{Z}_{-n}, \mathbf{X}, \boldsymbol{\theta}) - \log h(\mathbf{z}_n) - \boldsymbol{\phi}_n^T \mathbf{z}_n + A(\boldsymbol{\phi}_n)] \\ &= E_q[\log p(z_n | \mathbf{Z}_{-n}, \mathbf{X}, \boldsymbol{\theta}) - \log h(\mathbf{z}_n)] - \boldsymbol{\phi}_n^T A'(\boldsymbol{\phi}_n) + A(\boldsymbol{\phi}_n) \end{aligned} \quad (2.112)$$

since $E_q[\mathbf{z}_n] = A'(\boldsymbol{\phi}_n)$ (see Theorem 2.1.1). The derivative with respect to $\boldsymbol{\phi}_n$ is:

$$\frac{\partial f_n}{\partial \boldsymbol{\phi}_n} = \frac{\partial}{\partial \boldsymbol{\phi}_n} (E_q[\log p(z_n | \mathbf{Z}_{-n}, \mathbf{X}, \boldsymbol{\theta}) - \log h(\mathbf{z}_n)]) - \boldsymbol{\phi}_n^T A''(\boldsymbol{\phi}_n) \quad (2.113)$$

From this, we find that the optimal ϕ_n satisfies:

$$\phi_n^* = [A''(\phi_n)]^{-1} \frac{\partial}{\partial \phi_n} (E_q[\log p(z_n | \mathbf{Z}_{-n}, \mathbf{X}, \boldsymbol{\theta}) - \log h(\mathbf{z}_n)]) \quad (2.114)$$

When the conditional $p(\mathbf{z}_n | \mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X})$ belongs to an exponential family distribution, then we have:

$$p(\mathbf{z}_n | \mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X}) = h(\mathbf{z}_n) \exp \left\{ w_n(\mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X})^T \mathbf{z}_n - A(w_n(\mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X})) \right\} \quad (2.115)$$

where $w_n(\mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X})$ denotes the canonical parameter for \mathbf{z}_n when conditioning on the remaining latent variables and the observations. This yields simplified expressions for the expected log probability of \mathbf{z}_n and its first derivative:

$$\begin{aligned} E_q[\log p(z_n | \mathbf{Z}_{-n}, \mathbf{X}, \boldsymbol{\theta})] &= E_q[\log h(\mathbf{z}_n)] + E_q[w_n(\mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X})]^T A'(\phi_n) - E_q[A(w_n(\mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X}))] \\ \frac{\partial}{\partial \phi_n} E_q[\log p(z_n | \mathbf{Z}_{-n}, \mathbf{X}, \boldsymbol{\theta})] &= \frac{\partial}{\partial \phi_n} E_q[\log h(\mathbf{z}_n)] + E_q[w_n(\mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X})]^T A''(\phi_n) \end{aligned} \quad (2.116)$$

Using the first derivative in Equ. 2.114, the maximum is attained at:

$$\begin{aligned} \phi_n^* &= [A''(\phi_n)]^{-1} \left(E_q[w_n(\mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X})]^T \right) A''(\phi_n) \\ &= E_q[w_n(\mathbf{Z}_{-n}, \boldsymbol{\theta}, \mathbf{X})] \end{aligned} \quad (2.117)$$

Non Convexity of Mean Field: An important fact about the mean field approach is that the variational problem 2.107 may be nonconvex, so that there may be many local minima, and the mean field updates can have multiple solutions. The source of this nonconvexity can be understood in different ways [Wainwright and Jordan, 2008, Jaakkola, 2000a], depending on the formulation of the problem. Perhaps the most intuitive way to understand it is by observing the structure of mean parameter sets. The geometric perspective of the set $\mathcal{M}(G)$ and its inner approximation $\mathcal{M}_F(G)$ reveals that more generally, mean field optimization is always non-convex for any exponential family in which the state space \mathcal{X} is finite [Wainwright and Jordan, 2008].

Figure 2.4 shows some cartoon illustrations of the mean field principle of imposing of tractable distributions to solve for ML parameter estimates. Figure 2.4a shows that, for discrete random variables, although the realizable mean parameter set $\mathcal{M}(G)$ is convex, when limited to mean parameters of tractable distributions $\mathcal{M}_F(G)$, the set becomes non-convex. The non-convexity arises due to the δ functions connecting $\mathcal{M}_F(G)$ to the extreme points $\boldsymbol{\mu}_e$ in $\mathcal{M}(G)$.

An illustration of the inequality $\Upsilon(\mathbf{Z})^T \log \boldsymbol{\mu} \geq -c$ ($c > 0$ and $= 1$ here) for the random variable $\mathbf{Z} | \mathbf{x} \sim Mult(\boldsymbol{\mu})$ with the multinomial being degenerated to a binomial in two dimensions is shown in the top right corners of Figs. 2.4a, 2.4b and 2.4c where the set of mean parameters is the set of expected sufficient statistics of the hidden variables \mathbf{Z} given a particular set of samples \mathbf{x} . For a K -dimensional multinomial, we have the constraint $\sum_{k=1}^K \mu_k = 1$ and $0 \leq \mu_k \leq 1$. If the random variables \mathbf{Z} are the indicators for components in a mixture or mixed membership model, then efficiently solving this type of inequality subject to the number of constraints (which quickly becomes exponential), is the essence of mean field optimization for probabilistic models. The objective in this case is to find an extremum for $\boldsymbol{\mu}^* \in \mathcal{M}_{F_{\mathbf{Z}|\mathbf{x}}}(G)$ through the sufficient statistics $\Upsilon(\mathbf{Z}, \mathbf{x})$ but restricted only to the feasible positive

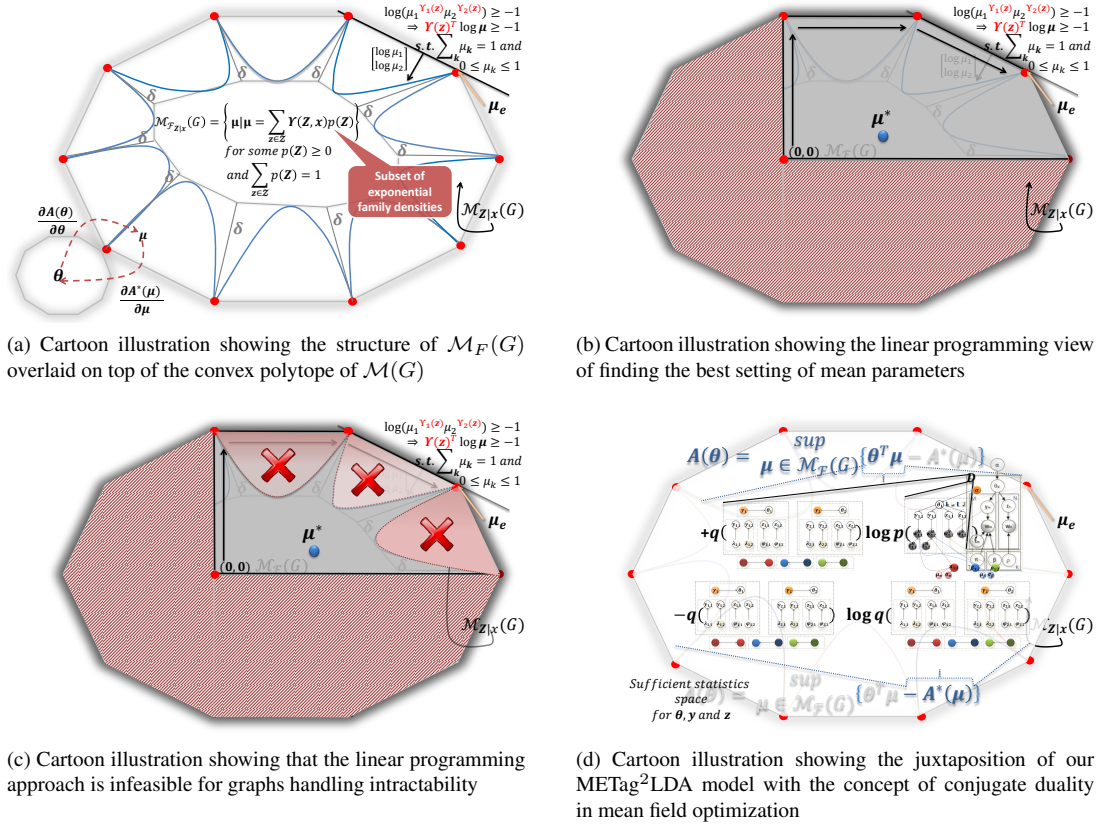


Figure 2.4: Cartoon illustrations highlighting the key features of mean field optimization

region of the coordinate system where \mathbf{Z} is defined.

Figure 2.4b shows the relation of such an optimization technique to the well known linear programming problem where a pivot is initially chosen as a basis which contains only slack variables. Linear programming optimization traces out a path shown in black arrows in Fig. 2.4b where the pivot is cycled through possible extreme points $\mu_e = \Upsilon(\mathbf{Z}, \mathbf{x})$ of the of the sufficient statistics for each possible configuration of the discrete random variable \mathbf{Z} holding the set of samples \mathbf{x} fixed. By doing so, it replaces each slack variable within a basis with the original variables μ_k as much as possible while maximizing a linear function of $\log \mu$ and $\Upsilon(\mathbf{Z}, \mathbf{x})$.

Unfortunately, although this is a viable choice for small graphs, Fig. 2.4c shows that due to the induction of tractable distributions, the effective realization mean parameter space becomes heavily constrained (constricted as shown in the illustration) and the cycle of pivots through the boundaries of the convex hull of $\mathcal{M}_{Z|X}(G)$ is no longer possible due to the possible chance of violating such constraints i.e. finding a solution to the maximum likelihood objective function for some μ_k which lands up in the regions marked by the red \times in Fig. 2.4c.

Finally, in Fig. 2.4d, we observe the close relationship of Equ. 2.107 to Equ. 2.102 visualized through our proposed METag²LDA (see Fig. 1.7e) model. The mean parameter space μ consists of the expected sufficient statistics of $\theta_{1:D}$, $\mathbf{Y}_{D \times \{1:M_d\}, d \in D}$ and $\mathbf{Z}_{D \times \{1:N_d\}, d \in D}$ computed under the parameters of the variational distributions $\gamma_d, \lambda_{d,m}, \phi_{d,n}$ that map to the model parameters $\alpha, \{\beta, \tau\}$ and

ρ respectively. The term $-q(\cdot) \log q(\cdot)$ is the entropy of the variational q distribution and is equal to $-A^*(\boldsymbol{\mu})$. The expanded expressions $\boldsymbol{\theta}^T \boldsymbol{\mu}$ can be found by inspecting the respective parameters in Equ. 4.26. The dimensionality of the variational parameters should not be confused with the set of three axes shown in the lower left corner in Fig. 2.4d that just conveys the idea that $\boldsymbol{\mu}$ consists of *three distinct sets* of mean parameters.

In the next section, we touch upon the basic topic model Latent Dirichlet Allocation (LDA) [Blei et al., 2003] which has been subsequently extended in this thesis to satisfy several criteria along the lines of incorporating domain knowledge at the word and document levels. The domain knowledge can arise from linguistic (such as part-of-speech) annotations, crowd-sourcing (such as Wikipedia article categories) or even multimedia (such as embedded videos).

2.7 Latent Dirichlet Allocation and Variational Bayesian EM

Topic models initially started out as an use case of generative Bayesian modeling for textual corpora where a corpus is a collection of documents. The basic assumption was that the documents are represented as a random mixture over latent topics, where each topic is characterized by a distribution on words. One such model, Latent Dirichlet Allocation (LDA) [Blei et al., 2003] is shown in Fig. 2.5a as a directed graphical model.

Graphical models allow us to graphically represent interactions between unobserved and observed random variables and parameters in a clear and concise way using directed arcs for causality relationships and square plate notation for identifying variable repetition (identified by the alphabet in one of the corners of the plate). Shaded nodes represent observed variables, while unshaded nodes represent latent variables. In Fig. 2.5a, unshaded nodes outside of the rectangular plates represent model parameters as they do not grow with the data.

In Figure 2.5a it is observed that the only observed random variables are the words with M being the number of word positions in document $d \in \{1, \dots, D\}$. The model parameters are $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_{1:K}$ where $\boldsymbol{\theta}$ represents the topic or theme proportions for a document represented as a K -dimensional vector with each dimension corresponding to a multinomial distribution over the vocabulary. Clearly, the use of proportions suggest that the only input to the model are documents represented in terms of their word counts. The likelihood of a document w.r.t. the model parameters is obtained by integrating out the beliefs encoded in the hidden variables as

$$\int p(\boldsymbol{\theta}_d | \boldsymbol{\alpha}) \left(\prod_{m=1}^{M_d} \sum_{z_{d,m}} p(z_{d,m} | \boldsymbol{\theta}_d) p(w_{d,m} | z_{d,m}, \boldsymbol{\beta}) \right) d\boldsymbol{\theta}_d \quad (2.118)$$

This likelihood is intractable to compute and approximate algorithms like variational Bayes [Beal, 2003] are used to overcome the intractability. To explain the phenomenon captured by the graphical model representation in Figure 2.5a, the document generation process is written in the following “statistical pseudocode” form. Let $\boldsymbol{\alpha}$ be a K -dimensional parameter, and let topics $\boldsymbol{\beta}_{1:K}$ be K multinomials over a fixed vocabulary of words. The number K is the same as the number of latent topics and equals the dimensionality of the topic indicator variable z_{d_n} . LDA assumes that *an* M -word document d arises from the following generative process:

For each document $d \in [1, \dots, D]$,

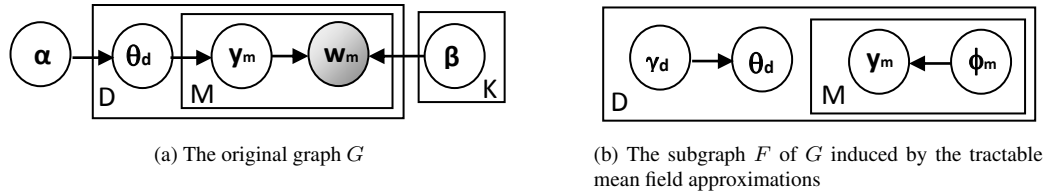


Figure 2.5: The graphical model for the Latent Dirichlet Allocation

Draw $\theta_d | \alpha \sim \text{Dir}(\alpha)$
 For each position $m \in \{1, \dots, M_d\}$ in document d :
 Draw a topic assignment $z_{d,m} | \theta_d \sim \text{Mult}(\theta_d)$
 Draw a word $w_{d,m} | \{z_{d,m}, \beta_{1:K}\} \sim \text{Mult}(\beta_{z_{d,m}})$

In the learning phase, the model parameters α and $\beta_{1:K}$ are estimated from training data and through inference on test data, the topic indices $z_{d,m}$'s are obtained using the estimated model parameters. In the test set, thus, the vocabulary remains the same and only the word counts (may) change for each document. The topics capture the discrete distributions whose modes highlight words that appear to have a semantic connection. The primary reason is the way how semantically related words appear in the documents and imposing a hierarchical document schema which exhibits a mixed membership to the latent topics in varying proportions allows for better separation of the modes than if no document schema was assumed or no more than one topic is allowed to account for the words in a document.

In fact, the motivation behind developing the LDA model was to overcome the limitations of mixture modeling through the simple mixture of unigrams model and the probabilistic Latent Semantic Indexing (pLSI) model [Hofmann, 1999]. For the simple mixture of unigram models, all words in a document are assumed to be sampled from a single topic while in the pLSI model, the assumption is that the proportion of topics for a new document match those in one or more of the training documents. This latter constraint is alleviated to some extent by “folding-in” in the new document by re-fitting the topic parameters for the new document anew. LDA overcomes both of these problems by allowing the topic proportions θ_d of a document d to be controlled by a random variable that can easily allow for the probability that the new document be endowed with a different set of topics than was seen in one or more training documents.

2.7.1 The Success Behind Latent Dirichlet Allocation

After analyzing a natural language text corpora, for the topics to appear intuitive, LDA must simultaneously satisfy two objectives:

1. For each document, its words must be allocated to as few topics as possible.
2. For each topic, only a few words must only be assigned high probabilities.

Unfortunately, these goals compete with each other: If a document is assigned only a single topic, as is the case with naive mixture of unigrams model, then that makes satisfying condition 2 hard—all of its words must have high probability under that topic. On the other hand, allocating only a few words in each topic makes satisfying condition 1 hard—to cover a document's words, it must assign many topics to it for e.g. a general Wikipedia article with its varied section contents consists of several different sub topics. Trading off these goals appears to find groups of tightly co-occurring words which appear semantically related. However, this semantic relatedness of words is a direct consequence of how

language has evolved to express ideas in a coherent fashion [Chang et al., 2009].

2.7.2 LDA: How much data is necessary to learn the model parameters?

In general it is an open problem to bound the amount of data required to learn the parameters of a topic model with probabilistic guarantees. Recently an algorithm which provably learns the parameters of a topic model *given samples from the model* has been proposed in [Arora et al., 2012]. For a topic model, there is an unknown topic-word matrix β^T with nonnegative entries of dimension $V \times K$, and a probabilistic unknown matrix $\theta_{1:D}^T$ that is dimension $K \times D$. Each column of $\beta^T \theta_{1:D}^T$ is viewed as a probability distribution on rows i.e. the vocabulary elements, and each column comprises of $M_d \ll V$ i.i.d. samples from the associated distribution. The inference machinery for a topic model attempts to reconstruct β and the parameters of the generating distribution for $\theta_{1:D} - \alpha$. The proof provides a bound on the data provided that the topic-word distributions $\beta_{1:K}$ satisfy a condition called *separability* as reproduced below from [Arora et al., 2013].

The topic-word matrix $\beta_{1:K}$ is p -separable for $p > 0$ if for each topic k , there is some word with index j such that $\beta_{k,j} \geq p$ and $\beta_{k',j} = 0$ for $k' \neq k$. This intuitively means that $\beta_{1:K}$ has a diagonally dominant structure i.e. has a diagonal matrix (upto row permutations). Such a word with index j is called an anchor word because when it occurs in a document, it is a perfect indicator that the document is at least partially about the corresponding topic, since there is no other topic that could have generated the word. Additionally suppose that each document is of length $L \geq 2$, and let $\mathbf{R} = E_{p(\theta_{1:D}|\alpha)}[\theta_{1:D}^T \theta_{1:D}]$ be the $[K \times D] \times [D \times K] = K \times K$ topic-topic covariance matrix. Also, as before denote $\theta_{d,k}$ to be the expected proportion of topic k in a document d generated according to α .

A recent theorem has been proved in [Arora et al., 2012] on the amount of training documents needed to learn the parameters of a topic model like LDA. The theorem states that there is a *polynomial time algorithm* that learns the parameters of a topic model if the number of documents is at least

$$D = \max \left\{ \mathcal{O} \left(\frac{a^4 K^6 \log V}{\epsilon^2 p^6 \gamma^2 L} \right), \mathcal{O} \left(\frac{a^2 K^4 \log K}{\gamma^2} \right), \mathcal{O} \left(\frac{K^2 \log K}{\epsilon^2} \right) \right\} \quad (2.119)$$

where p is the separability of β , γ is the condition number of \mathbf{R} , and $a = \max_{k,k'} \theta_k / \theta_{k'}$ is the topic imbalance of the model. The algorithm learns the word-topic matrix $\beta_{1:K}$ and the $K \times K$ topic-topic covariance matrix \mathbf{R} up to additive error ϵ .

Further denote \mathbf{Q} to be the $V \times V$ word-word covariance matrix $\beta^T \mathbf{R}(\Pi) \beta$ where Π is the distribution which generates the columns of $\theta_{1:D}^T$ [Arora et al., 2012]. When number of documents is large enough, it has been shown that \mathbf{Q} is close to $\beta^T \theta^T \theta \beta$; infact, when $m > \frac{50 \ln V}{L \epsilon_Q}$, with high probability all entries of $|\mathbf{Q} - \frac{1}{m} \beta^T \theta^T \theta \beta| \rightarrow \epsilon_Q$.

Hardness of computing Maximum Likelihood estimates for topic model parameters: It is NP-hard even with two topics to find the Maximum Likelihood estimates of the word-topic distributions in LDA [Arora et al., 2013] and the essence of the work in [Arora et al., 2012] has been to show that even in the case of a separable topic-word matrix in the general case, it is NP-hard to compute the Maximum Likelihood estimates. Sontag and Roy recently proved in [Sontag and Roy, 2011] that given the matrix $\beta_{1:K}$ and a document d , computing the Maximum A Posteriori (MAP) estimate for the distribution on topics that generated document d is NP-hard as well.

We next briefly touch upon the exponential family representation of LDA and its variational dual and

how independence assumptions lead to the development of an efficient approximate inference machinery to find optimal, albeit locally, estimates of β and α .

2.7.3 Exponential Family Representation, Mean Field and Variational Bayes

The basic topic model LDA as shown in Fig. 2.5a is a hierarchical mixed membership model which belongs to the exponential family of models. The following algebraic manipulations help us ascertain the exponential family inclusion of the model.

LDA in exponential family representation: Words \mathbf{w} are drawn from a multinomial distribution, $p(w = j|z = k, \beta) = \exp(\log \beta_{k,j})$, for $j \in \{1, \dots, V\}$ and $z \in \{1, \dots, K\}$, where $\beta_{k,j}$ is a parameter encoding the probability of the j^{th} word given the k^{th} topic. This conditional distribution can be expressed as an exponential family in terms of indicator functions as follows:

$$p_{\beta}(w|z) \propto \exp\left(\sum_{k=1}^K \sum_{j=1}^V \log \beta_{k,j} \delta(z, k) \delta(w, j)\right) \quad (2.120)$$

where $\delta(z, k)$ is an $\{0, 1\}$ -valued indicator for the event $\{Z = k\}$, and similarly for $\delta(w, j)$. The topic variable Z also follows a multinomial distribution whose parameters are determined by the Dirichlet variable as follows:

$$p(z|\theta) \propto \exp\left(\sum_{k=1}^K \log \theta_k \delta(z, k)\right) \quad (2.121)$$

Finally, at the top level of the hierarchy, the Dirichlet variable θ has a pdf of the form

$$p_{\alpha}(\theta) \propto \exp\left(\sum_{k=1}^K \alpha_k \log \theta_k\right) \quad (2.122)$$

Overall then, for a single triplet $(\theta, \mathbf{Z}, \mathbf{w})$, the LDA model is an exponential family with parameter vector $\Theta = (\alpha, \beta)$, and an associated density of the form:

$$p_{\alpha}(\theta)p(z|\theta)p_{\beta}(w|z) \propto \exp\left(\sum_{k=1}^K \log \theta_k \delta(z, k) + \sum_{k=1}^K \alpha_k \log \theta_k\right) \times \exp\left(\sum_{k=1}^K \sum_{j=1}^V \log \beta_{k,j} \delta(z, k) \delta(w, j)\right) \quad (2.123)$$

The sufficient statistics Υ consist of the collections of functions $\{\log \theta_k\}$, $\{\delta(z, k) \log \theta_k\}$, and $\{\delta(z, k) \delta(j, w)\}$. The full LDA model entails replicating these types of local structures many times. For each fixed \mathbf{Z} , the set \mathcal{M}_F of mean parameters is of the form $\mu = E_{q(\mathbf{Z}|\mu)}[\Upsilon(\mathbf{Z}, \mathbf{X})]$.

We now briefly discuss about how the mean field factorization principle (see Sect. 2.6) gets applied in this case and thus how the factorization shown in Fig. 2.5b refers to the tractable subgraph for LDA. In a general model, let the set of all observed random variables be denoted by $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ (for LDA, \mathbf{X} is denoted by \mathbf{W}). The model also have a set of latent variables which are denoted by $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$. Note that N and M can be different in general. For LDA, N and M are same for a document d and the set of latent variables for a document d is $\mathbf{Z}_d = \{\mathbf{z}_d, \theta_d\}$ while the parameters are $\Theta = \{\alpha, \beta_{1:K}\}$. The probabilistic model specifies the joint distribution $p(\mathbf{X}, \mathbf{Z})$, and our goal is to find an approximation for the posterior distribution $p(\mathbf{Z}|\mathbf{X})$ assuming conditional dependence on the parameter set Θ to be implicit. Since $p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{Z})}{p(\mathbf{X})}$, in order to infer $p(\mathbf{Z}|\mathbf{X})$ the normalizing constant $p(\mathbf{X})$ needs to be computed, which is the marginal probability of the observations. Using

Jensen's inequality, the following hold true with dependence on the parameters being implicit:

$$\begin{aligned}
\log p(\mathbf{X}) &= \log \int p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} = \log \int q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\
&\geq \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\
&= E_q[\log p(\mathbf{X}, \mathbf{Z})] - E_q[\log q(\mathbf{Z})] = \mathcal{L}(q, \mathbf{X})
\end{aligned} \tag{2.124}$$

where the bound \mathcal{L} is tight when $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X})$. Although q can be *any* distribution over the hidden variables, in practice q is usually chosen from a family of functions such that q factorizes over the set of latent variables and have easily computable moments. Then variational methods try to compute $\log p(\mathbf{X})$ by finding q that maximizes $\mathcal{L}(q, \mathbf{X})$,

$$\log p(\mathbf{X}) = \operatorname{argmax}_{q \in \mathcal{M}_{tract}} \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[q(\mathbf{Z})] \tag{2.125}$$

where \mathcal{M} is a family of distributions including $p(\mathbf{Z}|\mathbf{X})$ which assert *no more conditional independence* than those in $p(\mathbf{Z}|\mathbf{X})$. For instance, \mathcal{M} can be the set of all joint probability distributions on \mathbf{Z} that contains no conditional but only marginal independence. In general, this optimization cannot be performed over \mathcal{M} due to an exponential state space. So, in order to compute $\mathcal{L}(q, \mathbf{Z})$, \mathcal{M} is restricted to a simpler family \mathcal{M}_{tract} , which is tractable, so that

$$\log p(\mathbf{X}) \geq \operatorname{argmax}_{q \in \mathcal{M}_{tract}} \mathcal{L}(q, \mathbf{Z}) \tag{2.126}$$

Since the tractable subfamilies of distributions \mathcal{M}_{tract} are fully factorized, the distributions in \mathcal{M}_{tract} have the form

$$q(\mathbf{Z}) = q(z_1|\phi_1)q(z_2|\phi_2) \dots q(z_N|\phi_N) \tag{2.127}$$

where, under the assumptions of the naive mean field approach, each z_n can be any distribution governed by a variational parameter ϕ_n and there are N different distributions corresponding to the i.i.d z_1, z_2, \dots, z_N s. The approximating distribution q is such a distribution that maximizes the objective function $\mathcal{L}(q, \mathbf{Z})$ over $\mathcal{M}_{tract} \subset \mathcal{M}$.

In more complicated scenarios, dependencies between z_n 's can be considered by keeping some edges between the corresponding random variables in the graphical model as is. This is usually referred to as the structured mean field approach an example of which can be found in [Ghahramani and Jordan, 1997].

In LDA, the observation of the word $w_{d,m}$ makes the hidden variable as well as the parameters coupled under the D-separation criteria [Shachter, 1998] and this leads to the removal of the edges that lead to the coupling giving rise to the dual graph shown in Fig. 2.5b.

The removal of edges in the original graph for LDA (Fig. 2.5a) makes the hidden variables θ_d 's and $z_{d,m}$'s to be marginally independent and each governed by the same type of distributions as in the original graph but with *free* parameters (Dirichlet: γ_d for θ_d and Multinomial: $\phi_{d,m}$ for $\beta_{z_{d,m}}$) which are allowed to *vary* according to the statistical moments of the data.

Generally speaking, the naive mean field approximation is the case where each $q_{\mathbf{z}_n}(\mathbf{z}_n)$ is fully factorized over the hidden variables:

$$q(\mathbf{z}_n) = \prod_{j=1}^{|\mathbf{z}_n|} q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \quad (2.128)$$

where $|\mathbf{z}_n|$ is the number of hidden variables interacting with \mathbf{z}_n (for example if \mathbf{z}_n belongs to a clique and there are N such cliques). In this case the expression for $\mathcal{L}(q_{\mathbf{Z}}(\mathbf{Z}), \boldsymbol{\theta})$ given in Equ. 2.79 becomes:

$$\begin{aligned} \mathcal{L}(q_{\mathbf{Z}}(\mathbf{Z}), \boldsymbol{\theta}) &= \sum_n \int \left[\prod_{j=1}^{|\mathbf{z}_n|} q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}) - \prod_{j=1}^{|\mathbf{z}_n|} q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \ln \prod_{j=1}^{|\mathbf{z}_n|} q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \right] d\mathbf{z}_n \\ &= \sum_n \int \prod_{j=1}^{|\mathbf{z}_n|} \left[q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}) - \sum_{j=1}^{|\mathbf{z}_n|} q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \ln q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \right] d\mathbf{z}_n \\ &= \sum_n \int q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \prod_{j=1}^{|\mathbf{z}_n|} [q_{\mathbf{z}_n \neq j}(\mathbf{Z}_{n \neq j}) \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}) d\mathbf{Z}_{n \neq j}] dz_{n,j} \\ &\quad - \sum_n \sum_{j=1}^{|\mathbf{z}_n|} \int \prod_{j=1}^{|\mathbf{z}_n|} [q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \ln q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j})] d\mathbf{z}_n \\ \therefore \mathcal{L}(q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}), \boldsymbol{\theta}) &= \int \left\{ \prod_{j=1}^{|\mathbf{z}_n|} [q_{\mathbf{z}_n \neq j}(\mathbf{Z}_{n \neq j}) \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}) d\mathbf{Z}_{n \neq j}] \right\} q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) dz_{n,j} \\ &\quad - \int [q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) \ln q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j})] d\mathbf{z}_{n,j} + \text{const} \end{aligned} \quad (2.129)$$

Using a Lagrange multiplier to enforce normalization of the each of the approximate posteriors for z_n , we take the functional derivatives of $\mathcal{L}(q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}), \boldsymbol{\theta})$ with respect to each of the $q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j})$ s and equate to zero, obtaining:

$$q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j}) = \frac{1}{Z_{n,j}} \exp \left[\int \prod_{j' \neq j}^{|\mathbf{z}_n|} q_{\mathbf{z}_n,j'}(\mathbf{z}_{n,j'}) \ln p(\mathbf{z}_n, \mathbf{x}_n | \boldsymbol{\theta}) d\mathbf{Z}_{n \neq j} \right] \quad (2.130)$$

for each $n \in \{1, \dots, N\}$ and each $j \in \{1, \dots, |\mathbf{z}_n|\}$. We have used the notation $d\mathbf{Z}_{n \neq j}$ to denote the element of integration for all items in \mathbf{z}_n except $\mathbf{z}_{n,j}$, and the notation $\prod_{j' \neq j}$ to denote a product of all terms excluding j . For the n^{th} datum, the update equation 2.130 is applied to each hidden variable j in turn and represents a set of coupled equations for the approximate posterior over each hidden variable. These fixed point equations are called mean-field equations and examples of these variational approximations can be found in [Jaakkola, 2000b] and the references therein. A very similar set of update equations also hold for structured mean field where we do not assume the factorization of the form $\prod_{j=1}^{|\mathbf{z}_n|} q_{\mathbf{z}_n,j}(\mathbf{z}_{n,j})$ for the variables in the n^{th} partition (see Chapter 10 in [Bishop, 2006]) but the inference is albeit complicated and computationally very expensive. In all of the models which we consider in the following chapters, $|\mathbf{z}_n|$ is of cardinality 1.

In a truly Bayesian setting, we put a prior η (usually a symmetric Dirichlet but can be an asymmetric one in the form of a Dirichlet tree as well [Andrzejewski et al., 2009]) over the topic Multinomials $\beta_{1:K}$ (see Fig 2.6). In this setting, all references to β_k is replaced with $E_{q_\eta}[\log \beta_k | \eta]$ where there is yet another factorization of the form $q_\eta(\eta) = \prod_{k=1}^K (\beta \sim Dir(\eta))$. We will study a model with this kind of

“global” parameter level factorization in the context of Gaussian parameters in Chapter 6.

2.7.4 Kullback-Leibler divergence in LDA

An alternative interpretation to explain mean-field variational methods is to minimize the difference between the approximating distribution and the target distribution using KL divergence (Kullback-Leibler divergence) [Cover and Thomas, 2006]. KL divergence is an information theoretic measure of the “distance” between two distributions, defined as for $p(\mathbf{X})$ and $q(\mathbf{X})$,

$$KL(p(\mathbf{X})||q(\mathbf{X})) = \mathbb{E}_{p(\mathbf{X})} \left[\log \frac{p(\mathbf{X})}{q(\mathbf{X})} \right] \quad (2.131)$$

Maximizing the objective function $\mathcal{L}(q, \mathbf{X})$ with respect to q is equivalent to finding the $q^* \in \mathcal{M}_{tract}$ which minimizes $KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}))$. In other words,

$$\begin{aligned} q^* &= \operatorname{argmax}_{q \in \mathcal{M}_{tract}} \mathcal{L}(q, \mathbf{Z}) \\ &= \operatorname{argmax}_{q \in \mathcal{M}_{tract}} (\mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})]) \\ &= \operatorname{argmin}_{q \in \mathcal{M}_{tract}} (\mathbb{E}_q[\log q(\mathbf{Z})] - \mathbb{E}_q[p(\mathbf{Z}|\mathbf{X})]) \\ &= \operatorname{argmin}_{q \in \mathcal{M}_{tract}} KL(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{X})) \end{aligned} \quad (2.132)$$

Note that $KL(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{X}))$ is the difference in the lowerbound $\mathcal{L}(q, \mathbf{Z})$ and $\log p(\mathbf{X}|\Theta)$ where $p(\mathbf{X})$ is the true likelihood of the data and Θ are the parameters of the model.

2.7.5 The E Step Inner Loop in the Mean Field Optimization of LDA

Let us revisit the inner-loop that is applied to find the optimum of the *posterior distributions* over the free variational parameters for the document level random variables given the current setting of optimal parameters (see Fig. 2.5b).

Algorithm 1 doc_e_step

```

1:  $\gamma_{d,k} = \alpha_k + \frac{\text{corpus.document}[d].\text{num\_words}}{K}$ 
2:  $\phi_{d,m,k} = \frac{1.0}{K}$ 
3:  $\text{elbo\_current} \leftarrow 0$ ;  $v\_iter \leftarrow 0$ 
4: while not converged do
5:   for  $m = 1 \rightarrow M_d$  do
6:     for  $k = 1 \rightarrow K$  do
7:       update  $\phi_{d,m,k}$  as  $\log \phi_{d,m,k} = \log \beta_{k,w_{d,m}} + \psi(\gamma_{d,k}^{(t)})$ 
8:     end for
9:     Normalize  $\phi_{d,m}^{(t+1)}$ s to sum to 1
10:  end for
11:  update  $\gamma_{d,k}^{(t+1)}$  as  $\gamma_{d,k}^{(t+1)} = \alpha_k + \sum_{m=1}^{M_d} \phi_{d,m,k}^{(t+1)}$ 
12:   $\text{elbo\_current} \leftarrow \text{compute\_likelihood}()$  {To compute likelihoods, see [Blei et al., 2003]}
13:   $v\_iter \leftarrow v\_iter + 1$ 
14: end while
15: return  $\text{elbo\_current}$ ;
```

In Algo. 1, line 11 where it is stated “update $\gamma_{d,k}^{(t+1)}$ as $\gamma_{d,k}^{(t+1)} = \alpha_k + \sum_{m=1}^{M_d} \phi_{d,m,k}^{(t+1)}$,” as in [Blei et al., 2003] is actually modified in the implementation in [Blei et al., 2003] to conform to the iterative structure of the fixed point iteration algorithm [Conte and Boor, 1980]. Usually a fixed point iteration algorithm is used to find the roots of an equation $f(z) = 0$ by the following construction: An equation of the form

$$z = g(z) \tag{2.133}$$

is derived from $f(z) = 0$ so that any solution of Equ. 2.133, i.e. any **fixed point** of $g(z)$ is a solution of $f(z) = 0$. Given a function $f(z)$ of a variable z at a particular iteration (t), the value of the next fixed point iterate, z , can be obtained using iterative improvements as:

$$z_{n+1} = f(z_n) \tag{2.134}$$

Three assumptions need to hold to obtain a solution using the fixed point iteration method.

- i) There is an interval $I = [a, b]$ such that $\forall z \in I, g(z)$ is defined and $g(z) \in I$
- ii) The iteration function $g(z)$ is continuous on $I = [a, b]$
- iii) The iteration function $g(z)$ is differentiable on $I = [a, b]$. Further there exists a nonnegative constant $C < 1$ such that $\forall z \in I, |g'(z)| \leq C$

If the first and third assumptions hold then $g(z)$ has exactly one fixed point $\zeta \in I$ and starting with any point $z_0 \in I$, the sequence $\{z_1, z_2, \dots\}$ obtained using fixed point iteration converges to ζ . The third constraint is particularly useful to assess convergence properties of the iterates. In particular if $y = f(x)$ is a function whose fixed point is ζ , then ζ lies on both $y = f(x)$ and $y = x$. This means that if the slope of $f(x)$ is higher than 1 i.e. $\tan(\pi/4)$ in absolute value [Conte and Boor, 1980], then convergence is problematic in this framework and we need to use some early stopping criterion.

In our context of solving lower bounds for models such as LDA, the expressions for finding the optimum of the posterior distributions over the hidden variables (and parameters is priors are used) give rise to a coupled set of equations. This facilitates the use of fixed point iteration technique to obtain a possible solution depending on the initial conditions. The argument is simple: Let f and g be two continuous functions such that $y = f(z)$ and $z = g(y)$. Then

$$\begin{aligned} y &= f(z); & z &= g(y) \\ \implies y &= f(g(y)) = (f \circ g)(y) = h(y) \end{aligned} \tag{2.135}$$

where $h = (f \circ g)$ is the composition function. This scheme allows us to find $\gamma^{(t+1)}$ from $\gamma^{(t)}$ as follows:

$$\begin{aligned} \gamma_k^{(t+1)} &= \alpha_k + \sum_{n=1}^N \phi_{d,m,k}^{(t+1)} \\ \implies \alpha_k &= \gamma_k^{(t)} - \sum_{n=1}^N \phi_{d,m,k}^{(t)} \\ \implies \gamma_k^{(t+1)} &= \gamma_k^{(t)} + \sum_{n=1}^N \left(\phi_{d,m,k}^{(t+1)} - \phi_{d,m,k}^{(t)} \right) = h(\gamma_k^{(t)}) \end{aligned} \tag{2.136}$$

2.7.6 Gibbs Sampling versus Variational Bayes in Topic Models

In this section, we briefly go over some preliminary concepts on Gibbs sampling algorithm and also draw a connection between it and the EM algorithm [Bishop, 2006]. Sampling methods such as Gibbs sampling has been very popular in a variety of models that extend LDA [Griffiths and Steyvers, 2004, Rosen-Zvi et al., 2004, Li and McCallum, 2006, Andrzejewski et al., 2009, Mimno et al., 2009, Spiliopoulou and Storkey, 2012] and they have become so due to the avoidance of local minima properties of the EM algorithm as well as the ease of implementation when there is conjugacy of the distributions (mostly assumed exponential) of interacting random variables. A thorough treatise on sampling methods can be found in [Robert and Casella, 2005].

We now give an overview of the Gibbs sampling algorithm in the context of missing data. Given a probability density f , a density g that satisfies $\int_{\mathbf{Z}} g(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} = f(\mathbf{X})$ is called a *completion* of f . The density g is chosen so that the full conditionals of g are easy to simulate from. Denote P to be the dimensionality of \mathbf{Y} . For $P > 1$, if we write $\mathbf{y} = (\mathbf{X}, \mathbf{Z})$ and denote the conditional densities of $g(\mathbf{Y}) = g(y_1, \dots, y_P)$ by

$$\begin{aligned} Y_1 | y_2, y_3, \dots, y_P &\sim g(y_1 | y_2, y_3, \dots, y_P), \\ Y_2 | y_1, y_3, \dots, y_P &\sim g(y_2 | y_1, y_3, \dots, y_P), \\ &\dots \\ Y_P | y_1, y_2, \dots, y_{P-1} &\sim g(y_P | y_1, y_2, \dots, y_{P-1}) \end{aligned}$$

Then obtaining a sample $Y^{(t+1)}$ at time step $(t + 1)$ from a sample $Y^{(t)}$ at time step t is defined using the Gibbs sampling algorithm 2.

Algorithm 2 Gibbs Sampling

- 1: Given $(y_1^{(t)}, \dots, y_P^{(t)})$, simulate
 - 2: 1. $Y_1^{(t+1)} \sim g_1(y_1 | y_2^{(t)}, y_3^{(t)}, \dots, y_P^{(t)})$,
 - 3: 2. $Y_2^{(t+1)} \sim g_2(y_2 | y_1^{(t+1)}, y_3^{(t)}, \dots, y_P^{(t)})$,
 - 4: 2. $Y_3^{(t+1)} \sim g_3(y_3 | y_1^{(t+1)}, y_2^{(t+1)}, y_4^{(t)}, \dots, y_P^{(t)})$,
 - 5: \vdots
 - 6: P. $Y_P^{(t+1)} \sim g_P(y_P | y_1^{(t+1)}, \dots, y_{P-1}^{(t+1)})$
-

This Gibbs sampling method is equivalent to the composition of P Metropolis-Hastings algorithms with acceptance probabilities uniformly equal to 1 [Robert and Casella, 2005]. This means that every simulated value is accepted and that Gibbs sampling is inherently multi-dimensional. However, the Gibbs sampler does not apply to problems where the number of parameters vary because this destroys the irreducibility property of the resulting chain. This is the reason why Gibbs sampling technique is difficult to use in models like pLSA (probabilistic Latent Semantic Analysis) [Hofmann, 1999].

The practical applicability of Gibbs sampling depends on the ease with which samples can be obtained from the full conditional distributions $g_k(y_k | y_{*, -k})$. In the case of graphical models, the conditional distributions for the individual nodes depend only on the variables in the corresponding Markov blanket (for undirected graphs) or their parents and co-parents (directed graphs). If the parent-child relationships preserve conjugacy of the corresponding distributions which are assumed to belong to the exponential family, then the full conditional distributions arising in Gibbs sampling will have the same

functional form as the original conditional distributions (conditioned on the parents) defining each node. In the case of LDA [Griffiths and Steyvers, 2004] implemented using Gibbs sampling, this very property, independence of observations and conjugacy of Dirichlet and Multinomial distributions lead to extremely simple expressions for the updates of the posterior distribution over the parameters and hidden variables (see Algo. 3).

We now briefly comment upon the connection between the EM algorithm and Gibbs sampling. Let us consider a model with hidden variables \mathbf{Z} , observed variables \mathbf{X} and parameters Θ . The functional that is optimized w.r.t. Θ in the M-step is the expected complete data log likelihood given by:

$$Q(\Theta, \Theta^t) = \int p(\mathbf{Z}|\mathbf{X}, \Theta^t) \ln p(\mathbf{Z}, \mathbf{X}|\Theta) d\mathbf{Z} \quad (2.137)$$

If we draw samples $\{Z^{(l)}\}$ from the current estimate of the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \Theta^t)$, with t being the current time step, then we have the following:

$$Q(\Theta, \Theta^t) \approx \frac{1}{L} \sum_{l=1}^L \ln p(\mathbf{Z}^l, \mathbf{X}|\Theta) \quad (2.138)$$

The Q functional can then be optimized in the usual way in the M-step. This is the so-called *Monte Carlo EM Algorithm*. In general we have the following algorithm for Gibbs sampling:

IP (Imputation Posterior) algorithm:

I Step. We wish to sample from the posterior distribution of \mathbf{Z} given \mathbf{X} . However we cannot do this directly since the parameters are coupled with the hidden variables through the observations. We have:

$$p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z}|\Theta, \mathbf{X})p(\Theta|\mathbf{X})d\Theta \quad (2.139)$$

So, for $l = 1, \dots, N$ we first draw a sample Θ^l from the current estimate for $p(\Theta|\mathbf{X})$, and then using this to draw a sample \mathbf{Z}^l from $p(\mathbf{Z}|\Theta^l, \mathbf{X})$

P Step. Given the relation:

$$p(\Theta|\mathbf{X}) = \int p(\Theta|\mathbf{Z}, \mathbf{X})p(\mathbf{Z}|\mathbf{X})d\mathbf{Z} \quad (2.140)$$

we use the samples obtained $\{\mathbf{Z}^l\}$ in the I step to compute a revised estimate of the posterior distribution over θ as:

$$p(\Theta|\mathbf{X}) \approx \frac{1}{L} \sum_{l=1}^L p(\Theta|\mathbf{Z}^l, \mathbf{X}) \quad (2.141)$$

The I step is called the Imputation step since some latent factor is being ascribed to an observation through posterior inference i.e. *filling in* the missing values. This is similar to the E step in the EM algorithm. The naming of the Posterior step is obvious due to the estimation of the posterior distribution over Θ .

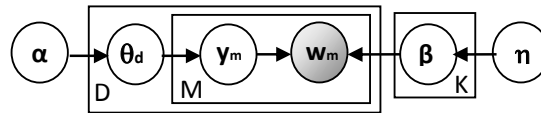


Figure 2.6: LDA with hyperparameters over topic multinomials

Let us revisit the generative process of LDA (see Figure 2.6) as mentioned before in Section 2.7 but

this time with priors $\boldsymbol{\eta}$ over the topic multinomials as well. As before let $\boldsymbol{\alpha}$ be a K -dimensional prior parameter, and let topics $\boldsymbol{\beta}_{1:K}$ be K multinomials over a fixed vocabulary of words. We now assume a prior for $\boldsymbol{\beta}_k$ denoted by $\boldsymbol{\eta}$ which can be a scalar when the Dirichlet is symmetric or can be V -dimensional if we assume an asymmetric Dirichlet. The number K is the same as the number of latent topics and equals the dimensionality of the topic indicator variable $z_{d,n}$. LDA assumes that a document d with N_d words arises from the following generative process:

For each topic $k \in \{1, \dots, K\}$,
 Draw $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\eta})$
 For each document $d \in [1, \dots, D]$,
 Draw $\boldsymbol{\theta}_d | \boldsymbol{\alpha} \sim \text{Dir}(\boldsymbol{\alpha})$
 For each position $m \in \{1, \dots, M_d\}$ in document d :
 Draw topic assignment $z_{d,m} | \boldsymbol{\theta}_d \sim \text{Mult}(\boldsymbol{\theta}_d)$
 Draw word $w_{d,m} | \{z_{d,m}, \boldsymbol{\beta}_{1:K}\} \sim \text{Mult}(\boldsymbol{\beta}_{z_{d,m}})$

In the learning phase, the principal parameters of the model that are learnt are the K topic multinomials $\boldsymbol{\beta}_{1:K}$, each with $V - 1$ independent parameters with V being the size of the vocabulary. Learning these parameters also lead to learning document specific distribution over topics which in turn are obtained from the individual word distributions of each document over the topics. The priors $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are usually optimized using a non-linear optimization strategy using directional derivatives such as choosing a Newton’s direction with inexact line search for step size optimization. In the test set, thus, the vocabulary remains the same only the word counts (may) change for each document. It is important to note that the marginal of the data distribution given the parameters is intractable to compute due to K^V possible configurations one of which fits the data best. Although VB-EM settles on finding a lower bound to the exact log likelihood function of the data, sampling algorithms tend to find samples from the exact posterior distribution of the hidden variables and parameters thereby finding samples that lead to globally optimal parameter estimation.

To this end, Gibbs sampling has been popularized for the LDA model by Griffiths et al. [Griffiths and Steyvers, 2004]. Gibbs sampling is a special case of Markov-chain Monte Carlo (MCMC) simulation. MCMC methods avoid random walk behavior and can emulate high-dimensional probability distributions $p(\mathbf{X})$ by setting up Markov chains to sample from those desired distributions which respect the invariance and ergodic properties. This means that one sample is generated for each transition in the chain after a stationary state of the chain has been reached, which happens after a so-called “burn-in period” which eliminates the auto-correlation of the initial samples thereby removing the influence of initialization parameters.

Gibbs sampling is a special case of MCMC where the dimensions x_i of the distribution are sampled alternately one at a time, conditioned on the values of all other dimensions, which we denote x_{-i} . The algorithm is shown in Algo. 3 where the counts $n_{*, -i}^*$ indicate that the datum i is excluded from the corresponding document or topic.

Referring to the Imputation Posterior scheme of sampling (see Algo. 2.7.6), the I step involves ascribing a state for variables $z_{d,n}$ and hence the per document proportions required to compute $\theta_{d,k}$ and the P step then fills in the values for the $\boldsymbol{\beta}_{1:K}$ parameters. It is standard practice [Neal, 2000] to integrate out the hidden variable for topic proportions $\boldsymbol{\theta}$ and the parameters $\boldsymbol{\beta}$ using the priors $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ respectively (see Fig. 2.6) and exploiting the Dirichlet-Multinomial conjugacy. This is so because they

can be interpreted as statistics of the associations between the observed $w_{d,n}$ and the corresponding $z_{d,n}$ which are the state variables of the Markov chain. The inference task here is to sample from the posterior under z given the observed \mathbf{w} s i.e. to find $p(\mathbf{Z}|\mathbf{w})$.

In Gibbs sampling for LDA, we can derive the full conditional distribution for a word token with index $i = (d, n)$, using the chain rule of probability and conditional independence of random variables in a directed graphical model [Shachter, 1998]. We note that $\mathbf{w} = \{w_i = v, \mathbf{w}_{-i}\}$ and $\mathbf{Z} = \{z_i = k, \mathbf{Z}_{-i}\}$. This yields:

$$p(z_i = k | \mathbf{Z}_{-i}, \mathbf{w}) = \frac{p(\mathbf{w}, \mathbf{Z})}{p(\mathbf{w}, \mathbf{Z}_{-i})} = \frac{p(\mathbf{w} | \mathbf{Z})}{p(\mathbf{w}_{-i} | \mathbf{Z}_{-i}) \cdot p(w_i)} \cdot \frac{p(\mathbf{Z})}{p(\mathbf{Z}_{-i})} \quad (2.142)$$

where the expression $\frac{p(\mathbf{w} | \mathbf{Z})}{p(\mathbf{w}_{-i} | \mathbf{Z}_{-i}) \cdot p(w_i)}$ arises due to the fact that $w_i \perp\!\!\!\perp \mathbf{w}_{-i} | \mathbf{Z}_{-i}$ and $p(w_i) = 1$. Note the striking similarity between Eqs. 2.142 and 2.117.

Algorithm 3 Gibbs Sampling for LDA

```

1: ▷ Initialisation
2: zero all count variables,  $n_d^k; n_d; n_k^w; n_k$ 
3: for documents  $d \in [1, \dots, D]$  do
4:   for words  $n \in [1, \dots, N_d]$  in document  $d$  do
5:     sample topic index  $z_{d,n} = k \sim Mult(1/K)$ 
6:     increment document-topic count:  $n_d^k + 1$ 
7:     increment document-topic sum:  $n_d + 1$ 
8:     increment topic-term count:  $n_k^w + 1$ 
9:     increment topic-term sum:  $n_k + 1$ 
10:   end for
11: end for
12: ▷ Gibbs sampling over burn-in period and sampling period
13: while not finished do
14:   for all documents  $d \in [1, \dots, D]$  do
15:     for all words  $n \in [1, \dots, N_d]$  in document  $d$  do
16:       ► for the current assignment of a topic  $k$  to a term  $w$  for word  $w_{d,n}$ :
17:       decrement counts and sums:  $n_d^k - 1; n_d - 1; n_k^w - 1; n_k - 1$ 
18:       ► multinomial sampling according to  $\frac{n_{k,-i}^w + \eta_w}{\sum_{w=1}^V n_{k,-i}^w + \eta_w} \cdot \frac{n_{d,-i}^k + \alpha_k}{[\sum_{k=1}^K n_{d,-i}^k + \alpha_k] - 1}$  (decrements from
19:       previous step; see [Griffiths and Steyvers, 2004] for details):
20:       sample topic index  $\hat{k} \sim p(z_i | \mathbf{Z}_{-i}, \mathbf{w})$ 
21:       ► use the new assignment of  $z_{d,n}$  to the term  $w$  for word  $w_{d,n}$  to:
22:       increment counts and sums:  $n_d^{\hat{k}} + 1; n_d + 1; n_k^w + 1; n_{\hat{k}} + 1$ 
23:     end for
24:   end for
25:   ▷ Check convergence (often using some heuristics) and output parameters
26:   if converged and L sampling iterations have been exhausted since last output then
27:     ► the different settings of the parameters are either averaged or the one yielding the best model
28:     log likelihood is chosen
29:     output parameter  $\beta$  according to  $\beta_{k,w} = \frac{n_k^w + \eta_w}{\sum_{w=1}^V n_k^w + \eta_w}$ 
30:     output hidden variables  $\theta$  according to  $\theta_{d,k} = \frac{n_d^k + \alpha_k}{\sum_{k=1}^K n_d^k + \alpha_k}$ 
31:   end if
32: end while

```

Variational methods are advantageous over sampling when latent variable pairs are not conjugate.

Usually this involves obtaining another expression for the Expected Lower Bound (ELBO) to the log likelihood functional using inequalities with Taylor expansions of functions and introduction of auxiliary variables (for e.g. see [Zhu et al., 2006]) and then optimizing the new lower bound instead. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods.

Additionally, Gibbs sampling for a topic modeling framework has a few disadvantages when it comes to parallelization [Zhai et al., 2012]. A major bottleneck for Gibbs sampling in parallel environment is distributed memory which makes synchronization of non-local data items such as the number of times a word type appears in a topic across all documents very difficult. Gibbs sampling for parallelization of LDA thus meets a trade-off between easier formulation of updates and complex engineering solutions needed to synchronize global counts to rectify for inconsistencies even if the problem is inherently document parallelizable.

A potential incentive for using Gibbs sampling formulation for creating new models that make use of LDAs modularity is the use of very short and simple iterations. For each word, there is a simple multiplication to build a sampling distribution of length K , sampling from that distribution, and updating an integer vector.

Sampling from a K -dimensional multinomial distribution with parameters $\theta_1, \theta_2, \dots, \theta_K$ is easy [Bishop, 2006]. We first choose a random number $r \in [0, 1]$ and then choose an indicator k such that $k = j : \arg \max_j \sum_{i=1}^j \theta_i \leq r$.

A K -dimensional Dirichlet distribution is a distribution belonging to the exponential family and is conjugate to the multinomial distribution. Its functional form is given by $p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}$ where $\sum_{k=1}^K \theta_k = 1$. To sample a random vector $\{\theta_1, \dots, \theta_K\}$ from a Dirichlet distribution with parameters $\{\alpha_1, \dots, \alpha_K\}$ requires a bit more effort and is done in the following way [Gelman et al., 2003]:

[i] Draw K independent random samples $\{y_1, \dots, y_K\}$ from Gamma distributions each with density

$$\gamma(\alpha_k, 1) = \frac{y_k^{\alpha_k - 1} \exp(-y_k)}{\Gamma(\alpha_k)} \quad (2.143)$$

[ii] Set $\theta_k = y_k / \sum_{k=1}^K y_k$

The Gamma distribution $\Gamma(\alpha, \beta)$ is defined by the following probability density:

$$\gamma(y, \alpha, \beta) = \frac{y^{\alpha - 1} \exp(-y/\beta)}{\beta^\alpha \Gamma(\alpha)} \quad (2.144)$$

The parameter $\alpha > 0$ is responsible for the shape of the distribution graph and called the *shape parameter* and $\beta > 0$ is called the *scale parameter*. There is an alternative commonly-used parametrization of the Gamma distribution denoted by $\Gamma(a, b)$ with pdf $f(y, a, b) = \frac{y^{a-1} \exp(-by)b^a}{\Gamma(a)}$. In this context, $b = 1/\beta$ is called the *rate parameter*. One important property of the Gamma distribution is that the sum of i.i.d $\Gamma(\alpha_k, \beta)$ random variates is a $\Gamma(\sum_k \alpha_k, \beta)$ random variate. Sampling from a Gamma distribution is a bit more convoluted starting with samples from uniform distribution (due to the non-integral values of the shape and scale parameters). Concrete implementations of classical algorithms for doing that can be found in [Press et al., 2007].

In contrast, each iteration of variational inference requires the evaluation of complicated functions that are not directly implemented in hardware. However, variational methods typically takes much less

number of iterations (in the order of tens compared to hundreds or thousands for Gibbs sampling) to converge and convergence is also easier to assess. It also enjoys the full benefits of local document specific calculations within the E step which can be easily be done in the Map phase of a Map-Reduce framework. On the other hand, a larger number of iterations for Gibbs sampling means that much more communication overhead for synchronization to maintain consistent global counts. The complete marginal independence assumptions in the dual graph for LDA using the variational Bayes approach makes the inference machinery easily amenable to document level parallelization with less synchronization overheads and thus may be preferred over the Gibbs sampling version in a distributed setting where maintaining consistency of counts across machines in the latter case causes runtime performance degradation due to many more synchronization steps. The only difference is that the Gibbs sampling scheme will sample from the exact posterior while VBEM will only find ML or MAP estimates of the approximate zero forcing modes of the true posterior modes.

Chapter 3

Learning to Summarize using Sparse Coherence Flows

*“Words, like nature, half reveal and half conceal the soul within.” - Alfred Lord
Tennyson*

3.1 Introduction

The single document summarization problem has been studied as early as the 1960s [Luhn, 1958, Edmundson, 1968] and recent papers on multi-document summarization using the basic results obtained in [Nenkova et al., 2006b] have shown that high frequency salient words (not including function words such as stopwords) in the input documents are covered to a large extent in the human summaries. Most summarization systems are extractive i.e. full sentences from the input documents are selected that maximize the coverage of salient words without sacrificing readability. Recently it has been also shown that solving the problem of multi-document summarization exactly in polynomial time is NP-hard i.e., simultaneously satisfying the constraints of relevancy to query, non-redundancy within the summary and total summary length is NP-hard [McDonald, 2007]. Adding local coherence flow to this list of constraints not only calls for even more complex optimization but also needs to quantify coherence in some fashion.

Topic models like LDA [Blei et al., 2003] have become the cornerstone for understanding the thematic organization of large text corpora in a completely unsupervised but robust fashion. The focus of this chapter is to extend LDA for extractive query focused multi-document summarization. We hypothesize that in addition to a bag of words, a document can also be viewed in a different manner—words in a sentence always carry syntactic and semantic information and often such information (for e.g., the Grammatical and Semantic Role (henceforth GSR) of a word such as subject, object, noun and verb concepts etc.) is carried across adjacent sentences to enhance coherence in different parts of a document.

In the realm of computational linguistics, there has been work in Centering theory including those by Grosz et al. [Grosz et al., 1995]. Their work mainly specifies how discourse interpretation depends on interactions among speaker intentions, attentional state and linguistic form. The discourse participants’ focus of attention at any given point in time is modeled by their “attentional state” which comprises of a *focus* in the current utterance being understood. This focus within the attentional state helps identify

“centers” of utterances which relate different parts of local discourse segments meaningfully and according to [Grosz et al., 1995], the “centers” are semantic objects, not just words, phrases, or syntactic forms. Centering theory helps formalize the constraints on the centers to maximize local coherence properties of a discourse. In the context of this chapter, the grammatical and semantic roles are approximated by the explicit realization of the roles induced by the sentential words such as a parts-of-speech, named entity classes etc. We augment this representation to include a *lemmatized* surface form of the word as well in chapter 5. In most datasets where there are no ground truth annotations of the words, the explicit realization of the GSRs are induced through some inference mechanism employed under the umbrella of structured prediction techniques. To better understand attentional state and centering, consider the following simple example:

- Discourse 1
 1. Martha finally went to visit Martha’s_Vineyard.
 2. She had been hoping to spend her vacation at Martha’s_Vineyard for many years.
 3. Martha packed her things and prepared to leave.
 4. She was excited to see the shores of Martha’s_Vineyard as she was boarding the ferry.
- Discourse 2
 1. Martha finally went to visit Martha’s_Vineyard.
 2. Martha’s_Vineyard had been her top choice for spending a vacation for many years.
 3. Martha packed her things and prepared to leave.
 4. The shores of Martha’s_Vineyard made Martha excited as she was boarding the ferry.

Discourse 1 is an example where the focus of attention is clearly on Martha. Discourse 2 highlights a shift of attention from Martha to Martha’s_Vineyard and vice versa. For e.g. in the first utterance if a reader perceives the focus of attention to be Martha’s_Vineyard, there is a retention of the focus in the second utterance. If, however, in the first utterance the focus of attention be Martha, then there is a focus shift in the next utterance. In any case, the focus is on Martha in the third utterance. Discourse 2 is thus less coherent than discourse 1 in terms of the effort to understand the discourse i.e. discourse 1 has less *inference load*.

In the words of Grosz et al. [Grosz et al., 1995], “It is well known from the study of complexity theory that the manner in which a class of problems is represented can significantly affect the time or space resources required by any procedure that solves the problem. Here too we conjecture that the manner, i.e., linguistic form, in which a discourse represents a particular propositional content can affect the resources required by any procedure that processes that discourse. We use the phrase *inference load* placed upon the hearer to refer to the resources required to extract information from a discourse because of particular choices of linguistic expression used in the discourse.”

In the example above, in discourse 1, the pair (Subject, “Martha”) approximates a center that is retained through the focus of attention in the utterances. Thus the propagation of these centers of utterances within discourse segments helps maintain the local coherence, which in turn is responsible for “*easy understanding*” of the discourse i.e. reducing the inference load on part of the reader.

Each word in a sentence of a document has an associated role (syntactic or semantic) with it for e.g., a noun helps identify a concept (abstract or concrete) and thus serves as a part or whole of a center of utterance. If two consecutive sentences contain the same word, then there is a GSR transition (henceforth **GSRt**) within the context of sentences. The change in attentional state in local discourse segments are approximated through these transitions. If the word is not present in the preceding (or

succeeding) sentence then there is still a transition from (to) a *null*, identified by “–” GSR to (from) the current GSR. A GSRt can thus be looked upon as a multinomial distribution over sentences in a document. Although only the use of entities are advocated by centering theory, verbs have also been used as GSRs in this chapter to understand the intents in attention i.e. actions.

Following the original versions of the Centering theory as is, the most coherent parts of a discourse are those where there are frequently occurring GSRts involving entities i.e. Subjects, Objects and Named-Entity GSRs. Selecting sentences which have a high frequency of such GSRts is helpful, however, it is not guaranteed that such a property is dominant in a general document collection.

Interestingly the implications of centering bear remarkable similarity to summarizing videos. Amongst the many objects present in each frame of the video, the observer is focused on the principle actions and the associated entities and objects which help generate a textual query for that video. This phenomenon is briefly implied in Figure 1.1 in Chapter 1 where the central question becomes “*Do we speak all that we see?*”

Section 3.3 and the sub-sections within it review how Centering theory [Grosz et al., 1995] has been adapted to develop our new Utterance Topic Model (UTM) [Das and Srihari, 2009] to include the notion of coherence as an auxiliary meta-perspective of a document which is assumed to influence its latent topical structure. In section 3.5 it is shown how UTM can be extended as a full summarization model. For a particular query, we rank sentences by a product of topical salience and as well as the topical influence over observed GSR transitions. The techniques of our proposed method are described in section 3.4 and results and analysis of the output of our model in terms of summarization are presented in Section 3.6.2.

3.2 Related Work

Topic models have been widely applied to text despite a willful ignorance of the underlying deep linguistic structures that exist in natural language due to computational time constraints. There have been a lot of work on either applying topic models directly to a certain problem as in [Blei and McAuliffe, 2008, Yano et al., 2009, Chen et al., 2009] or adapting basic LDA style topic modeling as in [Nallapati and Cohen, 2008, Mei et al., 2007a]. In a topic model, the words of each document are assumed to be exchangeable i.e., their probability is invariant to permutation of the positions of the words in a document. A workaround to this inadequacy was posed and addressed in [Graber and Blei, 2009]. It is also important to note that although a topic model *may* suggest documents relevant to a query (by treating the query as a short document), it can be very noisy due to the length of the query and further, finding particularly relevant phrases for question answering is still a challenging task. Our main focus in this chapter has been to build a new topic model based on the LDA framework which can use linguistic features and semantic roles of words in a discourse segment.

In the active research area of extractive multi-document summarization, most earlier methods had focused on clustering of sentences or building graphs of sentences from the relevant documents and then using graph mining algorithms such as Pagerank to select most authoritative sentences (as in [Wei et al., 2010]). Other approaches include algorithmic formulation of summary extraction using greedy, dynamic and integer linear programming methodologies. The work in [McDonald, 2007] compares these approaches and also proves that in general the inferring an extractive summary is NP-hard. More earlier works in summarization had focused on core natural language processing techniques and propositional

logic [Marcu, 2000b] however, all such methods lack exploratory topic analysis of the corpus.

Although many articles on summarization exist in the proceedings of different conferences and journals, one of the best places to find interesting ideas on this problem is in the proceedings for the Document Understanding Conference (DUC) and the Text Analysis Conference (TAC) [DUC, 2007, DUC, 2008]. The approaches of a few of the TAC 2008 teams who participated in the summarization track and obtained good results are discussed very briefly below.

The Thomson Reuters team¹ used a classifier based approach to identify good first sentences that are deemed ideal for summary construction. Note that this assumption may be valid if the data belongs to the newswire domain. Since the TAC datasets mainly consist of news articles, this assumption is valid. The ICSI Summarization system [Gillik and Favre, 2009] formulates the summarization problem as an integer linear programming problem where they maximize the coverage of n -grams (bi-grams in their paper) in summary sentences. In this setting, the goal is to assign binary integers (0 or 1) α_{ij} to a sentence s_j consisting of a concept (n -gram) c_i to indicate whether sentence s_j should really be a candidate summary sentence. This is achieved by formulating a set of constraints involving weights of concepts, the latter being calculated using its importance to the document collection as well as to the query. In both ILP formulations mentioned in [Gillik and Favre, 2009, McDonald, 2007], the inclusion of the concept of coherence is completely missing.

The system presented by Tsinghua University [Long et al., 2009] presents a new summarization technique using Kolmogorov complexity and information distance. However to approximately compute the information distances between n the summary units (sentences) they use simple weights of words like *tf-idf* etc. and functions involving these weights. They also used another method which used ranking of sentences through eigenvalue calculation of the sentence similarity matrix that also performs comparably to their former method. The system submitted by Zhang et al.² again focuses on identifying salient n -grams to be included in a summary and rank sentences based on such weights.

The work by Ye et. al. [Ye et al., 2005] calculates the semantic similarity among the sentences in the cluster, and between a sentence and the given cluster of documents. The semantic similarity between sentences is determined by the number of sentential concept overlaps calculated from the WordNet synset hierarchy including glosses, hypernyms and meronyms. Another interesting approach taken by [Li et al., 2005] where the sentences are scored by a weighted combination of several features including pattern based features which provide clue as to how to interpret an information need. It has been shown in [J et al., 2005], that using contextual language models and latent semantic indexing, the resulting summaries has been promising based on the results of the ROUGE evaluation tool. Their contextual language model essentially computed a language model within a window of words instead of an explicit n -gram. In yet another unique approach to summarization [Hovy et al., 2005], the syntactic structure of the parse trees was utilized to generate valid triples of basic elements (BEs) or (head—modifier—relation) triples and then summary sentences were extracted using a score directly related to computing important BEs in them. The focus in [Srihari et al., 2007] was more about finding hidden connections among query concepts using textual evidences through semantic cues rather than summarization. However, a final summarization was performed on the evidence trails and was therefore chosen as a system for comparison.

Some of the recent and notable Bayesian topic model approaches to summarization have been pre-

¹<http://www.nist.gov/tac/publications/2008/participant.papers/TOC.proceedings.pdf>

²<http://www.nist.gov/tac/publications/2008/participant.papers/ICTCAS.proceedings.pdf>

sented in [Chen et al., 2009] and [Daumé III and Marcu, 2006] that involve exploratory topic analysis.

3.3 Centering Theory and Sparse Coherence Flows

This section presents our preliminary attempts to model summarization as an unsupervised learning problem without resorting to complex question answering techniques using inference with parse trees. Although part-of-speech (POS) as well as syntactic information has been used, the proposed model ignores any syntactic tree structure of sentences at present unlike that in [Graber and Blei, 2009] which ignores inter-sentential clues of entity propagation from Centering theory. This has been done to model documents and sentences using the syntactic and semantic features of words *across* adjacent sentences and not *within* sentences.

3.3.1 Discourse Analysis: Centering Theory

Centering theory in discourse analysis comprises of a family of models which attempt to analyze the notion of “centering” [Grosz et al., 1995] with regards to *coherence*. Discourse structure mainly deals with modeling of the inherent linguistic structure, intentional structure and the local attentional state. Each of these modules manifest themselves in various forms depending on the inherent assumptions of the modeling choices. For example, one can derive both a dependency parse or a rhetorical parse from a syntactic parse where the inherent assumptions of each parsing model is very different (see Chapter 5 for an introduction on Rhetorical Structure Trees).

The syntactic parse is involved in finding the hierarchy of non-terminals in the grammar of a specified language that leads to the generation of an utterance or a sentence. The dependency parse often uses the syntactic parse to decompose an utterance or a sentence into a set of binary relation tuples where a pair of words are connected through a labeled relation identifying a governing constituent as well as a dependent constituent. Rhetorical structure [Mann and Thompson, 1988] on the other hand is involved in segmenting text into Elementary Discourse Units (EDUs) and finding labeled binary relations that joins two EDUs in a meaningful way. The meaning in this case originates from a set of pre-specified (albeit incomplete) rhetorical relations. Although, the incompleteness of such relations [Sibun, 1993] can be a cause of concern for deep analysis of discourse, however, in this chapter we are not concerned with such analysis and instead take a coarser view of the underlying discourse structures. This allows us to build statistical models which decouple deep linguistic assumptions from necessary exchangeability conditions on the observations thus providing an option for fast inference. Rhetorical Structure also models the inherent intentional structure as manifested through the rhetorical relations between adjoining constituents—we will visit Rhetorical Structure Trees in Chapter 5 in the context of “bulleted list” summary generation. We first provide a background on “centering” in the context of anaphora (or co-reference) resolution that will help us understand its importance with regards to the creation of controlled vocabularies or meta-information from documents.

Centering Theory [Grosz et al., 1995] concerns itself with modeling the attentional state in discourse analysis. There are two levels of attentional state—a global level, which is dependent on the intentional structure, is concerned with the relations between discourse segments and the ways in which attention shifts between them; The local level, on the other hand, is concerned with changes of “focus of attention” within discourse segments. **Centering**, an element of the local level, pertains to the interaction between

	$C_b(U_{n+1}) = C_b(U_n)$ or $C_b(U_n)$ is undefined	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

Table 3.1: Transition relations holding between pair of adjacent sentences due to the centers

the form of linguistic expression and local discourse coherence. In particular, it relates local coherence to the choice of referring expressions such as pronouns contrasted with definite descriptions or proper names, and argues that differences in coherence correspond in part to the different demands for inference made by different types of referring expressions, given a particular attentional state. We will explore an example from anaphora resolution i.e. co-reference resolution using a *Centering algorithm* to highlight this point.

In centering theory, utterances make up the fundamental constituents. In general, utterances can be full short sentences but can easily be smaller meaningful constituents of a larger sentence. Also, the attentional state in Centering Theory is reflected through the participation of entities within the loci of attentional focus across utterances. If we denote U_n and U_{n+1} to be adjacent utterances, the **backward-looking center** of U_n , denoted as $C_b(U_n)$, represents the entity currently being focused on *in the discourse* after U_n is interpreted. The **forward-looking centers** of U_n , denoted as $C_f(U_n)$, form an ordered list containing the entities mentioned in U_n , all of which can serve as $C_b(U_{n+1})$. In general, however, $C_b(U_{n+1})$ is the most highly ranked element of $C_f(U_n)$ mentioned in U_{n+1} . The C_b of the first utterance in a discourse is undefined. Brennan et al. [Brennan et al., 1987] uses the following ordering: Subject > Existential predicate nominal > Object > Indirect object or oblique > Demarcated adverbial PP. An example of each of the semantic roles of the noun concept “Ford Focus” is as follows:

- (a) A *Ford Focus* is parked in the lot. [*subject*]
- (b) There is a *Ford Focus* parked in the lot [*existential predicate nominal: The existential phrase is marked with a “There”*]
- (c) John parked a *Ford Focus* in the lot [*object*]
- (d) John gave his *Ford Focus* a facelift. [*indirect object: A noun or pronoun that indicates to whom or for whom the action of a verb in a sentence is performed*]
- (e) Inside his *Ford Focus*, John showed Susan his new CD player. [*demarcated adverbial PP*]

We now describe a centering based algorithm for pronoun resolution from [Brennan et al., 1987]. In their algorithm, referents of pronoun are computed from relations that hold between the forward- and backward-looking centers in *adjacent sentences*. The algorithm defines four inter-sentential relations between a pair of utterances U_n and U_{n+1} that depend on the relationship between $C_b(U_{n+1})$, $C_b(U_n)$ and $C_p(U_{n+1})$ where $C_p(U_n)$ is the highest ranked forward-looking center in U_n .

The following rules are used by the algorithm:

- Rule 1:** If any element of $C_f(U_n)$ is realized by a pronoun in utterance U_{n+1} , then $C_b(U_{n+1})$ must be realized as a pronoun also.
- Rule 2:** Transition states are ordered. Continue is preferred to Retain is preferred to Smooth-Shift is preferred to Rough-Shift.

With these rules and concepts the algorithm is as follows:

1. Generate possible $C_b - C_f$ combination for each possible set of reference assignments
2. Filter by constraints, e.g., syntactic coreference constraints, selectional restrictions, centering rules

and constraints

3. Rank by transition orderings.

The result is that the pronominal referents that get assigned are those that yield the most preferred relation in Rule 2 without violating Rule 1 or coreference constraints (gender, number, syntactic, selectional restrictions).

We now give an example of this approach where an error in pronoun resolution leads to a considerable higher inference load on part of the reader.

U_1 : Bob opened a new dealership last week.

U_2 : John took a look at the Fords in his lot.

U_3 : He ended up buying one.

The question here is to which entity should “he” in the utterance U_3 refer to so that the inference load i.e. the effort to understand the paragraph is minimum.

We have the following for the first sentence U_1 :

1. $C_f(U_1)$: {Bob, dealership}
2. $C_p(U_1)$: {Bob}
3. $C_b(U_1)$: undefined

Similarly, for the second sentence U_2 we have:

1. $C_f(U_2)$: {John, Fords}
2. $C_p(U_2)$: {John}
3. $C_b(U_2)$: {Bob} [*because of “his lot” and $C_p(U_1)$ is “Bob”*]

Since $C_b(U_2) \neq C_p(U_2)$ and $C_b(U_1)$ is undefined, we have a “Retain” relation between U_1 and U_2 . Now we wish to determine the referent for the subject pronoun “he” in U_3 .

If **he** refers to *John*, then we have:

Case1

1. $C_f(U_3)$: {John}
2. $C_p(U_3)$: {John}
3. $C_b(U_3)$: {John} (whichever entity “he” is referring to)

If **he** refers to *Bob*, then we have:

Case2

1. $C_f(U_3)$: {Bob}
2. $C_p(U_3)$: {Bob}
3. $C_b(U_3)$: {Bob} (whichever entity “he” is referring to)

In case 1, we have a “Smooth-shift” relation between U_2 and U_3 since $C_b(U_3)$ is “John” and the most highly ranked element of $C_f(U_2)$ i.e. $C_p(U_2)$ is “Bob.” Thus $C_p(U_3) = C_b(U_3) \neq C_b(U_2)$. However for case 2, we have a “Continue” relation which is preferred over “Smooth-shift” and thus “he” in U_3 is wrongly associated with “Bob.” This phenomenon not only makes the sequence of utterances difficult to understand but also makes Bob a more prominent entity in the discourse segment than John through frequency measurements. From the perspective of incorporating such features into a frequency based statistical model, such wrong substitutions can result in assigning higher importance to unimportant entities. We thus, initially, take a different approach to incorporate such local attentional state information relating to local coherence as observed variables in topic models which we discuss next.

The term *centers* of an utterance refer to those entities serving to link that utterance to other utterances in the discourse segment that contains it. Each utterance, which can be coarsely approximated

by a sentence, S in a discourse segment (DS) is assigned a set of forward-looking centers, $C_f(S, DS)$ and each utterance other than the segment initial utterance is assigned a *single* backward-looking center, $C_b(S, DS)$. The backward-looking center of utterance S_{n+1} connects with one of the forward-looking centers of utterance S_n . An illustration from [Grosz et al., 1995] below elucidates coherence through such center linkages. Such center linkages constitute what we name as “Sparse Coherence Flow.”

- (a) John has been having a lot of trouble arranging his vacation
- (b) He cannot find anyone to take over his responsibilities. (he = John) $C_b = \text{John}$; $C_f = \{\text{John}\}$
- (c) He called up Mike yesterday to work out a plan. (he = John) $C_b = \text{John}$; $C_f = \{\text{John}, \text{Mike}\}$

3.3.2 Sparse Coherence Flows

For building a statistical topic model that incorporates GSR transitions (henceforth GSRts) across utterances, words in a sentence were attributed with GSRs like subjects, objects, concepts from WordNet [Fellbaum, 1998] synset role assignments (wn), adjectives, VerbNet [Kipper, 2005] thematic role assignment (vn), adverbs and “other” (if the feature of the word does not fall into the previous GSR categories). Further if a word in a sentence is identified with two or more GSRs, only one GSR is chosen based on the left to right descending priority of the categories mentioned. In our experiments, we enforce the following ordering priority: $\langle \text{subj} > \text{nn} > \text{wn} > \text{obj} > \text{adj} > \text{vb} > \text{vn} > \text{adv} \rangle$. These roles (GSRs) have been extracted separately using the text analytics engine SemantexTM from Content Savvy Inc. (previously Janya Inc.).

In a window of sentences, there are potentially $(G + 1)^2$ GSRts for a total of G GSRs with the additional one representing a null role (denoted by “-”) if the word is not found in the contextual sentence. It is easy to see that with the introduction of the notion of the null GSR, the number of GSRts involving non-null GSRs is much much fewer than with one of the GSRs being null and arises due to suppression of a coreference resolution module. Referring back to the Anaphora resolution example in the previous section, the term “sparse” is used to signify a deliberate ignoring of co-reference resolution of the pronouns during document processing thus making the statistical topic model much less prone to errors arising out of wrong anaphora resolution.

If there are T_G GSRts in the corpus, then a sentence is represented as a vector over the GSRt counts only along with a count vector over the word vocabulary. In the extended model for summarization, one more role called “*ne*” (encompassing all Named Entity classes) has also been added with the highest priority.

Note that although Centering theory focuses more on the nominal entities and pronouns to highlight the transition of attentional state across the forward and backward looking center(s), we also consider words which play key roles in identifying the relationship between a subject and object like verbs and also words that promote/demote the quality of a noun or verb like adjectives and adverbs. This is because words other than nominal entities also play a major role in identifying the intention in the discourse segments.

Also since topic models like LDA work best by eliminating stopwords when a symmetric Dirichlet prior is assumed for the latent topic proportions of documents, some of these stopwords like pronouns etc. are important for signifying coherence and so we used anaphora resolution as offered by SemantexTM to substitute pronouns with the referent nouns as a preprocessing step. However, this turned out to be very noisy and we do not use anaphora resolution when we revisit this type of document perspective in Chapter 5.

For further insight on how GSRts have been used, a matrix has been constructed consisting of sentences as rows and words as columns; the entries in the matrix are filled up with a specific GSR for the word in the corresponding sentence following GSR priorities. Table 3.2 shows a slice of such a matrix taken from the DUC 2005 dataset which contains documents related to events concerning rules imposed on food labeling. Table 3.2 suggests, as in [Barzilay and Lapata, 2005a], that dense columns of the GSRs indicate potentially salient and coherent sentences (1 and 2 here) that present less inference load with respect to a query like “Food Labeling.”

↓SentenceIDs	words... →						
ID	food	consumers	health	confusion	label(ing)	FDA	regulations
1	<i>nn</i>	–	<i>nn</i>	<i>nn</i>	<i>nn</i>	<i>ne</i>	–
2	<i>nn</i>	–	<i>nn</i>	–	–	<i>ne</i>	–
3	–	<i>subj</i>	–	–	–	–	–
4	<i>subj</i>	<i>nn</i>	<i>subj</i>	–	–	–	<i>nn</i>

Table 3.2: Snapshot of a sentence-word GSR grid view of a document on “Health and Safety” category

where “*nn*” is a noun and “*ne*” is a Named Entity category. Sentences 1 through 4 in the document read as:

1. The Food and Drug Administration (FDA) has proposed a stringent set of rules governing the use of health claims on food labels and advertising, ending nearly six years of confusion over how companies may promote the health value of their products.
2. By narrowing standards for what is permissible and strengthening the FDA’s legal authority to act against misleading claims , the rules could curtail a trend in food marketing that has resulted in almost 40% of new products and a third of the \$ 3.6 billion in food advertising over the last year featuring health-related messages.
3. Most such messages are intended to make the consumer think that eating the product will reduce the risk of heart disease or cancer.
4. The regulations, which will be published next week in the Federal Register , were criticized by food industry officials , who said they would require health claims to meet an unrealistic standard of scientific proof and would hinder the ability of manufacturers to give consumers new information about nutrition.

Note that the counts for the GSRts “*nn*→–” and “*nn*→*nn*” for sentenceID 1 are both two in this snapshot. Thus this discourse is dominant in GSRts involving a noun GSR.

For another example on a different event, Table 3.3 shows a slice of the GSR grid obtained from the TAC2008 dataset which contains documents related to events concerning Christian minorities in Iraq and their current status. Table 3.3 again suggests that dense columns of the GSRs indicate potentially salient and coherent sentences (7 and 8 here) that present less inference load with respect to a query like “Baghdad attacks”. Note that in this chapter, the GSRt representation does not include the surface form of the words and thus the words “food” and “health” in Table 3.2 are indistinguishable for the perspective of GSRt proportions in the document. In this case the words are assumed to be generated from latent topics which are influenced by the local but coarse coherence properties of the documents. This type of GSRt can be looked upon as coarse coherence encoding of the discourse. Such an encoding does not turn out to be very effective w.r.t. held-out log likelihood calculations and we remedy this in Chapter 5.

↓SentenceIDs	words... →				
ID	protect	attacks	churches	Baghdad	Mosul
6	–	–	–	–	–
7	–	<i>subj</i>	<i>obj</i>	<i>wn</i>	–
8	–	<i>vn</i>	<i>wn</i>	<i>wn</i>	<i>wn</i>
9	–	<i>subj</i>	–	–	–
10	–	–	<i>subj</i>	<i>wn</i>	–

Table 3.3: Snapshot of a sentence-word GSR grid view of a document on “Attacks” category

where “*wn*” is a WordNet synset role assignment and “*vn*” is a VerbNet thematical role assignment. The sentences 6 through 10 in the document read as:

6. The major Christian groups include Chaldean - Assyrians, who make up Kana’s group, and Armenians.
7. On Oct. 16 , bomb **attacks** targeted five churches in **Baghdad** which damaged buildings but caused no casualties.
8. Officials estimate that as many as 15,000 of Iraq’s nearly one million Christians have left the country since August, when four churches in **Baghdad** and one in Mosul were **attacked** in a coordinated series of car bombings.
9. The **attacks** killed 12 people and injured 61 others.
10. Another church was bombed in **Baghdad** in September.

3.4 Learning to Summarize using Utterance Topic Models

This section describes our first attempt to model topics not only using word counts but also using GSRts. An extension that transforms the utterance topic model to a “Learning To Summarize” (henceforth LeToS) model is also presented.

3.4.1 Utterance Topic Model

The proposed probabilistic graphical Utterance Topic Model (henceforth UTM) is presented now. To describe the document generation process, it is assumed that there are K latent topics, T_G total number of possible GSRts and T GSRts associated with each document. Also denote θ and π to be the topic and topic-coupled GSRt proportions in each document. We treat the GSRt proportions per document in this model to be *topic-coupled* since the expected number of terms per GSRt also depend on their latent topic assignment. Had we treated the GSRts to be word level annotations, then a model like TagLDA [Zhu et al., 2006] would have been applicable. In that case we still have to make a decision which prefers only one GSRt in a list of GSRts for a particular word depending upon its contextual sentences and there is no principled approach to make this choice. In this chapter, we do not associate the surface form of the word in the GSRt representation in this preliminary investigation—a limitation which is refined in Chapter 5.

Denote $r_{d,t}$ to be the observed GSRt t for a particular window of three sentences in a document d ; $w_{d,m}$ is the observed word in the n^{th} position of the d^{th} document. Further denote, $z_{d,t}$ to be an indicator

variable for topic proportions, $y_{d,m}$ is the indicator variable for topic-coupled GSRt proportions. At the parameter level, each topic is a multinomial over the vocabulary V of words in the corpus and each topic is also a multinomial over the GSRts following the implicit relation of GSRts to words within sentence windows. Also these GSRts are the output of a separate natural language parsing system.

At a higher level, each document in the corpus has mixing proportions over both the latent topics and also over the topic-coupled GSRts. In our proposed model, a GSRt along with the topic is jointly responsible for selecting a word from the vocabulary. This intuition becomes clear by observing the GSRts of the words “attack,” “churches” and “Baghdad” in Table 3.3. Without the corresponding GSRts, the topic of “attacks on churches in Baghdad” would not have formed in that particular discourse and the choice of this particular topic led to the realization of GSRts such as $subj \rightarrow wn$, $wn \rightarrow subj$, $wn \rightarrow obj$, $wn \rightarrow wn$ etc. The document generation process is shown in Fig. 3.1a and is explained as a model below:

For each document $d \in 1, \dots, D$

- Choose a topic proportion $\theta | \alpha \sim Dir(\alpha)$
 - Choose topic indicator $z_t | \theta \sim Mult(\theta)$
 - Choose a GSRt $r_t | z_t = k, \rho \sim Mult(\rho_{z_t})$
- Choose a GSRt proportion $\pi | \eta \sim Dir(\eta)$
- For each position m in document d
 - Choose $y_m | \pi \sim Mult(\pi)$
 - Choose a word $w_m | y_m = t, \mathbf{z}, \beta \sim Mult(\beta_{z_{y_m}})$

where $m \in \{1, \dots, M_d\}$ is the number of words in document $d \in \{1, \dots, D\}$, t is an index into one of the T GSRts and k is an index into one of the K topics; β is a $K \times V$ matrix and ρ is a $K \times T_G$ matrix. The model can be viewed as a generative process that first generates the GSRts and subsequently generates the words which describes the GSRts. For each document d , we first generate T topic-coupled GSRts using a simple LDA model and then for each of the M_d word positions, a GSRt is sampled and a word $w_{d,m}$ is drawn conditioned on the same factor which generated the chosen the GSRt. The factor here is indicative of the latent topic. Instead of influencing the choice of a topic-coupled GSRt to be selected from an assumed distribution (e.g. uniform or Poisson) over the number of GSRts, the document specific proportions are used i.e. $\chi_{d,t} - \eta_t$ is the expected number of words assigned to a GSRt t influenced by the generating topics for document d .

Direct posterior inference over the latent variables is intractable because of coupling of the parameters to the latent factors conditioned on the observed variables and we resort to approximate inference through variational Bayes. Variational Bayes breaks the edges between coupled random variables and parameters, removes the observed variables that lead to coupling and introduces free variational parameters which act as surrogates to the causal distribution of the original latent variables. The resulting simpler tractable distribution is shown in Fig. 3.1b. In the variational setting, the constraints for each document are $\sum_{k=1}^K \phi_{d,t,k} = 1$ and $\sum_{t=1}^T \lambda_{d,m,t} = 1$. Note that θ is K -dimensional and π is T_G -dimensional.

3.4.2 Parameter Estimation and Inference

This section outlines the various updates of the latent variables and the parameters. In our model mean field variational inference is used to find as tight as possible an approximation to the log likelihood of the data (the joint distribution of the observed variables given the parameters) by minimizing the KL

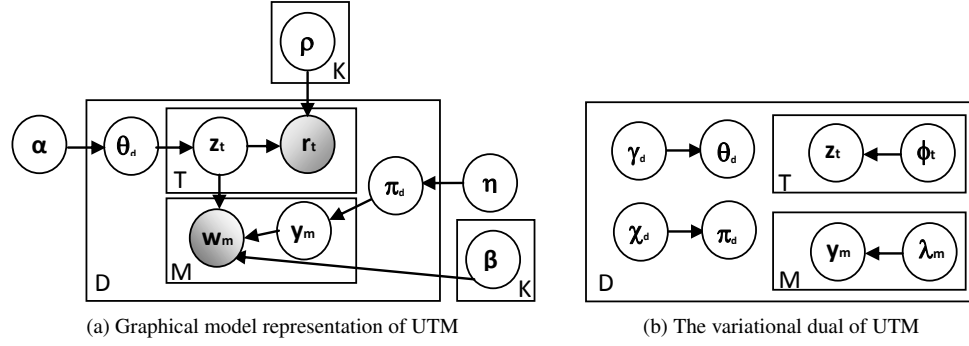


Figure 3.1: Utterance Topic Model: Extending LDA to include coarse coherence properties of discourse segments

divergence of the posterior distribution of the latent variables over the variational parameters to likelihood of the data. The details can be found in [Blei et al., 2003, Beal, 2003]. As discussed in Chapter 1, for tractability purposes, a fully factorized variational distribution over *each document* d is assumed as:

$$q(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y} | \boldsymbol{\gamma}, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda}) = q(\boldsymbol{\theta} | \boldsymbol{\gamma}) q(\boldsymbol{\pi} | \boldsymbol{\chi}) \prod_{t=1}^T q(z_t | \phi_t) \prod_{m=1}^{M_d} q(y_m | \lambda_m) \quad (3.1)$$

The variational functional to optimize can be shown to be:

$$\mathcal{L} = \mathbb{E}_q[\ln p(\mathbf{r}, \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\rho}, \boldsymbol{\beta})] - \mathbb{E}_q[\ln q(\boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{y} | \boldsymbol{\gamma}, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda})] \quad (3.2)$$

where $\mathbb{E}_q[f(\cdot)]$ is the expectation of $f(\cdot)$ under the q distribution.

3.4.3 Latent variable inference

The key inferential problem that is being solved here is to infer the posterior distribution over the latent variables conditional upon the observations and parameter values. As discussed in Chap. 1, the intractable integration problem of computing the log partition function is transformed into a tractable lower bound optimization problem. Figure 3.1b shows the variational parameters of the original model: $\boldsymbol{\gamma}, \boldsymbol{\chi}, \boldsymbol{\phi}, \boldsymbol{\lambda}$; These parameters are defined for *every individual instance* of the latent variables over which the integral of the log partition function is defined. The maximum likelihood estimators of these latent variables have the following forms:

$$\gamma_{d,k} = \alpha_k + \sum_{t=1}^T \phi_{d,t,k} \quad (3.3)$$

$$\chi_{d,t} = \eta_t + \sum_{m=1}^{M_d} \lambda_{d,m,t} \quad (3.4)$$

$$\lambda_{d,m,t} \propto \exp\left\{(\Psi(\chi_{d,t}) - \Psi(\sum_{f=1}^T \chi_{d,f})) + (\sum_{i=1}^K \phi_{d,t,i} \ln \beta_{z_{(y_m=t)=k,m}})\right\} \quad (3.5)$$

$$\phi_{d,t,k} \propto \exp\left\{\ln \rho_{k,t} + (\Psi(\gamma_{d,k}) - \Psi(\sum_{k=1}^K \gamma_{d,k})) + (\sum_{m=1}^{M_d} \lambda_{d,m,t} \ln \beta_{z_{(y_m=t)=k,m}})\right\} \quad (3.6)$$

3.4.4 Maximum Likelihood Parameter estimation

The expressions for the maximum likelihood estimators of the parameters of the original graphical model can be obtained using derivatives w.r.t the parameters of the functional \mathcal{L} . The following maximum likelihood expressions are obtained:

$$\rho_{k,g} \propto \sum_{d=1}^D \sum_{t=1}^T \sum_{g=1}^{T_G} \phi_{d,t,k} r_{d,t}^g \quad (3.7)$$

$$\beta_{k,j} \propto \sum_{d=1}^D \sum_{j=1}^V \sum_{m=1}^{M_d} \left(\sum_{t=1}^T \lambda_{d,m,t} \phi_{d,t,k} \right) w_{d,m}^j \quad (3.8)$$

$$(3.9)$$

where g and t are dummy index variables for all possible and document specific topic-coupled GSRts respectively. Further, $r_{d,t}^g$ is 1 iff $t = g$ and 0 otherwise i.e. a delta function. The updates of α and η following symmetric Dirichlet distributions are exactly the same as mentioned in [Blei et al., 2003].

3.5 The Learning To Summarize model—LeToS

This section investigates how a topic model like UTM which incorporates the proportions of the grammatical and semantic role transitions of words as a complementary view of a document indicating coherence can be extended to a model for multi-document summarization. The motivation to model the summarization process as generative model arises from an intuitive *psycholinguistic scenario*: Suppose in an exam, a student is asked to write an essay type answer based out of a large amount of preparatory reading materials. Now, under usual circumstances, she is less inclined to memorize the entire set of materials. Instead, for possible question scenarios, the student remembers only selected sentences (be directly extracted from text or paraphrased through natural language generation techniques) which are much like those found in the summary slide (section) of a lecture (chapter) about a particular topic. Then a coherent answer is constructed by expanding on the summary sentences and rearranging them.

Table 3.2 shows how dense columns (non “-” entries) of the document level sentence-term GSR grid identify potential coherent informative sentences w.r.t particular query words. Thus to extend UTM into a summarization model, each GSRt is treated as distribution over sentences. This makes the model to operate on only a fixed index of sentences which makes it amenable to extractive multi-document summarization but at the same time destroys the generalizability of UTM. This shortcoming is removed in Chapter 5 by addressing the problem of summarization from multiple perspectives including the topical significance of a sentence w.r.t the whole collection. To define a probabilistic topic summarization model, the document generation process is described as follows:

For each document $d \in 1, \dots, D$

- Choose a topic proportion $\theta | \alpha \sim Dir(\alpha)$
- Choose topic indicator $z_t | \theta \sim Mult(\theta)$
- Choose a GSRt $r_t | z_t = k, \rho \sim Mult(\rho_{z_t})$
- Choose a GSRt proportion $\pi | \eta \sim Dir(\eta)$

For each position m in document d :

For each instance of utterance s_p for which w_m occurs in s_p in document d :

Choose $v_p | \pi \sim \text{Mult}(\boldsymbol{\pi})$

Choose $y_m \sim v_p \delta(w_m \in s_p)$

Choose a sentence $s_p \sim \text{Mult}(\boldsymbol{\Omega}_{v_p})$

Choose a word $w_m | y_m = t, \mathbf{z}, \boldsymbol{\beta} \sim \text{Mult}(\boldsymbol{\beta}_{z_{y_m}})$

where, as before, M_d is the number of words in document $d \in 1, \dots, D$, P is the number of sentences in the same document and t is an index into one of the T GSRTs. The delta function $\delta(w_m \in s_p)$ is 1 iff the m^{th} word belongs to the p^{th} sentence and 0 otherwise. Under this extension, $\chi_{d,t} - \eta_t$ to be the expected number of words and sentences per topic-coupled GSRT in each document with $\chi_{d,t}$ being the variational surrogate for $\pi_{d,t}$ in the dual model. Each topic-coupled GSRT is also treated as a multinomial $\boldsymbol{\Omega}_t$ over the total number U of sentences in the corpus. Thus a GSRT is selected using π and a word $w_{d,m}$ is chosen to describe it and along with it a sentence $s_{d,p}$ containing $w_{d,m}$ is also sampled. In disjunction, π along with $v_{d,p}$, $s_{d,p}$ and $\boldsymbol{\Omega}$ focus mainly on topically induced coherence properties among the ‘‘coarser’’ units i.e. the sentences. However, the influence of a particular GSRT like ‘‘*subj*→*subj*’’ on coherence may be discounted if that is not the dominant trend in the transition topic. This fact is enforced through the coupling of empirical GSRT proportions to topics of the sentential words. Figure 3.2a give the depiction of the above process as a graphical model. The variational Bayesian counterpart of the model is exactly the same as in figure 3.1b but with an additional independent P plate inside of the D plate for sentence-GSRT multinomials i.e a plate with a directed arc from variational ζ_p to indicator $v_{d,p}$. For obtaining summaries, sentences are ordered w.r.t query words by accumulating the sentence-query

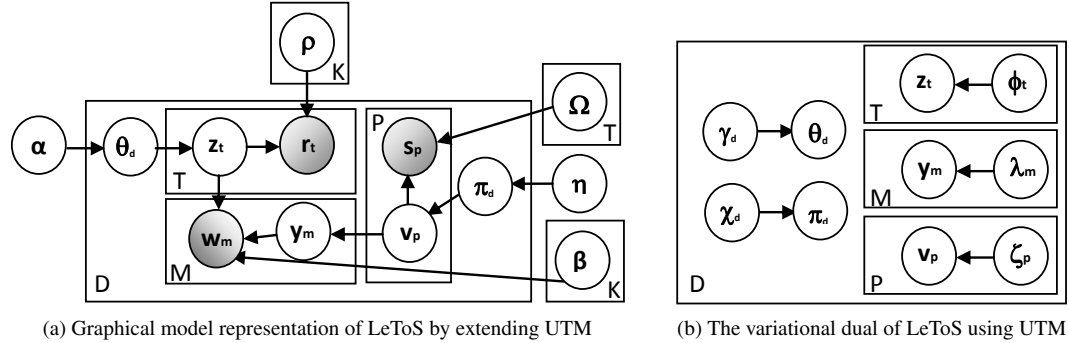


Figure 3.2: Extending the UTM model to the Learning To Summarize model (LeToS) by assuming that sentences are distributions over coarse coherence properties of discourses which in our case are GSRTs

word pair probability scores by computing:

$$p(s_{d,u} | \mathbf{q}) = \sum_{l=1}^Q \left(\sum_{t=1}^T \sum_{k=1}^K \zeta_{d,u,t} \phi_{d,t,k} (\lambda_{d,l,t} \phi_{d,t,k}) \gamma_{d,k} \chi_{d,t} \right) \delta(w_{d,l} \in s_{d,u}) \quad (3.10)$$

where Q is the number of the query words in query vector \mathbf{q} and s_u is the u^{th} sentence in the corpus that belongs to all such document d 's which are relevant to the query, w_l is the l^{th} query word, and k

and t and topic and GSRt indices respectively. Normally, under this model it can be enforced that each sentence in the summary be actually extracted from a unique document only, however, if larger more coherent summaries are needed, the sentences in the window of each most probable sentence can be included. Further, whenever possible, the sentences are scored over only “complete” GSRts which lack any “-” GSRs.

3.5.1 Parameter Estimation and Inference in the Extended Model

The set of equations in section 3.4.3 is augmented by the updates of the variational sentence multinomial and the posterior Dirichlet update for the topic coupled GSRt proportions as:

$$\chi_{d,t} = \eta_t + \sum_{m=1}^{M_d} \lambda_{d,m,t} + \sum_{p=1}^{P_d} \zeta_{d,p,t} \quad (3.11)$$

$$\zeta_{d,p,t} \propto \Omega_{t,p} \exp\{\Psi(\chi_{d,t}) - \Psi(\sum_{j=1}^T \chi_{d,j})\} \quad (3.12)$$

Note that these are again per-document updates. The only addition to the set of equations given in section 3.4.4 is:

$$\Omega_{t,u} \propto \sum_{d=1}^D \sum_{p=1}^{P_d} \zeta_{d,p,t} s_{d,p}^u \quad (3.13)$$

where u is an index into one of the S sentences in the corpus and $s_{d,p}^u = 1$ if the p^{th} sentence in document d is one among S .

3.6 Experimental Setup and Results

3.6.1 Description of the Datasets

The datasets that are used for finding topics as well as subsequent summarization are the DUC 2005, TAC 2008, TAC 2009 as well as a more practical real-life data from [Yahoo! Answers](#). The DUC 2005 dataset had 50 folders with at least 25 documents in each folder. Each such folder corresponded to a particular “topic focus” or “cluster” or “docset” representing varying human information needs.

The TAC 2008 dataset is organized into 48 folders as in DUC 2005, however, it also has documents in each folder grouped into two timelines which we merge for the sake of theme detection and summarization without modeling any temporal aspect. The organization for the TAC 2009 dataset is also similar with 44 folders. The manually collected Yahoo! Answers dataset consists of 10 such topic focuses with each topic focus pertaining to a particular real-life question. For each such topic focus, 10 relevant answers were collected and each answer was archived as a separate document.

Fig. 3.3 shows the empirical proportions of such GSRts, calculated across contexts of three sentences. The red dotted boxes show the histogram heights for some GSRts involving subjects, noun and WordNet concepts for two datasets: TAC 2008 newswire and the in-house collection of Yahoo! Answers. The histogram trends are more or less similar for both datasets and shows that the proportions of *complete* GSRts (i.e. GSRts with no null GSRs) are significantly less than those for incomplete GSRts. We next study the qualitative aspect of the topics of our UTM model and the summarization performance

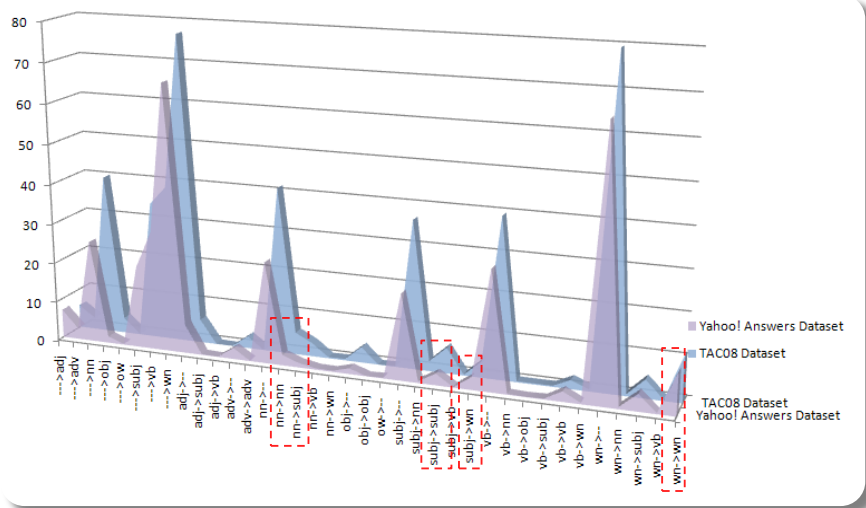


Figure 3.3: Empirical proportions of GSRts based on a maximum of three-sentence window for the Yahoo! Answers and TAC08 Newswire datasets

of its extension—the LeToS model.

3.6.2 Qualitative Topic Analysis and Summarization Performance

topic19	topic0	topic2	topic7	topic39	topic33	topic32	topic34
mines	company	Armstrong	ice	London	pope	Felt	drugs
coal	Fannie Mae	samples	glaciers	bomb	Vatican	Throat	planned
safety	executive	tested	years	police	John Paul II	Deep	Medicare
accidents	financial	Tour	sea	people	funeral	WoodWard	coverage
China	officer	L'Equipe	warming	attacks	words	Watergate	benefit
coal	chief	EPO	scientists	trains	Poland	FBI	part
mine	account	doping	Antarctica	subway	Rome	post	retirees
years	Raines	times	Alaska	bus	GMT	Mark	health
produced	billion	years	global	station	catholic	source	senior
officials	top	French	levels	killed	world	secret	cost
killed							

Table 3.4: Some topics from LDA for TAC 2008

Tables 3.4 and 3.5 present a few sample topics as distributions over the vocabulary by manually matching the topics from LDA and UTM—the latter incorporating the notion of coherence in a coarser way by depending only on the grammatical and syntactic role transitions without incorporating the surface forms of the words. Majority of the words in these topics come from documents that contained information regarding the following events:

- [i] “Describe the coal mine accidents in China and actions taken”
- [ii] “Give an account of the criminal investigation of Franklin Raines”

topic35	topic5	topic58	topic47	topic22	topic6	topic8	topic9
mine	Fannie Mae	Armstrong	planet	London	pope	Felt	drugs
coal China safety year	company account Raines Fannie	steroids tested samples years	Pluto ice scientists glaciers	bomb police attacks trains	Vatican funeral word John Paul II	Obama Throat Deep president	plan Medicare coverage retirees
accident officials panda state coal mine	financial executive Howard chief years	drugs Tour doping L'Equipe EPO	objects Earth warming sea melting	subway bus station killed city	GMT Rome move Catholic world	Watergate Woodward state FBI post	employment benefit prescription health subsidies

Table 3.5: Some topics from UTM for TAC 2008 matching those in Table 3.4

- [iii] “Describe accusations that seven-time Tour de France winner Lance Armstrong used the performance-enhancing drug EPO”
- [iv] “Describe the developments and impact of the continuing Arctic and Antarctic ice melts”
- [v] “Describe the July 7, 2005 bombings in London, England and the events, casualties and investigation resulting from the attack”
- [vi] “Follow the events connected with the death of Pope John Paul II”
- [vii] “Describe the revelation of the identity of *Deep Throat* and ensuing reactions”
- [viii] “Describe the developments in the Medicare Part D implementation”

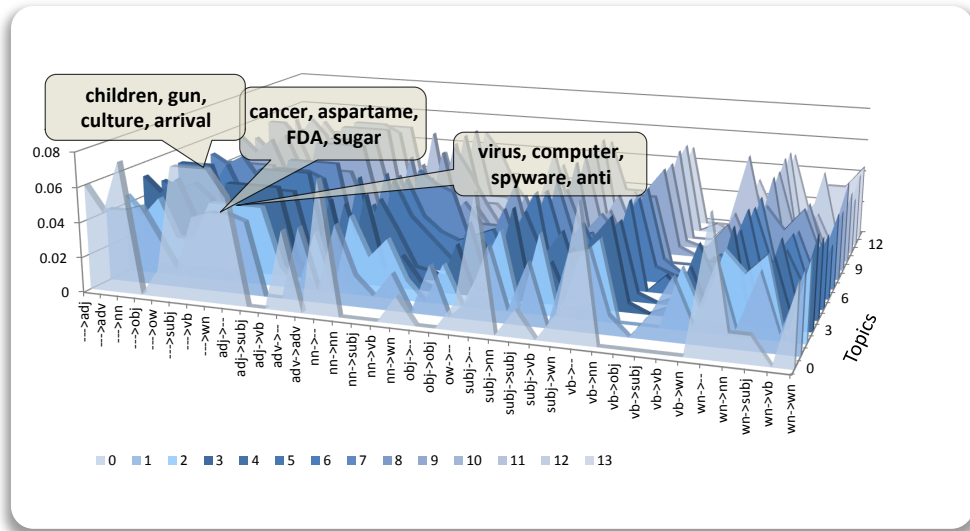
There is one important observation that becomes very conspicuous just by eyeballing table 3.5. There are few *intrusive* words such as “panda,” “Pluto” and “Obama” in topics 35, 47 and 8 respectively. In our Utterance Topic Model, the topics are not directly generated from word counts. Instead, the words are chosen to best describe a GSRt that a topic generates. Thus, word co-occurrence is not the only factor in determining the thematical structure of the documents for UTM. For example, the word Pluto has been observed in topic 47 because “Pluto” was found to describe the same GSRts as the other words in that topic. This phenomenon happened because there were a smaller set of documents which dealt with scientists discovering another planet outside of Pluto’s orbit. Similarly there are a few documents reporting shortage of bamboos as food for pandas in China. Thus in the UTM model, the influence of our coarse representation of contextual transitions i.e. center shifting or retention of GSRs such as “*subj*”, “*obj*” and “*wn*”s etc. shifts any chance of p -separability [Arora et al., 2013] of the topic-GSRt matrix ρ which in turn displaces the ideal p -separability properties of the topic-word matrix β more than that of a much simpler LDA model.

This is indeed a disadvantage of UTM. Similar minor such inclusions has been observed for other datasets as well. From a summarization point of view, however, there are no downsides to the effects of such inclusion of a few *intrusive* words. In terms of an information need like “effects of global warming,” there likely be no relevant documents containing Pluto and so these words don’t affect sentence probability calculation w.r.t. the query words.

Recall that LeToS is a fixed index (i.e no train-test split because of unique sentenceIDs) summarization system where the data is not allowed to change but the queries are i.e., once the model parameters

are calculated on the fixed data, variational inference is used to determine the topics of the free form queries, perform query expansion using related topical words and finally select the top sentences which are weighted by products of other variational parameters. To determine the number of topics fit to the data, one way is to run UTM on the dataset, and decide the best number of topics from an ELBO plot.

The “intrusive” nature of the topics is indeed a shortcoming of the coarse coherence encoding and we make use of a finer coherence structure by simply using a normalized surface form of the word in the GSRt in Chapter 5. Nevertheless, the indirect generation of words through document level metadata (here the GSRt proportion perspective) causes problems when there are more than a few topics significantly responsible for generating the document’s metadata. This fact also leads to lowering of the likelihood of the model to the data in terms of held-out log likelihood compared to LDA [Blei et al., 2003] and CTM [Blei and Lafferty, 2005].



Query 8: Has **gun culture** arrived in India? If yes, how to stop it?
 Query 1: Are **Sugar substitutes** bad for you?
 Query 2: What do you do to **protect** your **computer** from being infected by a **virus, worm** or **spyware** attack?

Figure 3.4: Strengths of the multinomial parameters of the topic distributions over GSRts for the in-house Yahoo! Answers dataset. Three sample queries are shown just beneath the plot that highlight the words which are highly probable for the topics over GSRts: ρ_1 , ρ_2 and ρ_3

Fig. 3.4 shows the plot of the parameters $\rho_{1:K}$ for $K = 14$ for the in-house Yahoo! Answers dataset. All topics have assigned much more probability masses over incomplete GSRts which is also a dominant trend in the documents themselves. Topic 1 which focuses on “gun culture” have relatively more mass over the “ \rightarrow [GSR]” and the “ $vb \rightarrow$ ” transitions than topics 2 and 3 which focus on “sugar substitutes” and “computer security.”

Similar to the plot in Fig. 3.4, Fig. 3.5 shows the plot of the parameter $\rho_{1:K}$ for the TAC 2008 dataset. The graphs are a little different from that in Fig. 3.4 for the complete GSRts. For the TAC 2008 dataset, the local coherence for words are highlighted more when conditioned on the topics. This is intuitive due to the better writing style and length of the documents in the newswire dataset. For the incomplete GSRts, we still have approximately the same trend for the in-house Yahoo! Answers dataset but a bit more pronounced in the latter dataset. In both cases, however, the empirical GSRt proportions

Systems	Rouge-2 Recall	95% Conf. Interval	Rouge-SU4 Recall	95% Conf. Interval
ModelA	0.34367	0.30939 - 0.37816	0.39876	0.36682 - 0.43142
ModelB	0.36794	0.33509 - 0.40440	0.43518	0.40642 - 0.46766
ModelC	0.30019	0.26992 - 0.33272	0.37335	0.34434 - 0.40416
ModelD	0.31269	0.28657 - 0.34182	0.38028	0.35551 - 0.40693
NUS	0.14632	0.12305 - 0.17200	0.23557	0.21646 - 0.25593
HK Poly	0.13984	0.11951 - 0.16282	0.23066	0.21194 - 0.25070
IIITH	0.14127	0.11740 - 0.16612	0.22849	0.20762 - 0.25163
LeToS-60-qE-U	0.13213	0.11064 - 0.15452	0.21425	0.19610 - 0.23395
LeToS-70-qE-U	0.12799	0.10648 - 0.14990	0.21448	0.19711 - 0.23455
LeToS-80-qE-U	0.13888	0.11332 - 0.16617	0.22302	0.20023 - 0.24589
LeToS-90-qE-U	0.12318	0.10329 - 0.14607	0.21242	0.19394 - 0.23263
LeToS-60-NE-qE-U	0.12556	0.10551 - 0.14537	0.21409	0.20009 - 0.22944
LeToS-70-NE-qE-U	0.12904	0.10692 - 0.15211	0.21747	0.20005 - 0.23662
LeToS-80-NE-qE-U	0.12481	0.10604 - 0.14501	0.21166	0.19586 - 0.22867
LeToS-90-NE-qE-U	0.12512	0.10679 - 0.14575	0.21385	0.19699 - 0.23102
LeToS-60-qE-NU	0.11320	0.09531 - 0.13337	0.19659	0.17934 - 0.21604
LeToS-70-qE-NU	0.11198	0.09233 - 0.13352	0.19710	0.18001 - 0.21641
LeToS-80-qE-NU	0.11767	0.09757 - 0.13863	0.20317	0.18336 - 0.22364
LeToS-90-qE-NU	0.11586	0.09764 - 0.13678	0.20264	0.18524 - 0.22224
LeToS-60-NE-qE-NU	0.10837	0.08754 - 0.13308	0.19365	0.17555 - 0.21414
LeToS-70-NE-qE-NU	0.08939	0.07229 - 0.10976	0.18461	0.16862 - 0.20149
LeToS-80-NE-qE-NU	0.09289	0.07617 - 0.11173	0.18546	0.17052 - 0.20204
LeToS-90-NE-qE-NU	0.09252	0.07710 - 0.10863	0.18788	0.17356 - 0.20317
BQFS	0.12976	0.10834 - 0.15281	0.21703	0.19938 - 0.23647
BE-ISI	0.11973	0.09801 - 0.14425	0.21084	0.19337 - 0.22957
UIR	0.09622	0.07994 - 0.11504	0.17894	0.16544 - 0.19240

Table 3.6: Comparison of DUC 2005 ROUGE Results

over 48 queries were obtained as **0.3089** with a rank of 14 out of 58 submissions. For TAC 2009 also, using the manual Pyramid [Nenkova and Passonneau, 2004] scoring for summaries, the average Pyramid scores for the 100 word summaries over 44 queries were obtained as **0.3024** for the A timeline and **0.2601** for the B timeline for LeToS and ranked 13th and 9th of 52 submissions. Note that the score is lower overall due to the extractive nature of summarization and a short 100 word limit. The phenomenon of summarization systems scoring less on shorter length summaries has been well explored in [Nenkova and Louis, 2008]. The scores for the system in [Srihari et al., 2007] that uses coherence to some extent and a baseline returning all the leading sentences (up to 100 words) in the most recent document are (0.1756 and 0.1601) and (0.175 and 0.160) respectively for the A and B timelines. The score for the B timeline is lower due to redundancy which was not addressed in our model. These scores indicate that performance of our model was consistent with the development (TAC 2008) dataset and test (TAC 2009) datasets. Although the *update task* in TAC 2009 is also to find the temporal novelty in the summaries of two timelines, we do not address that problem.

Role transition proportion analysis on the summaries raised a few interesting questions on using such graphs as a possible way to measure fluency in the final system generated summaries w.r.t. the model summaries. Although the trends of the proportions of the incomplete GSRts closely in Fig. 3.6

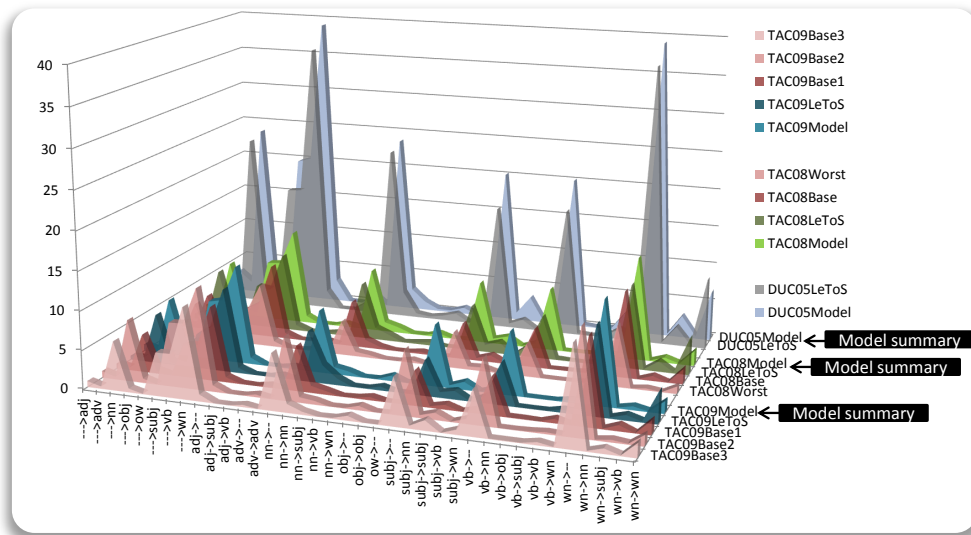


Figure 3.6: Empirical proportions of GSRs in the summaries obtained by various models on the Yahoo! Answers and TAC08/09 newswire datasets. Summaries from LeToS have each sentence coming from a different document.

match those of the model summaries for the DUC 2005, TAC 2008 and TAC 2009 datasets, the proportions for the complete GSRs involving nominal concepts remain the key discriminators. It is possible, then, to define a separate objective function just to reorder the candidate summary sentences by formulating an optimization framework that takes into account not only some measure of dissimilarity between adjacent sentences but also includes some constraint on the type of role transitions to maximize coherence within the spans of the adjacent sentences.

Southern	California	mudslides	mud	rain	man	vehicle	deaths	killed
<i>nn</i>	<i>wn</i>	<i>subj</i>	–	<i>wn</i>	–	–	<i>wn</i>	–
–	–	<i>subj</i>	–	–	<i>subj</i>	–	–	<i>vb</i>
–	–	–	<i>wn</i>	–	<i>subj</i>	<i>subj</i>	–	<i>vb</i>

Table 3.7: A snapshot of sentences that focuses on mudslides killing men

Table 3.7 shows some sample sentences from the D0906B folder of the TAC2009 dataset which concerns itself with documents related to the information need: “Describe the effects and responses to the heavy rainfall and mudslides in Southern California.” Note that “*vb*” denotes a verb and “*wn*” denotes a WordNet synset role assignment. This snapshot of sentences once more verifies the claim of Centering theory that center propagation across sentences e.g. (mudslides, *subj*), (man, *subj*) and (killed, *vb*) seem to select those sentences that best represents a scenario which is easy to remember. In other words, these sentences will make an impact on a reader the most as she reads through the discourse. The corresponding sentences below also seem to qualitatively support the high relevancy of these sentences w.r.t. the query.

1. “A fourth day of thrashing thunderstorms began to take a heavier toll on **southern California** on Sunday with at least three **deaths** blamed on the **rain**, as flooding and **mudslides** forced road closures and emergency crews carried out harrowing rescue operations.”

2. “In Elysian_Park, just north of downtown, a 42-year-old homeless **man** was **killed** and another injured when a **mudslide** swept away their makeshift encampment.”

3. “Another **man** was **killed** on Pacific_Coast_Highway in Malibu when his sport utility **vehicle** skidded into a **mud** patch and plunged into the Pacific Ocean.”

<p>Short ≈120 words summarized answer for “Are Sugar substitutes bad for you?” [with query expansion and each sentence belongs to a different document]</p>
<p>“The show stated Aspartame turns into METHANOL in your body and is like drinking FORMALDEHYDE! Splenda is another popular one, but because the body doesn’t recognize, the body won’t digest it, and can actually make you GAIN weight. The FDA has approved it for 5 mg/Kg body weight, which is the least of all the sweeteners and comes out to 6 cans of diet cola per day. Aspartame is at the root of diseases such as: aspartame fibromyalgia, aspartame restless leg syndrome, aspartame and migraines, aspartame and tumors, aspartame allergy, aspartame multiple sclerosis, bladder cancer aspartame, aspartame and central nervous system, aspartame and infertility, aspartame and weight gain,....”</p>
<p>Short ≈120 words summarized answer for “Are Sugar substitutes bad for you?” [without query expansion and each sentence belongs to a different document]</p>
<p>“OK, so why are sugar subs so bad? Sugar has a bad reputation because most people consume too much of it and don’t realize Sugar. It’s made by substituting three atoms of chlorine for three hydroxyl groups on the sugar molecule. This sugar substitute, sold commercially as Equal and NutraSweet, was hailed as the savior for dieters who for decades had put up with saccharine’s unpleasant after taste. i am severely addicted to sugar substitutes, initially worrying about my weight. My advice, speak to a medical professional and seek their advice on sugar substitutes as far as their recommendations and opinion on different brands, etc.. because it’s natural, it’s better for you than equal or sweet n low, but because the liver can’t process it as quickly as pure sugar, it converts into triglycerides, increasing your risk of heart disease.”</p>
<p>Short ≈120 words baseline summary for “Are Sugar substitutes bad for you?”</p>
<p>“Fructose is another extract of sugar. Hope that shed a little light on your questions. It’s made by substituting three atoms of chlorine for three hydroxyl groups on the sugar molecule. Honey enters the bloodstream slowly, 2 calories per minute, while sugar enters quickly at 10 calories per minute, causing blood sugars to fluctuate rapidly and wildly. This sugar substitute, sold commercially as Equal and NutraSweet, was hailed as the savior for dieters who for decades had put up with saccharine’s unpleasant after taste. Too much phenylalanine causes seizures, elevated blood plasma, is dangerous for pregnancy causing retardation, PMS caused by phenylalanine’s blockage of serotonin, insomnia, and severe mood swings. Sugar substitutes, turn into formaldehyde in the body..”</p>

Table 3.8: Different short ≈120 words summarized answers for “Are Sugar substitutes bad for you?”

A sample summarized ≈ 120 words answers for a small Yahoo! Answers dataset³ is also presented here which had been obtained using our proposed extension of UTM to LeToS. A particularly popular question was selected as an example for answer summarization - “**Are Sugar substitutes bad for you?**” The questions were fed into the model with standard stopword removal and stemming and was thus transformed into “sugar substitut bad”

Table 3.8 shows the summary to the question on sugar substitutes w.r.t. the query being input as is or with topic expansion. A set of best fit topics from this small experimental real-world dataset can be found at the author’s website⁴. A baseline summary is also included for qualitative comparison. The baseline summaries are generated such that two sentences are extracted with at least one query word overlap from the beginning and end of each document till the length constraint is satisfied. For this baseline, the documents have been considered in the lexicographic order of their filenames.

We observe in Table 3.8, that in the summary w.r.t. the expanded query, the word “aspartame” received prominence due to topic based query expansion. This expansion is an unsupervised step where the query is expanded from its original bag-of-words form to an expanded bag-of-words form where the extra words are appended which are very likely to appear in the topic of the original query words in the relevant documents. Additionally the summaries consist of one single **unique** sentence from each document to reflect an ideal multi-document summarization task scenario. Although experiments on the DUC 2005 dataset show that the ROUGE scores increase by following this approach, however, the overall quality of coherence decreases. One can introduce coherence by using the context around the selected sentence as was chosen to select the GSRts, but then the topical relevance of the context remains in question. Also, a length of 100 words or even 250 words is too restrictive to introduce contextual paragraphs in summaries.

3.7 Summary

The summarization model discussed in this chapter chooses sentences which contain words which best describe the coarser GSRts. The lack of any structure in the query e.g. viewing the query as a bag of words like “sugar substitute bad” gives rise to serious drawbacks. There is no way by which the topic model can understand the real intention of the query: “How or why are sugar substitutes bad for you?” This is a severe limitation of all topic models but is also intuitive since nowhere in the objective function are there expressions which suggest any “intentional” view of document generation. Summarization through topic models thus only answers part of the objective—one that relates to ascribing significance to words through document topic analysis. We address this issue in Chapter 3 not through modeling the query but by exploiting rhetorical relations which relate different spans of a sentence and reflect intentions to some degree as well.

From the success of higher order n-grams in information retrieval [Buttcher et al., 2010], one can potentially envision the benefit of using at least bi-grams within the purview of the topic modeling framework. Indeed, there has been some work [Wallach, 2008] on topic models with word bi-grams, however, even not all word bigrams of “How or why are sugar substitutes bad for you?” can solve the problem of *understanding* the query. Also in a bigram topic model, there are $K \times V \times (V - 1)$ parameters to represent topics, where K is the number of topics and V is the size of the unigram vocabulary. For an

³<http://www.acsu.buffalo.edu/~pdas3/research/datasets.html>

⁴<http://www.acsu.buffalo.edu/~pdas3/research/LeToS.html>

n-gram word topic model, there will thus be $K \times \prod_{i=0}^{n-1} (V - i)$ topic n-gram parameters which can lead to severe overfitting when relevant data is not abundant.

In the next chapter we develop topic models that are general enough to handle an ubiquitous document structure—documents containing a high level metadata at the document level (such as the GSRs used in this chapter, any controlled vocabulary, any user assigned document tags, embedded multimedia captions, etc.) as well as fine-grained word level annotations. We use these models in a novel way along with several simpler to complex local models to address the problem of multi-document summarization in Chapter 5.

3.8 Appendix

This section gives partially complete derivations to find out the optimal settings of the hidden variables and the parameters of the utterance topic model following the framework laid out in Chapter 1. Note that the inference part i.e. inferring variational distributions for hidden variables (E-step) is document specific, while the model parameter estimation (M-step) is corpus wide. We start out with some initial values of the parameters and we then find the posterior distribution over the latent variables parameterized by the free variational parameters in the VBE step and holding this distribution fixed, optimize the parameters of the model in the VBM step. In each of these steps, we select out only those terms from \mathcal{L} that depend on the variable being optimized.

3.8.1 Derivations for the UTM model

$$(\gamma^*, \chi^*, \phi^*, \lambda^*) = \arg \min_{(\gamma, \chi, \phi, \lambda)} KL(q(\theta, \pi, \mathbf{z}, \mathbf{y} | \gamma, \chi, \phi, \lambda) || p(\theta, \pi, \mathbf{z}, \mathbf{y} | \mathbf{r}, \mathbf{w}, \alpha, \eta, \rho, \beta)) \quad (3.14)$$

By Jensen's inequality, we have

$$\ln p(\mathbf{r}, \mathbf{w} | \alpha, \eta, \rho, \beta) \geq \{\mathbb{E}_q[p(\mathbf{r}, \mathbf{w}, \theta, \pi, \mathbf{z}, \mathbf{y} | \alpha, \eta, \rho, \beta)] - \mathbb{E}_q[q(\theta, \pi, \mathbf{z}, \mathbf{y} | \gamma, \chi, \phi, \lambda)]\} = \mathcal{L} \quad (3.15)$$

We thus have:

$$\begin{aligned} \mathcal{L}(\gamma, \chi, \phi, \lambda) = & \mathbb{E}_q[\ln p(\theta | \alpha)] + \mathbb{E}_q[\ln p(\pi | \eta)] + \mathbb{E}_q[\ln p(\mathbf{z} | \theta)] + \mathbb{E}_q[\ln p(\mathbf{r} | \mathbf{z}, \rho)] + \mathbb{E}_q[\ln p(\mathbf{y} | \pi)] \\ & + \mathbb{E}_q[\ln p(\mathbf{w} | \mathbf{y}, \mathbf{z}, \beta)] - \mathbb{E}_q[\ln q(\theta | \gamma)] - \mathbb{E}_q[\ln q(\pi | \chi)] - \mathbb{E}_q[\ln q(\mathbf{z} | \phi)] - \mathbb{E}_q[\ln q(\mathbf{y} | \lambda)] \end{aligned} \quad (3.16)$$

Each of the terms in the equation (3.16) expands out to:

$$\ln \Gamma\left(\sum_{j=1}^K \alpha_j\right) - \sum_{k=1}^K \ln \Gamma(\alpha_k) + \sum_{k=1}^K (\alpha_k - 1) \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right) \right) \quad (3.17)$$

$$+ \ln \Gamma\left(\sum_{f=1}^T \eta_f\right) - \sum_{t=1}^T \ln \Gamma(\eta_t) + \sum_{t=1}^T (\eta_t - 1) \left(\Psi(\chi_{d,t}) - \Psi\left(\sum_{f=1}^T \chi_{d,f}\right) \right) \quad (3.18)$$

$$+ \sum_{t=1}^T \sum_{k=1}^K \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right) \right) \phi_{d,t,k} \quad (3.19)$$

$$+ \sum_{t=1}^T \sum_{k=1}^K \sum_{g=1}^{T_G} \phi_{d,t,k} \ln \rho_{k,t} r_{d,t}^g \quad (3.20)$$

$$+ \sum_{m=1}^M \sum_{t=1}^T \left(\Psi(\chi_{d,t}) - \Psi\left(\sum_{j=1}^T \chi_{d,j}\right) \right) \lambda_{d,m,t} \quad (3.21)$$

$$+ \sum_{m=1}^M \sum_{k=1}^K \sum_{j=1}^V \left(\sum_{t=1}^T \lambda_{d,m,t} \phi_{d,t,k} \right) \ln \beta_{z_{(y_{d,m}=t)=k,j}} w_{d,m}^j \quad (3.22)$$

$$- \ln \Gamma\left(\sum_{j=1}^K \gamma_{d,j}\right) + \sum_{k=1}^K \ln \Gamma(\gamma_{d,k}) - \sum_{k=1}^K (\gamma_{d,k} - 1) \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right) \right) \quad (3.23)$$

$$- \ln \Gamma\left(\sum_{j=1}^T \chi_j\right) + \sum_{t=1}^T \ln \Gamma(\chi_{d,t}) - \sum_{t=1}^T (\chi_{d,t} - 1) \left(\Psi(\chi_{d,t}) - \Psi\left(\sum_{j=1}^T \chi_{d,j}\right) \right) \quad (3.24)$$

$$- \sum_{t=1}^T \sum_{k=1}^K \phi_{d,t,k} \ln \phi_{d,t,k} \quad (3.25)$$

$$- \sum_{m=1}^M \sum_{t=1}^T \lambda_{d,m,t} \ln \lambda_{d,m,t} \quad (3.26)$$

Where, each term in a document is represented as a binary vector $w_n^j, j \in \{1, \dots, V\}$, V being the number of terms in the vocabulary. The total number of GSR transitions is fixed at T_G and Ψ is the digamma function. It is to be understood that the t index for variational parameter updates is specific to the GSRt IDs in a document d and that for the global parameters like ρ , g is a global index into one of the possible T_G GSRts. M is the number of terms in a document.

3.8.1.1 INFERENCE ON VARIATIONAL PARAMETERS

Here we estimate the free variational parameters for the variational model depicted in Fig. 3.1b following the constraints on ϕ and λ .

For γ :

$$\mathcal{L}[\gamma] = - \ln \Gamma\left(\sum_{j=1}^K \gamma_{d,j}\right) + \sum_{k=1}^K \ln \Gamma(\gamma_{d,k}) + \sum_{k=1}^K (\alpha_k + \sum_{t=1}^T \phi_{d,t,k} - \gamma_{d,k}) (\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right)) \quad (3.27)$$

$$\begin{aligned} \frac{\partial \mathcal{L}[\gamma]}{\partial \gamma_{d,k}} = & (\alpha_k + \sum_{t=1}^T \phi_{d,t,k} - \gamma_{d,k}) \left(\Psi'(\gamma_{d,k}) - \Psi'\left(\sum_{j=1}^K \gamma_{d,j}\right) \right) \\ & - (\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right)) + (\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right)) \end{aligned}$$

Setting the above derivative to 0, we get,

$$\gamma_{d,k} = \alpha_k + \sum_{t=1}^T \phi_{d,t,k} \quad \text{since} \quad \left(\Psi'(\gamma_{d,k}) - \Psi'\left(\sum_{j=1}^K \gamma_{d,j}\right) \right) \neq 0 \quad (3.28)$$

For χ :

We follow the procedure above exactly and by taking derivative of $\mathcal{L}[\chi]$ w.r.t. $\chi_{d,t}$, we have:

$$\chi_{d,t} = \eta_t + \sum_{m=1}^M \lambda_{d,m,t} \quad (3.29)$$

For λ :

$$\begin{aligned} \mathcal{L}_{[\lambda]} = & \sum_{m=1}^M \sum_{t=1}^T (\Psi(\chi_{d,t}) - \Psi(\sum_{j=1}^T \chi_{d,j})) \lambda_{d,m,t} + \sum_{m=1}^M \sum_{k=1}^K \sum_{j=1}^V \sum_{t=1}^T (\sum_{d,m,t} \lambda_{d,m,t} \phi_{d,t,k}) \ln \beta_{z_{y_{d,m}},j} w_{d,m}^j \\ & - \sum_{m=1}^M \sum_{t=1}^T \lambda_{d,m,t} \ln \lambda_{d,m,t} + \sum_{m=1}^M \mu_m (\sum_{t=1}^T \lambda_{d,m,t} - 1) \end{aligned} \quad (3.30)$$

where μ are the m Lagrange multipliers in (3.30) for document d .

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \lambda_{d,m,t}} = 0 & \implies (\Psi(\chi_{d,t}) - \Psi(\sum_{j=1}^T \chi_{d,j})) + (\sum_{t=1}^T \phi_{d,t,k} \ln \beta_{z_{y_{d,m}},j}) - 1 - \ln \lambda_{d,m,t} + \mu_m = 0 \\ \implies \lambda_{d,m,t} & = \exp \left\{ \Psi(\chi_{d,t}) - \Psi(\sum_{f=1}^T \chi_{d,f}) + (\sum_{k=1}^K \phi_{d,t,k} \ln \beta_{z_{y_{d,m}},j}) - 1 + \mu_m \right\} \\ \implies \exp\{\mu_m - 1\} & = \frac{1}{\sum_{t=1}^T \exp \left\{ (\Psi(\chi_{d,t}) - \Psi(\sum_{f=1}^T \chi_{d,f})) + (\sum_{k=1}^K \phi_{d,t,k} \ln \beta_{z_{y_{d,m}},j}) \right\}} \end{aligned}$$

Setting the derivative $\frac{\partial \mathcal{L}}{\partial \lambda_{d,m,t}}$ to 0 gives us:

$$\lambda_{d,m,t} \propto \exp \left\{ (\Psi(\chi_{d,t}) - \Psi(\sum_{f=1}^T \chi_{d,f})) + (\sum_{k=1}^K \phi_{d,t,k} \ln \beta_{z_{y_{d,m}},j}) \right\} \quad (3.31)$$

For ϕ :

$$\begin{aligned} \mathcal{L}_{[\phi]} = & \sum_{t=1}^T \sum_{k=1}^K (\Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) \phi_{d,t,k} + \sum_{t=1}^T \sum_{k=1}^K \phi_{d,t,k} \ln \rho_{d,k,t} \\ & + \sum_{m=1}^M \sum_{k=1}^K \sum_{j=1}^V \sum_{t=1}^T (\sum_{d,m,t} \lambda_{d,m,t} \phi_{d,t,k}) \ln \beta_{z_{(y_{d,m}),j}} w_{d,m}^j - \sum_{t=1}^T \sum_{k=1}^K \phi_{d,t,k} \ln \phi_{d,t,k} + \mu_t (\sum_{k=1}^K \phi_{d,t,k} - 1) \end{aligned} \quad (3.32)$$

where μ are the t Lagrange multipliers in $\mathcal{L}_{[\phi]}$.

As before, we have:

$$\frac{\partial \mathcal{L}}{\partial \phi_{d,t,k}} = 0 \implies \phi_{d,t,k} \propto \exp \{ \ln \rho_{d,t,k} + (\Psi(\gamma_{d,k}) - \Psi(\sum_{k=1}^K \gamma_{d,k})) + (\sum_{m=1}^M \lambda_{d,m,t} \ln \beta_{z_{(y_{d,m}),j}}) \} \quad (3.33)$$

3.8.1.2 MODEL PARAMETER ESTIMATION

Here we calculate the maximum likelihood settings of the parameters that do not grow with the data. So we need to take into account the contribution of these for all the documents and not just a single document.

For ρ :

$$\mathcal{L}_{[\rho]} = \sum_{d=1}^D \sum_{t=1}^{T_d} \sum_{k=1}^K \sum_{g=1}^{T_G} \phi_{d,t,k} \ln \rho_{k,t} r_{d,t}^g + \sum_{k=1}^K \mu_k \left(\sum_{g=1}^{T_G} \rho_{k,g} - 1 \right) \quad (3.34)$$

where the μ_k 's are the K Lagrange multipliers in (3.34).

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \rho_{k,g}} &= \sum_{d=1}^D \sum_{t=1}^{T_d} \phi_{d,t,k} r_{d,t}^g \frac{1}{\rho_{k,g}} + \mu_k \\ \frac{\partial \mathcal{L}}{\partial \rho_{k,g}} = 0 &\implies \rho_{k,g} = - \frac{\sum_{d=1}^D \sum_{t=1}^{T_d} \phi_{d,t,k} r_{d,t}^g}{\mu_k} \implies \mu_k = - \sum_{g=1}^{T_G} \sum_{d=1}^D \sum_{t=1}^{T_d} \phi_{d,t,k} r_{d,t}^g \\ \therefore \frac{\partial \mathcal{L}}{\partial \rho_{k,g}} = 0 &\implies \rho_{k,g} \propto \sum_{d=1}^D \sum_{t=1}^{T_d} \phi_{d,t,k} r_{d,t}^g \end{aligned} \quad (3.35)$$

For β :

$$\mathcal{L}_{[\beta]} = \sum_{m=1}^M \sum_{k=1}^K \sum_{j=1}^V \left(\sum_{t=1}^T \lambda_{d,m,t} \phi_{d,t,k} \right) \ln \beta_{z_{y_{d,m,j}}} w_{d,m}^j + \sum_{k=1}^K \mu_k \left(\sum_{j=1}^V \beta_{k,j} - 1 \right) \quad (3.36)$$

where μ_k s are the K Lagrange Multipliers in (3.36)

$$\frac{\partial \mathcal{L}}{\partial \beta_{k,j}} = 0 \implies \beta_{k,j} \propto \sum_{d=1}^D \sum_{m=1}^{M_d} \left(\sum_{t=1}^T \lambda_{d,m,t} \phi_{d,t,k} \right) w_{d,m}^j \quad (3.37)$$

For α :

$$\begin{aligned} \mathcal{L}_{[\alpha]} &= \sum_{d=1}^D \left(\ln \Gamma \left(\sum_{k=1}^K \alpha_k \right) - \sum_{k=1}^K \ln \Gamma(\alpha_k) + \sum_{k=1}^K (\alpha_k - 1) (\Psi(\gamma_{d,k}) - \Psi \left(\sum_{j=1}^K \gamma_{d,j} \right)) \right) \\ &\implies \frac{\partial \mathcal{L}}{\partial \alpha_k} = D \left(-\Psi(\alpha_k) + \Psi \left(\sum_{j=1}^K \alpha_j \right) \right) + \sum_{d=1}^D (\Psi(\gamma_{d,k}) - \Psi \left(\sum_{j=1}^K \gamma_{d,j} \right)) \\ &\quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \alpha_k \alpha_j} = \partial(k, j) D \left(\Psi'(\alpha_k) - \Psi' \left(\sum_{j=1}^K \alpha_j \right) \right) \end{aligned} \quad (3.38)$$

The derivative w.r.t. α_k depends on α_j and thus we can resort to Newton's iterative method to find out the maximal α using the gradient and Hessian vector and matrix respectively as in [Blei et al., 2003]. Ψ' is the trigamma function.

For η :

The update is similar to α update

3.8.2 Derivations for the LeToS Model for Summarization

This section gives partially complete derivations to find out the optimal settings of the hidden variables and the parameters of the Learning To Summarize (LeToS) model for summarization extended from Utterance Topic Model. As before, the inference part i.e. inferring variational distributions for hidden variables (E-step) is document specific, while the model parameter estimation (M-step) is corpus wide. In this model we set out to define each topic coupled GSRt proportion to be distribution over the sentence vocabulary in addition to all model parameters in UTM. In this case, thus, the Evidence Lower Bound (ELBO) \mathcal{L} from Equ. 3.15 includes terms that represent the dependency of the topic coupled GSRts to sentences in the training set. This dependence on a fixed index of sentences in the training set is the principle cause for making LeToS a fixed index model without the ability to generalize to new documents.

$$(\gamma^*, \chi^*, \phi^*, \lambda^*, \zeta^*) = \arg \min_{(\gamma, \chi, \phi, \lambda, \zeta)} KL(q(\theta, \pi, \mathbf{z}, \mathbf{y}, \mathbf{v} | \gamma, \chi, \phi, \lambda, \zeta) || p(\mathbf{r}, \mathbf{w}, \mathbf{s} | \alpha, \eta, \rho, \beta, \Omega)) \quad (3.39)$$

By Jensen's inequality, we have

$$\ln p(\mathbf{r}, \mathbf{w}, \mathbf{s} | \alpha, \eta, \rho, \beta, \Omega) \geq \{\mathbb{E}_q[p(\mathbf{r}, \mathbf{w}, \mathbf{s}, \theta, \pi, \mathbf{z}, \mathbf{y}, \mathbf{v} | \alpha, \eta, \rho, \beta, \Omega)] - \mathbb{E}_q[q(\theta, \pi, \mathbf{z}, \mathbf{y}, \mathbf{v} | \gamma, \chi, \phi, \lambda, \zeta)]\} \quad (3.40)$$

We thus have:

$$\begin{aligned} \mathcal{L}(\gamma, \chi, \phi, \lambda, \zeta; \alpha, \eta, \rho, \beta, \Omega) &= \mathbb{E}_q[\ln p(\theta | \alpha)] + \mathbb{E}_q[\ln p(\pi | \eta)] + \mathbb{E}_q[\ln p(\mathbf{z} | \theta)] + \mathbb{E}_q[\ln p(\mathbf{r} | \mathbf{z}, \rho)] \\ &+ \mathbb{E}_q[\ln p(\mathbf{y} | \mathbf{v})] + \mathbb{E}_q[\ln p(\mathbf{w} | \mathbf{y}, \mathbf{z}, \beta)] + \mathbb{E}_q[\ln p(\mathbf{v} | \pi)] + \mathbb{E}_q[\ln p(\mathbf{s} | \mathbf{v}, \Omega)] \\ &- \mathbb{E}_q[\ln q(\theta | \gamma)] - \mathbb{E}_q[\ln q(\pi | \chi)] - \mathbb{E}_q[\ln q(\mathbf{z} | \phi)] - \mathbb{E}_q[\ln q(\mathbf{y} | \lambda)] - \mathbb{E}_q[\ln q(\mathbf{v} | \zeta)] \end{aligned} \quad (3.41)$$

Each of the terms in the equation (3.41) expands out to:

$$\ln \Gamma\left(\sum_{j=1}^K \alpha_j\right) - \sum_{k=1}^K \ln \Gamma(\alpha_k) + \sum_{k=1}^K (\alpha_k - 1) \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right) \right) \quad (3.42)$$

$$+ \ln \Gamma\left(\sum_{f=1}^T \eta_f\right) - \sum_{t=1}^T \ln \Gamma(\eta_t) + \sum_{t=1}^T (\eta_t - 1) \left(\Psi(\chi_{d,t}) - \Psi\left(\sum_{f=1}^T \chi_{d,f}\right) \right) \quad (3.43)$$

$$+ \sum_{t=1}^T \sum_{k=1}^K \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right) \right) \phi_{d,t,k} \quad (3.44)$$

$$+ \sum_{t=1}^T \sum_{k=1}^K \sum_{g=1}^{T_G} \phi_{d,t,k} r_{d,t}^g \ln \rho_{k,t} \quad (3.45)$$

$$+ \sum_{m=1}^M \sum_{t=1}^T \left(\Psi(\chi_{d,t}) - \Psi\left(\sum_{j=1}^T \chi_{d,t}\right) \right) \lambda_{d,m,t} \quad (3.46)$$

$$+ \sum_{m=1}^M \sum_{k=1}^K \sum_{j=1}^V \left(\sum_{t=1}^T \lambda_{d,m,t} \phi_{d,t,k} \right) \ln \beta_{z_{(y_{d,m}=t)=k}, j} w_{d,m}^j \quad (3.47)$$

$$+ \sum_{p=1}^P \sum_{t=1}^T \left(\Psi(\chi_{d,t}) - \Psi\left(\sum_{j=1}^T \chi_{d,j}\right) \right) \zeta_{d,p,t} \quad (3.48)$$

$$+ \sum_{p=1}^P \sum_{t=1}^T \sum_{u=1}^U \zeta_{d,p,t} s_{d,p}^u \ln \Omega_{t,p} \quad (3.49)$$

$$- \ln \Gamma\left(\sum_{j=1}^K \gamma_{d,j}\right) + \sum_{k=1}^K \ln \Gamma(\gamma_{d,k}) - \sum_{k=1}^K (\gamma_{d,k} - 1) \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right) \right) \quad (3.50)$$

$$- \ln \Gamma\left(\sum_{j=1}^T \chi_{d,j}\right) + \sum_{t=1}^T \ln \Gamma(\chi_{d,t}) - \sum_{t=1}^T (\chi_{d,t} - 1) \left(\Psi(\chi_{d,t}) - \Psi\left(\sum_{j=1}^T \chi_{d,j}\right) \right) \quad (3.51)$$

$$- \sum_{t=1}^T \sum_{k=1}^K \phi_{d,t,k} \ln \phi_{d,t,k} \quad (3.52)$$

$$- \sum_{m=1}^M \sum_{t=1}^T \lambda_{d,m,t} \ln \lambda_{d,m,t} \quad (3.53)$$

$$- \sum_{p=1}^P \sum_{t=1}^T \zeta_{d,p,t} \ln \zeta_{d,p,t} \quad (3.54)$$

Where, each term in a document is represented as a binary vector $w_{d,m}^j$, $j \in \{1, \dots, V\}$, V being the number of terms in the vocabulary. The number of GSR transitions is fixed at T and Ψ is the digamma function. It is to be understood that the t index for variational parameter updates is specific to the GSRt IDs in a document d and that for the global parameters like ρ , g is a global index into one of the possible T_G GSRts. M is the document length w.r.t. terms and P is the document length in terms of sentences.

3.8.2.1 INFERENCE ON VARIATIONAL PARAMETERS

Here we estimate the free variational parameters for the variational model depicted in Fig. 3.2b following the constraints on ϕ and λ .

For γ : The derivations for γ in the case of the LeToS model is the same as in Eqs. 3.27 and 3.28.

For χ : Following the procedure above and taking derivative of $\mathcal{L}[\chi]$ w.r.t. χ_t , we have

$$\chi_{d,t} = \eta_t + \sum_{m=1}^M \lambda_{d,m,t} + \sum_{p=1}^P \zeta_{d,p,t} \quad (3.55)$$

For λ :

$$\begin{aligned} \mathcal{L}[\lambda] = & \sum_{m=1}^M \sum_{t=1}^T \left(\Psi(\chi_{d,t}) - \Psi\left(\sum_{j=1}^T \chi_{d,j}\right) \right) \lambda_{d,m,t} + \sum_{m=1}^M \sum_{k=1}^K \sum_{j=1}^V \sum_{t=1}^T \left(\sum_{t=1}^T \lambda_{d,m,t} \phi_{d,t,k} \right) \ln \beta_{z_{y_{d,m}}} w_{d,m}^j \\ & - \sum_{m=1}^M \sum_{t=1}^T \lambda_{d,m,t} \ln \lambda_{d,m,t} + \sum_{m=1}^M \mu_m \left(\sum_{t=1}^T \lambda_{d,m,t} - 1 \right) \end{aligned} \quad (3.56)$$

where μ is the Lagrange multiplier in (3.56)

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \lambda_{d,m,t}} = 0 &\implies \Psi(\chi_{d,t}) - \Psi\left(\sum_{j=1}^T \chi_{d,j}\right) + \left(\sum_{t=1}^T \phi_{d,t,k} \ln \beta_{z_{y_{d,m},j}}\right) - 1 - \ln \lambda_{d,m,t} + \mu_m = 0 \\
&\implies \lambda_{d,m,t} = \exp\left\{\Psi(\chi_{d,t}) - \Psi\left(\sum_{f=1}^T \chi_{d,f}\right) + \left(\sum_{k=1}^K \phi_{d,t,k} \ln \beta_{z_{y_{d,m},j}}\right) - 1 + \mu_m\right\} \\
&\implies \sum_{t=1}^T \lambda_{d,m,t} = 1 \implies 1 = \sum_{t=1}^T \exp\left\{\Psi(\chi_{d,t}) - \Psi\left(\sum_{j=1}^T \chi_{d,j}\right) + \left(\sum_{k=1}^K \phi_{d,t,k} \ln \beta_{z_{(y_{d,m}=t)=k}}\right) - 1 + \mu_m\right\} \\
&\implies \exp\{\mu_m - 1\} = \frac{1}{\sum_{t=1}^T \exp\left\{\Psi(\chi_{d,t}) - \Psi\left(\sum_{f=1}^T \chi_{d,f}\right) + \left(\sum_{k=1}^K \phi_{d,t,i} \ln \beta_{z_{y_{d,m},j}}\right)\right\}}
\end{aligned} \tag{3.57}$$

Setting the derivative $\frac{\partial \mathcal{L}}{\partial \lambda_{d,m,t}}$ to 0 gives us,

$$\lambda_{d,m,t} \propto \exp\left\{\Psi(\chi_{d,t}) - \Psi\left(\sum_{f=1}^T \chi_{d,f}\right) + \left(\sum_{k=1}^K \phi_{d,t,k} \ln \beta_{z_{y_{d,m},j}}\right)\right\} \tag{3.58}$$

For ϕ :

$$\begin{aligned}
F_{[\phi]} &= \sum_{t=1}^T \sum_{k=1}^K (\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^K \gamma_j\right)) \phi_{d,t,k} + \sum_{t=1}^T \sum_{k=1}^K \phi_{d,t,k} \ln \rho_{k,t} \\
&\quad + \sum_{m=1}^M \sum_{k=1}^K \sum_{j=1}^V \sum_{t=1}^T \lambda_{d,m,t} \phi_{d,t,k} w_{d,m}^j \ln \beta_{z_{y_{d,m},j}} - \sum_{t=1}^T \sum_{k=1}^K \phi_{d,t,k} \ln \phi_{d,t,k} + \sum_{t=1}^T \mu_t \left(\sum_{k=1}^K \phi_{d,t,k} - 1\right)
\end{aligned} \tag{3.59}$$

where μ are the T Lagrange multipliers in $\mathcal{L}_{[\phi]}$. As before,

$$\frac{\partial \mathcal{L}}{\partial \phi_{d,t,k}} = 0 \implies \phi_{d,t,k} \propto \exp\left\{\ln \rho_{k,t} + \left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right)\right) + \left(\sum_{m=1}^M \lambda_{d,m,t} \ln \beta_{z_{y_{d,m},j}}\right)\right\} \tag{3.60}$$

Clearly upto this point there has not been any change in the LeToS model from the UTM model. This has both advantages as well as disadvantages. The obvious advantage is that we do not incur much computational complexity which we would have had we introduced a sentence hierarchy over words as a hierarchy of “is-a” relationships. A a multi-level topic hierarchy [Li and McCallum, 2006, Celikyilmaz and Hakkani-Tür, 2011] in this case is possible but still does not reflect a way to incorporate notions of coherence and attention within the original text. The disadvantage is that introducing sentences as observations introduces a pLSA type of constraint and the model lacks a proper inference scheme outside of documents input for summarization. We remove this constraint in chapter 5 where we address the problem of summarization through both unsupervised topic modeling as well as introduction of more sophisticated linguistic features.

For ζ :

$$\begin{aligned} \mathcal{L}[\zeta] = & \sum_{p=1}^P \sum_{t=1}^T (\Psi(\chi_{d,t}) - \Psi(\sum_{j=1}^T \chi_{d,j})) \zeta_{d,p,t} + \sum_{p=1}^P \sum_{t=1}^T \sum_{u=1}^U \zeta_{d,p,t} s_{d,p}^u \ln \Omega_{t,p} \\ & - \sum_{p=1}^P \sum_{t=1}^T \zeta_{d,p,t} \ln \zeta_{d,p,t} + \sum_{p=1}^P \mu_p (\sum_{t=1}^T \zeta_{d,p,t} - 1) \end{aligned} \quad (3.61)$$

where μ_p are the P Lagrange multipliers in (3.61), one for each sentence in the document d .

$$\frac{\partial \mathcal{L}}{\partial \zeta_{d,p,t}} = 0 \implies \zeta_{d,p,t} \propto \Omega_{t,p} \exp\{\Psi(\chi_{d,t}) - \Psi(\sum_{j=1}^T \chi_{d,j})\} \quad (3.62)$$

3.8.2.2 MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

Here we calculate the maximum likelihood settings of the parameters that do not grow with the data. So we need to take into account the contribution of these for all the documents and not just a single document.

For ρ :

$$F[\rho] = \sum_{d=1}^D \sum_{t=1}^{T_d} \sum_{k=1}^K \sum_{g=1}^{T_G} \phi_{d,t,k} (\ln \rho_{k,t}) r_{dt}^g + \sum_{k=1}^K \mu_k \left(\sum_{g=1}^{T_G} \rho_{k,g} - 1 \right) \quad (3.63)$$

where the μ_t 's are the K Lagrange multipliers in (3.63)

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \rho_{k,g}} = & \sum_{d=1}^D \sum_{t=1}^T \phi_{d,t,k} r_{dt}^g \frac{1}{\rho_{k,g}} + \mu_k \frac{\partial \mathcal{L}}{\partial \rho_{k,g}} = 0 \implies \rho_{k,g} = - \frac{\sum_{d=1}^D \sum_{t=1}^T \phi_{d,t,k} r_{dt}^g}{\mu_k} \\ \implies & \mu_k = - \sum_{g=1}^{T_G} \sum_{d=1}^D \sum_{t=1}^T \phi_{d,t,k} r_{dt}^g \\ \therefore \frac{\partial \mathcal{L}}{\partial \rho_{k,g}} = 0 \implies & \rho_{k,g} \propto \sum_{d=1}^D \sum_{t=1}^T \phi_{d,t,k} r_{dt}^g \end{aligned} \quad (3.64)$$

For β :

$$\mathcal{L}[\beta] = \sum_{m=1}^M \sum_{k=1}^K \sum_{j=1}^V \sum_{t=1}^T \lambda_{d,m,t} \phi_{d,t,k} \ln \beta_{z_{y_{d,m,j}}} w_{d,m}^j + \sum_{k=1}^K \mu_k \left(\sum_{j=1}^V \beta_{k,j} - 1 \right) \quad (3.65)$$

where μ_k s are the K Lagrange Multipliers in (3.65)

$$\frac{\partial \mathcal{L}}{\partial \beta_{k,j}} = 0 \implies \beta_{k,j} \propto \sum_{d=1}^D \sum_{m=1}^M \left(\sum_{t=1}^T \lambda_{d,m,t} \phi_{d,t,k} \right) w_{d,m}^j \quad (3.66)$$

For Ω :

$$\mathcal{L}_{[\Omega]} = \sum_{d=1}^D \sum_{p=1}^{P_d} \sum_{t=1}^{T_G} \sum_{u=1}^U \zeta_{d,p,t} s_{d,p}^u \ln \Omega_{t,u} + \sum_{t=1}^{T_G} \mu_t \left(\sum_{u=1}^U \Omega_{t,u} - 1 \right) \quad (3.67)$$

where μ_t s are the T_G Lagrange Multipliers in (3.67)

$$\frac{\partial F}{\partial \Omega_{t,u}} = 0 \Rightarrow \Omega_{t,u} \propto \sum_{d=1}^D \sum_{p=1}^{P_d} \zeta_{d,p,t} s_{d,p}^u \quad (3.68)$$

For α :

For η :

The updates are exactly the same as in the UTM model.

This is the general form of derivations which we will be using in deriving the fixed point update equations for the rest of our new topic models in Chapters 4 and 6.

Chapter 4

Bi-Perspective Topic Models

“If you have an apple and I have an apple and we exchange these apples then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas.” -

George Bernard Shaw

4.1 Introduction

In this chapter we lay down a robust topic modeling framework to solve the problem of discovering latent topics from documents tagged from two different perspectives. Documents usually consist of at least two perspectives—a document level perspective and a word level perspective—these perspectives being implicit in general. The document level perspective summarizes the contents as a small bag-of-words, while the other perspective annotates the content in different ways. Tables 4.1, 4.2 and 4.3 show different examples of such perspectives. In this paper, it is assumed that tags are non-hierarchical concepts. These concepts can be represented by words or by some other higher order representation that eventually denotes a concept. Many times, for example, in plain text documents, these two perspectives are not explicitly shown. However, documents hosted by today’s interactive websites are a rich source of document tagging from *at least* two perspectives.

Word level annotations	Image captions	Category labels
Syringa/0 (Lilac)/0 is/0 a/0 genus/0 of/0 about/0 ... Lilac/1 bushes/1 can/1 be/1 prone/1 to/1 powdery/1 mildew/1 disease/1 ... RHS/2 Dictionary/2 of/2 Gardening./2 Macmillan/2 ISBN/2 0-333-47494-5/2	Syringa josikaea, Syringa vulgaris shrub in flower, etc.	Syringa, Garden plants, Flowers, Shrubs

Table 4.1: Document with word level “Position” tags and document level image caption word tags. The ellipses (...) indicate substantial skip of paragraphs [source: <http://en.wikipedia.org/wiki/Syringa>]

Table 4.1 shows an article on lilac flower in Wikipedia. Words in the document body annotated with a slash ‘/’ denotes a word level perspective which in this case is the position of the section in which the word appears. The section offsets are binned into three positions relative to the beginning of the

document - begin(0), middle(1) and end(2). Positions signify the importance of the choice of words that constitute sections in a document and they are also useful in tasks like multi-document summarization [Yih et al., 2007]. The document level perspective is assumed to be captured by the images which are described by the corresponding captions. In table 4.1, the third column represents the “ground truth” category labels which are a set of manually edited tags that summarize the Wikipedia article. With this structure of multimedia documents, several questions come to the forefront: “Can there be some way of using image captions to automatically suggest specific category labels for new articles? If so how good will those suggestions be? Further, can we discover latent topics or themes and label each theme with a multimedia object?”

The generative process of word generation in these kinds of documents hinges on the following intuitions: Documents are distributions over latent topics which are further conditional on the word level annotation classes. Latent topics, in turn, are distributions over observed document level tags and the main content words such that the most probable observed variable ensembles for a topic are related through the assumptions of the generative process. For one particular assumption, a word conditioned on the word level annotation observed at the word’s position is sampled independently of the document level tags from a topic. The topic proportions for that document only depends on the expected number of document level tags and the annotated words being assigned to each topic. In another assumption, topics generate the document level tags first. Then a document level tag position is chosen and a word conditioned on the word level tag observed at the word’s position is sampled from the corresponding topic in the position of the document level tag. For this second assumption, there is a more stricter enforcement of words to document level tags and is thus more intuitive from the document generation point of view. For example, a writer often “thinks” of a mental image/concept and then writes words that elaborate that image/concept. Although modeling multimedia Wikipedia articles serves as the primary motivation for developing the proposed models, the models are extremely generic and has been applied to a variety of other datasets for different tasks. For brevity, the document level tags are dubbed DL tags and the word level annotations i.e annotation classes are dubbed WL tags.

Table 4.2 shows an example where DL tags are abstracted at a level higher than words and we have come across this perspective in Chap. 3. In this example, the sample sentence in the table is an excerpt from a newswire article in the dataset used in the DUC2005 Summarization track [Dang, 2006a].

Word level (WL) tags	Document level (DL) tags
Some 167/NE-NUM people were arrested in the US/NE-LOC, including a senior executive of Columbia/NE-LOC’s national bank.	→ne, ne→-, →subj, subj→subj, nn→vb, vb→vb, →adj, adj→-, etc.

Table 4.2: Document with word level “Named Entity” annotations and document level “Named Entity as well as semantic and syntactic role transition” tags

The WL tags denote a particular named entity class like PERSON, LOCATION, ORGANIZATION, NUMBER, etc. being ascribed to each particular word or not. The DL tags, however, are indicative of discourse coherence markers. These markers are constructed following the techniques used in [Barzilay and Lapata, 2005a, Das and Srihari, 2009]. Each word in the document is associated with a grammatical or semantic role (GSR in short) like named entities (ne), nouns (nn), adjectives (adj), verbs (vb), subjects (subj), objects (obj), etc. As in Chapter 3, a GSR transition (GSRt in short) is a relation between the same normal form of a word that is either present in two contextual sentences or in a single sentence,

e.g. a GSRt for the word “car” can be $(subj,obj)$ leading to $(subj \rightarrow obj)$ or $(subj \rightarrow -)$ if the word “car” is not seen in the succeeding sentence. It is reported in [Barzilay and Lapata, 2005a] that a set of sentences with the same entities in roles like “subj,” “obj,” etc. are indicative of coherent passages. Although in [Barzilay and Lapata, 2005a], only entities are involved in GSRts, however, in this paper, words that are not entities are also considered since in quite a few cases, the foci of attentions are based not just on entities. Thus, the document level perspective for DUC 2005 newswire data mentioned in Section 4.4 is that of syntactic coherence. These kinds of user choice specific document level perspective emphasizes the fact that the proposed models are extremely flexible to incorporate any word annotation classes and document level tags.

4.1.1 Descriptions of Annotations in Datasets

In this section we touch upon the word level annotations for the three datasets which we have used for our experiments. All such annotations have a bag-of-words representation with no inter annotation dependencies modeled. For Wikipedia, we collected only those documents which had embedded images. The category labels of the Wikipedia articles are not used as DL tags, rather the image caption words are used. Generally, if captions are not available, an initial preprocessing can be done using the work in [Feng and Lapata, 2010a]. Words in the article title is also added to the list of DL tags. Each word in the main body of the article is annotated with the *positional* information of the sections they appeared in and has been labeled as $\{Begin, Begin_Middle, Middle, Middle_End, End\}$.

Unprocessed Amazon product reviews from the dataset used in [Blitzer et al., 2007] (henceforth the AR dataset) has been used in the experiments. The words in each review were tagged with affect labels using a simple lexicon lookup from the dataset created in [Bradley and Lang, 1999]. The lexicon consists of 2476 words that elicit human emotions in some form. The emotions were labeled with $\{Unhappy, Unsatisfactory, Melancholic, Despair, Hopeful, Contended, Satisfied, Pleased, Happy, Untagged\}$ tags based on the maximum valence values of the affect words. Non-affect words were tagged as “*Untagged*”. The AR dataset did not have product tags and hence the product name and the review title were used as “captions” for reviews which served as DL tags. Finally note that for WL tagging, a word can be conditioned on **only a singleton tag annotation**. Table 4.3 shows an example for the AR dataset used.

Word level tags	DL tags
What I like/CONTENDED is the exceptional zooming without loss/MELANCHOLIC in clarity.	compact camera, Ikon 550, 18X zoom {Rating: 4.0}

Table 4.3: Document with word level emotion tags and document level product feature tags

The DUC (Document Understanding Conference) 2005 dataset consists of newswire articles organized in 50 folders or document sets (docsets) with each folder consisting of at least 25 sizable news reports. The documents are processed to extract named entities and roles of the words using the Stanford CoreNLP toolkit¹. The GSRs are obtained using the dependency parse information. However, co-reference resolution is not performed due to unsatisfactory results. An example of this kind of tagging has been shown previously in Table 4.2. Lemmatized forms of the words are used e.g., “arrested” (verb) and “arrests” (noun) have the lemmatized form “arrest.” A total of 9 GSRs are chosen (Named Entities or ne, Subjects or subj, Objects or obj, Nouns or nn, Verb or vb, Adjective or adj, Adverb or

¹<http://nlp.stanford.edu/software/corenlp.shtml>

adv, Other as ow and Null as “-”) resulting in a total of 81 GSRts. Note that if a word has several GSRs associated with it, only one is chosen using the priority rule: ne > subj > obj > nn > adj > vb > adv > ow. The task in the DUC 2005 Summarization task has been the creation of 250 word multi-document summaries for each of the docsets in response to the corresponding information needs. However, in this paper the docsets from DUC 2005 dataset are used to validate entity-pair relationship discovery and not for summarization.

4.1.2 Improving Existing Tag Topic Models

Tag-topic models have been explored recently [Ramage et al., 2009b, Si and Sun, 2009, Bao et al., 2009, Zhu et al., 2006] as ways of improving word-based topic models with additional information in the form of tags, usually arising out of a **single** perspective. Existing mixed membership tag topic models [Ramage et al., 2009b, Si and Sun, 2009] (c.f. fig. 4.1b) can fit a number of latent topics to the documents without words and DL tags having direct correspondence with each other. However, there can be additional word level annotation information which are implicitly attributed to the words.

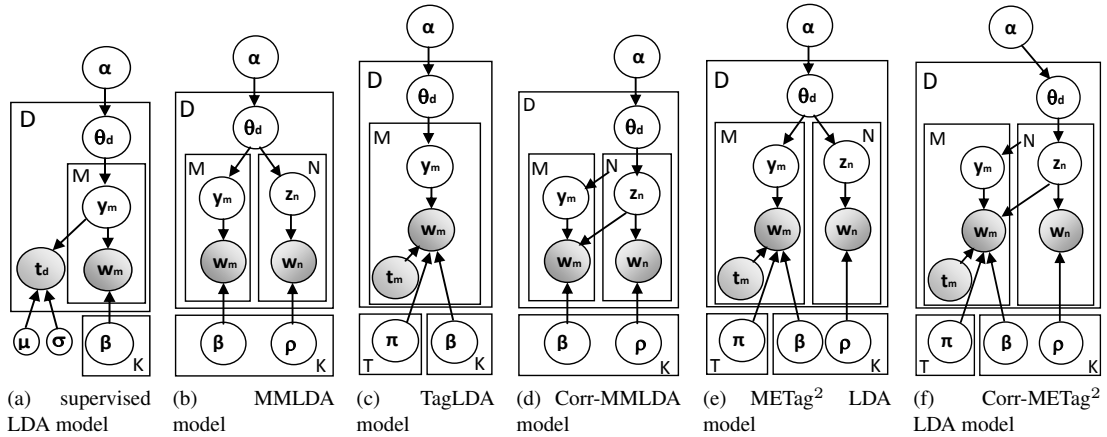


Figure 4.1: Graphical model representations of one supervised topic model, two existing tag topic models, one extended tag topic model and two new tag squared topic models

Thus existing tag topic models of documents focus either on document level tags [Ramage et al., 2009b, Si and Sun, 2009] or on word level annotations [Zhu et al., 2006] (c.f. fig. 4.1c). For models like those in [Ramage et al., 2009b, Si and Sun, 2009] (which are referred to as MMLDA - short for Multi(nomial) Multinomial LDA) each content word and DL tag is generated independently by choosing a topic and then choosing a content or DL tag. The expected number of words in the document’s topic thus depends on the counts of both the content and the DL words ascribed to that topic. The intuition behind the MMLDA model is as follows: We imagine a scale where the MMLDA model puts probability masses of the most probable topics over content words on one pan and the masses of the most probable topics over document level tags in another pan in such a way so as to keep the pans as balanced and the arms as horizontal as possible.

On the other hand, in the TagLDA model in [Zhu et al., 2006] the content words are generated by choosing a topic and drawing a word from the topic’s distribution but conditioned on the WL tag associated with that content word. This conditional aspect allows one to explore related words sharing the same semantic relatedness but conditioned along a particular facet - for e.g. all “PERSON” named

Model Highlights	sLDA	MMLDA	TagLDA	Corr - MMLDA	ME Tag ² LDA	Corr-ME Tag ² LDA
Generate words and DL tags from same topic?	×	✓	×	✓	✓	✓
Suggest related DL tags?	×	✓	×	✓	✓	✓
Associate words to DL tags probabilistically?	×	×	×	✓	×	✓
Decompose topics conditioned on WL annotations?	×	×	✓	×	✓	✓
Find topical-WL annotation orientation of new documents?	×	×	✓	×	✓	✓
Document “label” prediction?	✓	×	×	×	×	×

Table 4.4: Model features and their comparison

entities in a dataset that are semantically related through some hidden topic. Figures 4.1b and 4.1c show the existing tag topic models - MMLDA [Ramage et al., 2009b] and TagLDA [Zhu et al., 2006]. The model in figure 4.1d is implemented in this study as an improvement over MMLDA following [Blei and Jordan, 2003] for the text domain and is referred to as Corr-MMLDA. However, none of MMLDA, TagLDA or Corr-MMLDA addresses a tag space that is split across two different perspectives. The proposed TagSquaredLDA (abbreviated as Tag²LDA) models: METag²LDA model (ME is abbreviated form of Multinomial Exponential) (Fig. 4.1e) and Corr-METag²LDA model (Fig. 4.1f) allows topic modeling of documents with both DL and WL tags. Table 4.4 shows the relative merits and de-merits of each model discussed in this paper. Fig. 4.1a shows a supervised LDA topic model (sLDA) [Blei and McAuliffe, 2007] that is only used for predictive power comparison on the AR dataset. Experiments reveal the improvements of the Tag²LDA models over current tag topic models through better log likelihoods, or more predictive power for topical inference. Also an HMM type of model is not suitable for positional WL tagging, since, then during inference, there is nothing to *infer* on position “states”—they are implicit in any document.

Table 4.5 shows two topics from our small in-house Wikipedia collection. The topics are conditioned on facets that represent the position of the sections (binned into five major categories) to which the words belong. The topic marginals show only the topic distributions over words which is the parameter β_k in Fig. 4.1f. Words from the image captions which are possible candidates for tagging the document (as document metadata) are shown the rows beginning with “Tag suggestion from image captions.” The rows beginning with “Document content and image caption correspondences” show some possible correspondences of words that relate words in the main document body to those in the image captions within the same document.

4.1.3 Applications and Quantitative Measures

Measuring model perplexity [Blei et al., 2003] or equivalently log likelihood on held out test data is an established way of showing how good a model explains the observations. For reasonable sized datasets, the lower bounds on the true log likelihoods of held out test data (ELBO in the case of approximate inference) are also used and this is one of the measures we use to evaluate our models. However, while applying the models for a specific task, the goal is not only to measure held-out test data likelihood for a model. For example, for the Wikipedia data, it is important to have a quantitative measure of confidence

	Beginning →	Beginning To Middle →	Middle →	Middle To End →	End
Topic 175	galaxy, largely, result, Star, early, groups, region, production, active, observed	galaxy, Star, largely, groups, production, active, region, Universe, gas, structure	galaxy, Star, Universe, active, production, small, high, structure, study, Formation	Universe, largely, History, billion, small, production, mass, study, hydrogen, dwarf	University, press, Formation, study, active, remained, Andromeda, space, deep, History
	<i>Topic marginal:</i> galaxy Star spiraled milky matter cluster Hubble gas Universe structure Formation elliptical active galactic nebula dwarf				
	<i>Tag suggestion from image captions:</i> Galaxy, spiral, stars, Hubble, classification, Andromeda, rings, core, Great, compared				
	<i>Document content and image caption correspondences:</i> (Planet, Hubble) (Planet, object) (Planet, galaxy) (Herschel, Hubble) (ring, galaxy) (Heat, galaxy) (discoveries, Hubble)				
Topic 196	air, waters, Ice, light, fog, Areas, temperature, Day, creates, surface, layer, power, nations, cool, salt	air, fog, waters, common, Ice, light, temperature, point, Areas, formed, million, pressure, cloud, ground	air, waters, Ice, light, fog, million, Day, nations, temperature, small, cloud South, ground, region, Ocean	Ice, fog, air, light, waters, Day, Sea, nations, surface, pressure, temperature, Gallery, layer, winter, high, California	air, Gallery, pressure, forced, rises, shuttle, low, million, space, Ca., Florida, Shadow, cases, fog, Ice, Press, ISBN
	<i>Topic marginal:</i> fog air Shadow Ice condensation light vapor Humidity layer temperature freeze particle cool waters moisture evaporation salt				
	<i>Tag suggestion from image captions:</i> fog, Francisco, San, visible, high, temperature, streets, photo, Bai, lake, California, bridge, air				
	<i>Document content and image caption correspondences:</i> (dimensions, high) (beam, visible) (parallel, bridge) (droplets, fog) (combustion, temperature) (invisible, visible) (absorbed, air)				

Table 4.5: Topics and correspondences from the Corr-METag²LDA for the Wikipedia data for $K = 200$

between probable document tags from image caption words and ground truth category labels. To enable this kind of a comparison, a measure of semantic relatedness using path separation between concept pairs [Pedersen et al., 2004] in WordNet ontology was chosen as an evaluation tool. As an example, the connection between “fire_engines” and “fire_extinguisher” can be described by a shortest path linking these two concepts in WordNet as “*fire_extinguisher* ↔ device ↔ instrumentality ↔ container ↔ wheeled_vehicle ↔ self-propelled_vehicle ↔ motor_vehicle ↔ truck ↔ *fire_engine*” with a path length of nine and a simple “inverse of path length” similarity score of 0.11. Under this measure, a value of 1 indicates exact match or parent/child relationship. Using this kind of evaluation, users can be *explained* a “chain of reasoning” that relates a probable DL tag to a ground truth category label for a new document. For N suggested DL tags and C category word labels, scores for all possible $N \times C$ pairs P were obtained. The highest score served as a measure of DL tag suggestions. If a model captures caption words that happens to have shorter path distances to ground truth labels, then the model is scored higher. Note that WordNet is chosen since it is widely accepted—any other customized ontology can easily be pugged in depending upon the application.

Also note that Newman et al. [Newman et al., 2010] attempted to measure cohesiveness of topic

“labels” consisting of top 10 high probability words, where the “results over WordNet are patchy at best.” The WordNet evaluation presented here is not to measure topic cohesiveness but to measure and explain the goodness of a probable DL tag. Although measuring topic coherence is equivalent to measuring topic intrusion [Chang et al., 2009], the notion of a topic is only a mathematical convenience for a low-dimensional subspace that tries to capture the assumptions of the statistical generative model. So qualitatively, a topic is best interpreted by the task on which the model is adapted and its corresponding assumptions.

For the DUC 2005 dataset where the WL tags come from named entity classes, all pairs of PERSON named entities from documents in each of the 50 docsets are collected. For a particular proposed Tag²LDA model, hidden topics are inferred for these documents and pairs from top N entities from the “PERSON” facet of the topic were collected. Entity pairs that co-occur in a sentence are chosen as ground truth pairs that are strongly related—this is the baseline. The average of ratio of the counts of the PERSON entity pairs from topics to those from the baseline serve as a quantitative measure of improvement over the latter in the entity relationship discovery application. Some qualitative results are shown in Table 4.10. Note that the same named entity can occur across multiple docsets. This is particularly true of NUMBER and LOCATION classes and entities related to governments.

4.2 Related Work

Joint topic and tag analysis has been used in some recent works including [Ramage et al., 2009b, Si and Sun, 2009, Zhu et al., 2006] which has culminated in the creation of variants of topic models like LDA [Blei et al., 2003]. The principle shortcoming of these papers are the use of a single tagging perspective - either document level tags or word level tags. While the models in [Ramage et al., 2009b] and [Si and Sun, 2009] are essentially the same, the purposes of the models are a little different. Both use generative models of words and document tags to discover latent topics. In [Ramage et al., 2009b] the topic-tag and the topic-word features were used to better cluster tagged documents. In [Si and Sun, 2009] new documents are “folded-in” in the latent topic space and tags are predicted based on the inferred topic tag distribution. The work in [Zhu et al., 2006] is useful in the sense that the topics are discovered w.r.t to words being conditional on WL tags.

A recent work on topic-perspective modeling has been done in [Lu et al., 2010] where the authors have tried to use perspectives as hidden states that represent a discreet distribution over tags. It is important to note that the “corrLDA” model referred to in [Lu et al., 2010] is not a true correspondence model as there happens to be no direct correspondence between words and tags. Further, although there is a connection between the user-perspective and perspective-tag distributions, the connections between the perspective-tag distribution and the topic-word/topic-tag distributions weakly depend on a binary switching variable. The sLDA [Blei and McAuliffe, 2007] model discovers topics based on the ensembles of document contents and the response variables.

The values of the response variables are explained by the frequency counts of the words in the corresponding documents only. The labeledLDA [Ramage et al., 2009a] model establishes a one-to-one correspondence to the latent topics and the actual document tags. This is done in a manner similar to imposing a non-uniform prior on the latent topic proportions per document [Wallach et al., 2009]. Further the words in text are corresponded to topic labels which precludes any possibility of using WL conditional tags. The proposed Tag²LDA models use both joint (words and DL tags) and conditional

(words and WL tags) modeling, thereby allowing a richer document structure to be captured.

4.3 The Proposed Tag Squared LDA Models

This section introduces the model description and the model parameters for the Tag²LDA models. In all model figures in Fig. 4.1, the symbol notations and their meanings given in table 4.6 are adhered to.

Symbol	Meaning (<i>r.v.</i> = <i>random variable</i>)
D	total number of documents
N	total number of unique document level tags per document $d \in D$
M	total number of unique words per document $d \in D$
α	r.v. for Dirichlet prior for the document level topic proportions
θ_d	r.v. for document level latent topic proportions
ρ	r.v. for corpus level topic-DL_tag multinomial
β	r.v. for corpus level (marginal in figs. 4.1c, 4.1e and 4.1f) topic-word distribution
π	r.v. for corpus level marginal tag-word distribution
z_n	indicator variable for DL topic proportion
y_m in figs. 4.1b, 4.1c and 4.1e	indicator variable for DL topic proportion
y_m in figs. 4.1d and 4.1f	indicator variable for DL tag correspondence
w_n	r.v. for DL tag at position n ; vocabulary size $corrV$
w_m	r.v. for word at position m ; vocabulary size V
t_m in figs. 4.1c, 4.1e and 4.1f	r.v. denoting tag at position m , on which word w_m is conditioned; vocabulary size T
t_d in fig. 4.1a	r.v. denoting observed response for document d
μ, σ in fig. 4.1a	r.v.s denoting mean and standard deviation for the observed response for document d [Blei and McAuliffe, 2007]

Table 4.6: Symbols used in this chapter and their meaning

Note that in the Tag²LDA models, $p(w_{d,m}|C = i, \beta, \pi, t_{w_{d,m}})$, where $C = y_{d,m}$ or $C = z_{y_{d,m}}$, is not a simple topic multinomial anymore, but is distributed as

$$p(w_{d,m}|C = i, \beta, \pi, t_{w_{d,m}}) = \frac{\exp(\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{w_{d,m}},w_{d,m}})}{\sum_{v=1}^V \exp(\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{w_{d,m}},w_{d,m}})} \quad (4.1)$$

Note that each π_t is also a distribution over V . In essence the π parameter imparts domain knowledge to the observations in the form of word level annotation classes being ascribed to each and every word. Simplified Gibbs sampling exploiting Multinomial-Dirichlet conjugacy (as in [Griffiths and Steyvers, 2004]) cannot be applied in this setting. For notational convenience, in the correspondence models where $y_{d,m} \sim Unif(N_d)$, “Unif” is short for Uniform distribution, as in [Blei and Jordan, 2003]. We now illustrate the generative processes for the proposed Tag²LDA models:

For each document $d \in 1, \dots, D$

Choose a topic proportion $\theta|\alpha \sim Dir(\alpha)$

For each “document level” position n in document d

Choose topic indicator $z_n|\theta \sim Mult(\theta)$

Choose a “document level” tag $w_n|z_n = k, \rho \sim Mult(\rho_{z_n})$

For each “word level” position m in document d

Choose $y_m \sim Unif(\mathbf{1}, \dots, \mathbf{N})$ (for Corr-METag²LDA - fig. 4.1f)
or Choose $y_m | \theta \sim Mult(\theta)$ (for METag²LDA - fig. 4.1e)
Choose a word $w_m | y_m, \mathbf{z}, \mathbf{t}, \beta, \pi \sim p(w_m | z_{y_m}, \beta, \pi, t_m)$ (fig. 4.1f)
or Choose a word $w_m | y_m, \mathbf{t}, \beta, \pi \sim p(w_m | y_m, \beta, \pi, t_m)$ (fig. 4.1e)

In the correspondence models, the DL perspective plays a significant role in the quality of topic coherence. For example, when the GSRt perspective (c.f. table 4.2) is chosen as a DL perspective, the topics will capture words that are both co-occurring and generated from similar roles. To reiterate the example from Chapter 3, there can be a docset on “Global warming” which will tie together words like planet and ice based on co-occurrence alone. Consider another docset concerning discovery of ice on Pluto’s surface. Thus, if the GSR of “ice” is taken to be a subject, then a “Global warming” topic can include Pluto as a probable word because of the GSRts for the word “ice” that involve “subj”. This type of tagging is beneficial where the task is not to replicate the docset structure based on co-occurrence but to provide deeper insights into data for tasks such as summarization, relationship extraction, etc. This effect is hardly observed when the DL tags tersely summarize the main contents of the documents and the DL tag-topic matrix is close to p -separable [Arora et al., 2013].

In general, if the features of the DL perspective have low variance, then the assumption that topics generate the observations in the DL perspective first and then a word in the main document body is generated by the topic of a randomly selected DL observation. In that respect if there are too many competing topics for the DL perspective then the topical correspondence of the DL perspective to the WL perspective degrades in quality in the sense that the average entropy of the variational word distributions over topics increases.

4.3.1 Latent variable inference

The variational Bayesian Expectation Maximization algorithm (see Chapter 1, Section 2.6) has been used to maximize the lower bound to the true intractable likelihood of the data w.r.t. the model parameters. This section outlines the various updates of the latent variables and the parameters and subsection 4.3.3 outlines a general plan of implementation. To find as tight as possible an approximation to the log likelihood of the data (the joint and conditional distribution of the observed variables given the parameters), the KL divergence of an approximate factorized mean field distribution is minimized to the true posterior distribution of the latent variables given the data. A fully factorized q distribution with “free” variational parameters γ , ϕ and λ is imposed as

$$q(\boldsymbol{\theta}, \mathbf{z}, \mathbf{y} | \gamma, \phi, \lambda) = \prod_{d=1}^D q(\boldsymbol{\theta}_d | \gamma_d) \left[\prod_{n=1}^{N_d} q(z_{d,n} | \phi_{d,n}) \prod_{m=1}^{M_d} q(y_{d,m} | \lambda_{d,m}) \right] \quad (4.2)$$

and then optimal values of free variables and parameters are found by optimizing the lower bound on $\ln p(\mathbf{w}_m, \mathbf{w}_n | \alpha, \beta, \rho, \pi, \mathbf{t})$. The variational functional to optimize can be shown to be (as in [Beal, 2003])

$$\mathcal{F} = E_q[\ln p(\mathbf{w}_M, \mathbf{w}_N, \boldsymbol{\theta}, \mathbf{z}, \mathbf{y} | \alpha, \beta, \rho, \pi, \mathbf{t})] - E_q[\ln q(\boldsymbol{\theta}, \mathbf{z}, \mathbf{y} | \gamma, \phi, \lambda)] \quad (4.3)$$

where $E_q[f(\cdot)]$ is the expectation of $f(\cdot)$ over the q distribution and \mathcal{F} is the Evidence Lower Bound (ELBO) to true likelihood. This ELBO is directly related to measuring perplexity [Blei et al., 2003]. In the following subsections, it is assumed that K is the number of topics, ϕ to be free parameters of

the variational DL_tag-topic distribution and λ to be the free parameters of the variational word-topic or word-DL_tag distributions. These free parameters are defined for every document $d \in D$.

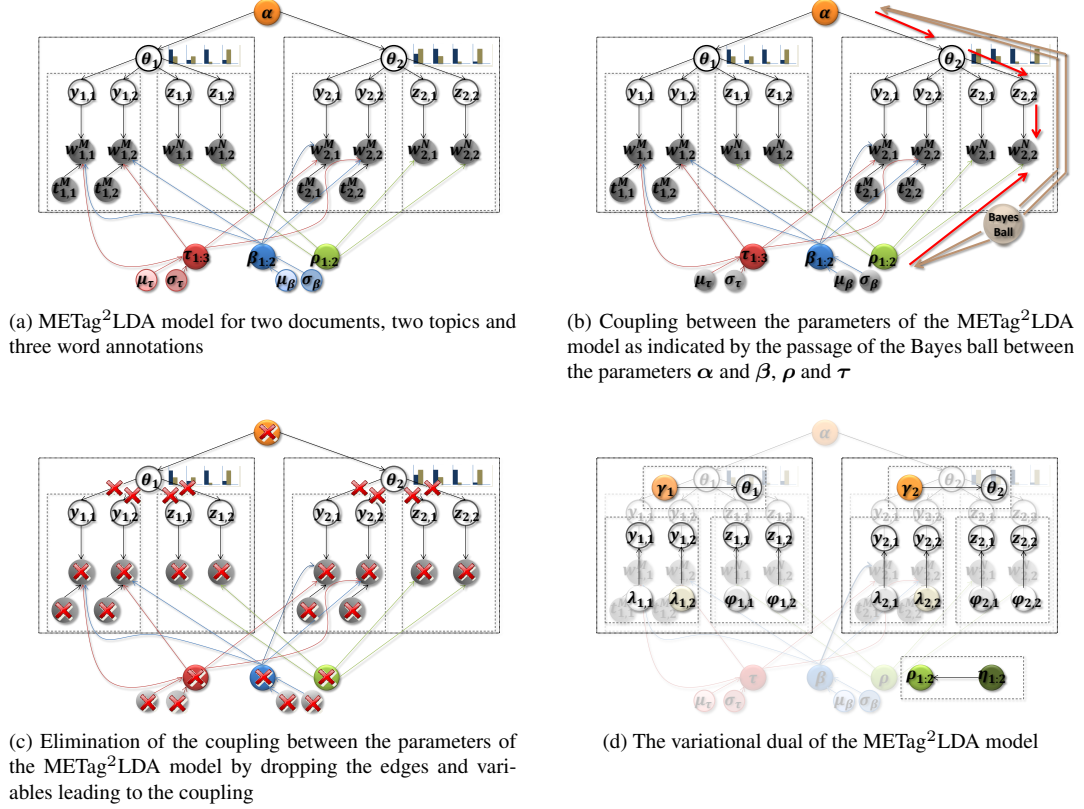


Figure 4.2: Mean field representation of the METag²LDA model

Figure 4.2 shows the juxtaposition of the METag²LDA model with its variational dual. The graphical model representation of METag²LDA is shown in Fig. 4.2a for two documents and two words per document. The number of topics in the illustration is set to two and the number of words in the document level perspective is also set to two in each of the documents. The dependence of the parameters of the model are highlighted by the head-to-head causal arcs to the observed variables which allows the Bayes Ball [Shachter, 1998] to pass through. This coupling within the parameters leads to an exponential state-space exploration to compute the log partition function over the different configurations of the parameters as realized by the indicator variables \mathbf{y} , \mathbf{z} . To eliminate the coupling, the arcs leading upto the observed nodes that allow the Bayes ball to pass through need to be removed (see Fig. 4.2c). The hidden variables which are removed as a side effect of deleting the arcs in the original model are represented as *independent* random variables drawn from distributions which belong to the same family as in the original model but with their parameters allowed to *freely vary* to best fit the observations. This is shown in Fig. 4.2d where θ_d follows Dirichlet γ_d (a surrogate for Dirichlet(α)); \mathbf{y} follows Multinomial(λ) and \mathbf{z} follows Multinomial(ϕ). If there is a prior for ρ , then the variational distribution for ρ in the dual model is Dirichlet(η).

The key inferential problem that is solved here is the learning of the posterior distribution of the

latent variables given the observations and parameters of the models on data that are new on count proportions. Following the inequality, $\ln(x) \leq \zeta^{-1}x + \ln(\zeta) - 1, \forall \zeta > 0$, the ELBO \mathcal{F} is changed to further lower bounds \mathcal{L} for the two models. The inequality is obtained using Taylor series expansion of the logarithm function as follows:

$$\begin{aligned} f(x) &= f(\zeta) + f'(\zeta)(x - \zeta) + \frac{f''(\zeta)}{2!}(x - \zeta)^2 + \dots + \epsilon \text{ [by Taylor series expansion with error term } \epsilon] \\ \implies \ln(x) &= \ln(\zeta) + \frac{1}{\zeta}(x - \zeta) + O(\zeta^2) \\ \implies \ln(x) &\leq \ln(\zeta) + \zeta^{-1}x - 1, \forall \zeta > 0 \end{aligned} \quad (4.4)$$

Thus, for the METag²LDA model, the lower bound to the log likelihood can be written as:

$$\begin{aligned} \mathcal{L}_{ME} &= E_q[\ln p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\ln p(\mathbf{Z}|\boldsymbol{\theta})] + E_q[\ln p(\mathbf{W}|\mathbf{Z}, \rho)] \\ &\quad + E_q[\ln p(\mathbf{Y}|\boldsymbol{\theta})] + E_q[\ln p(\mathbf{W}|\mathbf{Y}, \beta, \boldsymbol{\pi}, \mathbf{t})] - E_q[\ln q(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{Y}, |\gamma, \phi, \boldsymbol{\lambda})] \end{aligned} \quad (4.5)$$

The expression for $E_q[\ln p(\mathbf{W}|\mathbf{Y}, \beta, \boldsymbol{\pi}, \mathbf{t})]$ can be written for a document d as:

$$\begin{aligned} E_q[\ln p(\mathbf{w}_{d,m}|\mathbf{y}_{d,m}, \beta, \boldsymbol{\pi}, \mathbf{t})] &\geq \sum_{m=1}^{M_d} \sum_{i=1}^K \lambda_{d,m,i} (\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}}) \\ &\quad - \sum_{m=1}^{M_d} \left\{ \zeta_{d,m}^{-1} \left(\sum_{i=1}^K \sum_{v=1}^V \lambda_{d,m,i} \exp(\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}}) \right) + \ln \zeta_{d,m} - 1 \right\} \end{aligned} \quad (4.6)$$

Using the new lower bound, the maximum likelihood estimates of the hidden variables in document d are as follows:

$$\zeta_{d,m} = \sum_{v=1}^V \sum_{i=1}^K \lambda_{d,m,i} \exp \{ \ln \beta_{i,v} + \ln \pi_{t_{d,m},v} \} \quad (4.7)$$

$$\phi_{d,n,i} \propto \exp \left\{ \psi(\gamma_{d,i}) - \psi \left(\sum_{j=1}^K \gamma_{d,j} \right) + \ln \rho_{i,w_{d,n}} \right\} \quad (4.8)$$

$$\begin{aligned} \lambda_{d,m,i} \propto \exp \left\{ \psi(\gamma_{d,i}) - \psi \left(\sum_{j=1}^K \gamma_{d,j} \right) + (\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}}) \right. \\ \left. - \zeta_{d,m}^{-1} \sum_{v=1}^V \exp(\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}}) \right\} \end{aligned} \quad (4.9)$$

$$\gamma_{d,i} = \alpha_i + \sum_{n=1}^{N_d} \phi_{d,n,i} + \sum_{m=1}^{M_d} \lambda_{d,m,i} \quad (4.10)$$

For the Corr-METag²LDA model:

$$\begin{aligned} \mathcal{L}_{corrME} &= E_q[\ln p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\ln p(\mathbf{z}_n|\boldsymbol{\theta})] + E_q[\ln p(\mathbf{w}_n|\mathbf{z}_n, \rho)] + E_q[\ln p(\mathbf{y}_m|N)] \\ &\quad + E_q[\ln p(\mathbf{w}_m|\mathbf{y}_m, \beta, \boldsymbol{\pi}, \mathbf{t})] - E_q[\ln q(\boldsymbol{\theta}, \mathbf{z}, \mathbf{y}, |\gamma, \phi, \boldsymbol{\lambda})] \end{aligned} \quad (4.11)$$

The expression for $E_q[\ln p(\mathbf{w}_m|\mathbf{y}_m, \beta, \boldsymbol{\pi}, \mathbf{t})]$ can be written as:

$$E_q[\ln p(\mathbf{w}_m|\mathbf{y}_m, \beta, \boldsymbol{\pi}, \mathbf{t})] \geq \sum_{m=1}^{M_d} \sum_{i=1}^K \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) (\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}})$$

$$- \sum_{m=1}^{M_d} \{ \zeta_{d,m}^{-1} \left(\sum_{i=1}^K \sum_{v=1}^V \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) \exp(\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}}) + \ln \zeta_{d,m} - 1 \right) \} \quad (4.12)$$

Using these lower bounds and the maximum likelihood estimations of the hidden variables in document d are as follows:

$$\zeta_{d,m} = \sum_{v=1}^V \sum_{i=1}^K \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) \exp \{ \ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}} \} \quad (4.13)$$

$$\phi_{d,n,i} \propto \exp \left\{ \psi(\gamma_{d,i}) - \psi \left(\sum_{j=1}^K \gamma_{d,j} \right) + \ln \rho_{i,w_{d,n}} + \sum_{m=1}^{M_d} \lambda_{d,m,n} (\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}}) - \sum_{m=1}^{M_d} \zeta_{d,m}^{-1} \lambda_{d,m,n} \left[\sum_{v=1}^V \exp(\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}}) \right] \right\} \quad (4.14)$$

$$\lambda_{d,m,n} \propto \exp \left\{ \sum_{i=1}^K \phi_{d,n,i} (\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}}) - \zeta_{d,m}^{-1} \left(\sum_{v=1}^V \sum_{i=1}^K \phi_{d,n,i} \exp(\ln \beta_{i,w_{d,m}} + \ln \pi_{t_{d,m},w_{d,m}}) \right) \right\} \quad (4.15)$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N_d} \phi_{d,n,i} \quad (4.16)$$

where $\zeta_{d,m} \geq 0$ is an additional free variable used in the Taylor expansion of $\ln(x)$ to obtain a tractable second lower bound on the probability of word generation given the topic and tag parameters of the model. Note that $\zeta_{d,m}$ is defined for each document $d \in D$ and does not need to be initialized in the routines described in Section 4.3.3.

4.3.2 Maximum Likelihood Parameter estimation

The expressions for the maximum likelihood of the parameters of the original graphical model using derivatives w.r.t the parameters of the functional $\mathcal{L}_{(\cdot)}$ are obtained as follows:

For the METag²LDA model:

$$\rho_{i,j} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{j=1}^{corrV} \phi_{d,n,i} \delta(w_{d,n}, j) \quad (4.17)$$

$$\begin{aligned} \ln \beta_{i,v} &= \ln \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \lambda_{d,m,i} \delta(w_{d,m}, v) \right) - \ln \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \zeta_{d,m}^{-1} \lambda_{d,m,i} \exp(\ln \pi_{t_{d,m},w_{d,m}}) \delta(w_{d,m}, v) \right) \\ &= \ln(\text{term}_1^\beta) - \ln(\text{term}_2^\beta) \end{aligned} \quad (4.18)$$

$$\begin{aligned} \ln \pi_{t,v} &= \ln \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{i=1}^K \lambda_{d,m,i} \delta(w_{d,m}, v) \delta(t_{d,m}, t') \right) \\ &\quad - \ln \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \zeta_{d,m}^{-1} \sum_{i=1}^K \lambda_{d,m,i} \exp(\ln \beta_{i,w_{d,m}}) \delta(w_{d,m}, v) \delta(t_{d,m}, t') \right) \\ &= \ln(\text{term}_1^\pi) - \ln(\text{term}_2^\pi) \end{aligned} \quad (4.19)$$

For the Corr-METag²LDA model:

$$\rho_{i,j} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{j=1}^{corrV} \phi_{d,n,i} \delta(w_{d,n}^j) \quad (4.20)$$

$$\begin{aligned} \ln \beta_{i,v} &= \ln \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) \delta(w_{d,m}, v) \right) \\ &\quad - \ln \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \zeta_{d,m}^{-1} \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) \exp(\ln \pi_{t_{d,m},v}) \delta(w_{d,m}, v) \right) \\ &= \ln(\text{term}_1^\beta) - \ln(\text{term}_2^\beta) \end{aligned} \quad (4.21)$$

$$\begin{aligned} \ln \pi_{t,v} &= \ln \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{i=1}^K \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) \delta(w_{d,m}, v) \delta(t_{d,m}, t') \right) \\ &\quad - \ln \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \zeta_{d,m}^{-1} \sum_{i=1}^K \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) \exp(\ln \beta_{i,v}) \delta(w_{d,m}^v) \delta(t_{d,m}, t') \right) \\ &= \ln(\text{term}_1^\pi) - \ln(\text{term}_2^\pi) \end{aligned} \quad (4.22)$$

where $\delta(x_z, y) = 1$ iff $x_z == y$ and 0 otherwise and $t' \in \{1, \dots, T\}$. Since the updates for β and π are unconstrained, a Gaussian regularizer with 0 mean and constant standard deviation (set to 2 in this paper) is used for **every** $\beta_{i,v}$ and $\pi_{t,v}$. If β and π are in log space as β^ℓ and π^ℓ , then $\mathcal{L}_{i,t}$ is transformed to

$$\widehat{\mathcal{L}}_{i,t} = \mathcal{L}_{i,t} - \frac{1}{2\sigma^2} \left(\sum_{v=1}^V \left(\exp(\beta_{i,v}^\ell) \right)^2 \right) - \frac{1}{2\sigma^2} \left(\sum_{v=1}^V \left(\exp(\pi_{t,v}^\ell) \right)^2 \right) \quad (4.23)$$

So, in the derivative of $\widehat{\mathcal{L}}$ w.r.t $\ln \beta$ or $\ln \pi$ results in a quadratic in $e^{\beta_{i,v}^\ell}$ or $e^{\pi_{t,v}^\ell}$ as $\text{term}_1^{(\cdot)} - \text{term}_2^{(\cdot)} \exp(\cdot) - \frac{1}{2\sigma^2} (2 \times [\exp(\cdot)]^2) = 0$ as a necessary condition for extrema, where (\cdot) is $\beta_{i,v}^\ell$ or $\pi_{t,v}^\ell$. For $\exp(\cdot)$ to be ≥ 0 , the positive root is taken as the only solution. So the solution becomes (letting $A = \exp(\cdot)$),

$$2A = -\sigma^2 \text{term}_2^{(\cdot)} + \sigma \sqrt{\sigma^2 (\text{term}_2^{(\cdot)})^2 + 4 \text{term}_1^{(\cdot)}} \quad (4.24)$$

which is ≥ 0 . In the derivative of $\widehat{\mathcal{L}}$ w.r.t the \ln of β or π , if the regularizer is not used then convergence is not achieved² arising possibly out of the boundaries of the fixed point surface. **This derivation is different from that used in [Zhu et al., 2006]**. Further, while initializing marginal statistics for β and π , random initialization works best. A complete derivation of the extrema expressions for the hidden variables and model parameters is shown in Section 4.6. Note that number of Lagrange multipliers used in the optimization for $\phi_{d,n,i}$ is N_d , that for $\lambda_{d,m,i}$ or $\lambda_{d,m,n}$ is M_d and that for ρ is K . These free(ϕ, λ) and model(ρ) parameters follow multinomial distributions and hence sum to one.

4.3.3 Algorithms for Implementation

Algorithms 4, 5, 6 and 7 outline some computational procedures for implementing the model and corresponding time complexities (given as $\mathcal{O}(\cdot)$). If a procedure is not defined, comments in $\{\cdot\}$ explain the functionality of the procedures.

²Authors thank Jordan Boyd-Graber for the hint on using regularizers

Algorithm 4 VB EM

```
1: if algorithm_mode == "training" then
2:   initialize_statistics(); {use seeded initialization for  $\rho$  and random initialization for  $\beta$  and  $\pi$ }
3:   vb_m_step();
4: end if
5: elbo_prev  $\leftarrow$  0
6: elbo_current  $\leftarrow$  0; iters  $\leftarrow$  0
7: while converged  $\geq$  EM_CONVERGED do
8:   elbo_current  $\leftarrow$  vb_e_step() {update hidden variables}
9:   vb_m_step() {update model parameters}
10:  converged  $\leftarrow$  (elbo_prev - elbo_current)/(elbo_prev)
11:  elbo_prev  $\leftarrow$  elbo_current; iters  $\leftarrow$  iters+1
12: end while [ $\mathcal{O}(\text{iters} \times (\text{vb\_e\_step} + \text{vb\_m\_step}))$ ]
```

Algorithm 5 vb_e_step

```
1: zero_initialize_statistics(); [ $\mathcal{O}(K \cdot \text{corrV} + K \cdot \text{V} + T \cdot \text{V})$ ]
2: precompute_beta_and_pi_row_sums() {precompute  $\sum_{v=1}^V \exp\{\ln \beta_{i,v} + \ln \pi_{t,v}\} \forall i \in \{1, \dots, K\}$  and  $\forall t \in \{1, \dots, T\}$  in an  $K \times T$  matrix } [ $\mathcal{O}(K \cdot T \cdot V)$ ]  $\leftarrow$  This is a necessary computational bottleneck in the TagLDA and Tag2LDA family of models during each VBE Step
3: elbo_current  $\leftarrow$  0
4: for d = 0 to D do
5:   doc  $\leftarrow$  corpus  $\rightarrow$  document_vec  $\rightarrow$  at(d)
6:   elbo_current += doc_e_step(d, doc) {also accumulate term1 $\beta$ , term2 $\beta$ , term1 $\pi$  and term1 $\pi$  of the marginal statistics for  $\beta$  and  $\pi \forall d, w_{d,m}$  and  $t_{d,m}$  c.f. eqs. 4.18, 4.21, 4.19 and 4.22}
7: end for [ $\mathcal{O}(D(\text{doc\_e\_step}))$ ]
8: return elbo_current;
```

Algorithm 6 doc_e_step

```
1:  $\gamma_{d,i} = \alpha + \frac{(\text{documents}[d].\text{total\_num\_words} + \text{documents}[d].\text{total\_num\_corr\_words})}{K}$ 
2:  $\phi_{d,n,i} = \frac{1.0}{K}$ 
3:  $\lambda_{d,m,i} = \frac{1.0}{K}$  {If model is METag2LDA} {OR}  $\lambda_{d,m,n} = \frac{1.0}{\text{documents}[d].\text{unique\_num\_corr\_words}}$  {If model is Corr-METag2LDA}
4: elbo_current  $\leftarrow$  0; v_iter  $\leftarrow$  0
5: while not converged do
6:   update  $\zeta_{d,m}$ 
7:   update  $\phi_{d,n,i}$ 
8:   update  $\lambda_{d,m,i}$  {If model is METag2LDA} {OR} update  $\lambda_{d,m,n}$  {If model is Corr-METag2LDA}
9:   update  $\gamma_{d,i}$ 
10:  elbo_current  $\leftarrow$  compute_likelihood() {To compute likelihoods c.f. equations 4.5 for METag2LDA and 4.11 for Corr-METag2LDA}
11:  v_iter  $\leftarrow$  v_iter + 1
12: end while
13: return elbo_current; [ $\mathcal{O}(K + KN + KM + v\_iter(MK + NK + MK + KN))$ ] for METag2LDA or [ $\mathcal{O}(K + NK + MN + v\_iter(MKN + NKM + MKN + KN))$ ] for Corr-METag2LDA
```

Algorithm 7 vb_m_step

```
1: for all  $i \in 1, \dots, K, v \in 1, \dots, V$  and  $corr\_v \in 1, \dots, corrV$  do
2:   update  $\rho_{i,corr\_v}$  from sufficient statistics
3:   update  $\beta_{i,v}$  from marginal statistics
4:   update  $\pi_{t,v}$  from marginal statistics
5:   update  $\alpha$  {Follow the Newton-Raphson method in [Blei et al., 2003]}
6: end for [ $\mathcal{O}(K \cdot corrV + K \cdot V + T \cdot V)$ ]
```

4.4 Results and Discussions

This section shows the relative performances of the proposed models on DUC 2005, Wikipedia and the Amazon Review (AR) [Blitzer et al., 2007] datasets (see subsection 4.1.1). The AR dataset was further processed to extract not more than 400 reviews per category. The reviews belong to 25 category labels including {apparel, software, magazine, food, etc.}. The Wikipedia documents were crawled using the special export url³ mostly along the categories of {food, animal, countries, sport, war, transportation, natural, weapon, universe and ethnic groups}. The relative positions of the sections were binned into 5 categories which served as WL tags. Standard English stopwords were removed for the Wikipedia data and after processing, it contained 33,261 unique words and 6,902 unique DL tags (bag-of-words from image captions and Wikipedia article names). The AR dataset contained 6017 unique words and 4271 unique DL tags from product names and review titles after processing. Tags from the affect lexicon were used as WL tags.

For both datasets, words occurring once or more than a thousand times across the entire corpus were also removed. Also note that the main document word vocabulary V and the document level tag vocabulary $corrV$ were processed independently using the same token processing rules but without any correspondence. To compare the proposed models with sLDA [Blei and McAuliffe, 2007] on the AR data, all DL and WL tags were discarded for sLDA. Instead, the ratings served as values of the response random variables. For both the datasets the number of topics K were set to {20, 50, 100, 200}.

For the DUC 2005 dataset, the DL tags were the GSRts found in respective documents and their counts (see subsection 4.1.1 for GSRts). Two types of WL tags were considered: word position bins like WL tags for Wikipedia dataset and named entity classes - PERSON, ORGANIZATION, LOCATION, NUMBER and MISC. Together, DL+WL tagging for DUC 2005 data is named as GSRTPos and GSRTNe respectively (see fig. 4.3). Altogether, there were 36725 unique words for the DUC 2005 dataset and 81 corresponding terms which were just the GSRts. K was set to {40, 60, 80, 100} for the DUC2005 data based on human intuitions.

4.4.1 Model Loglikelihoods on Held-out Test Data

To measure predictive power of METag²LDA and Corr-METag²LDA, a 10-fold cross validation was performed on the DUC 2005 dataset. Figures 4.3a and 4.3c show the ELBO's (higher is better) on validation sets averaged over all folds. Figures 4.3b and 4.3d show the **minimum** of the differences in the ELBO's per topic across all folds for the Corr-METag²LDA model vs. Corr-MMLDA and METag²LDA models. Clearly the differences prove that the improved performance of correspondence Tag² model is statistically significant. This is also intuitive since words in a document are always generated from a

³http://en.wikipedia.org/wiki/Special:Export/<Wiki_article_name>

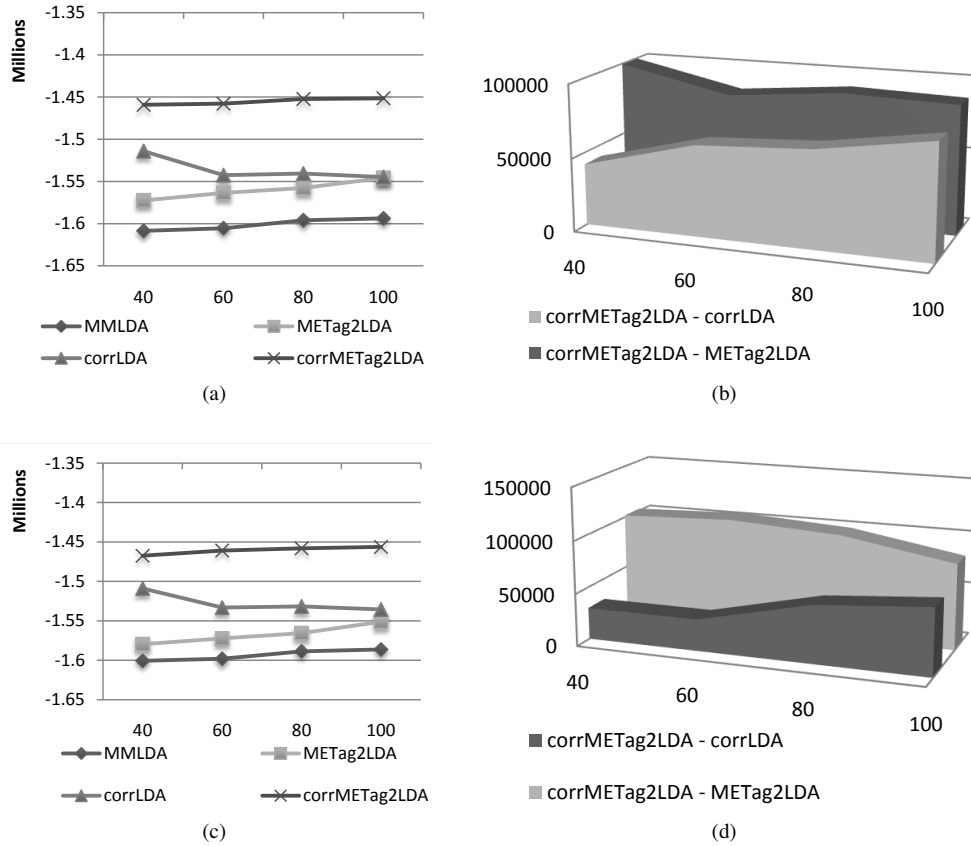


Figure 4.3: Cross-Validation results on DUC 2005 newswire data (**higher is better** in 4.3a and 4.3c): (4.3a) ELBO-Validation DUC 2005 GSRtNe; (4.3b) Minimum of differences in ELBO across topics of Corr-METag²LDA to corrLDA and METag²LDA for GSRtNe tagging; (4.3c) ELBO-Validation DUC 2005 GSRtPos; (4.3d) Minimum of differences in ELBO across topics of Corr-METag²LDA to corrLDA and METag²LDA for GSRtPos tagging

corresponding process, like visualizing an image or action role for a concept. There is a very slightly improved performance when the WL tags are chosen to be named entity classes.

The TagLDA model [Zhu et al., 2006] was not compared for this dataset since the concept of multiple GSRts at the word level breaks down for TagLDA. However, empirically it is seen that the nature of DL tags influences the predictive power of the proposed Tag²LDA models vs. TagLDA. For the DUC 2005 dataset, the DL tags were represented by coherence markers like “subj→subj” etc. as in [Barzilay and Lapata, 2005a]. Typically this type of coherence marker smooths out variations like “landslide:subj→subj,” “car:subj→subj” etc. under a common “subj→subj” abstraction. On the other hand, words like landslide and car signify concepts that allow for identifying specific centers in coherent sentences. In this respect, the WL perspective is more important (primary) over the more abstract DL perspective, the latter capturing a coarser notion of document level coherence. The counts of document level GSRts in the form of “GSR→GSR” do not allow for much variance to be exhibited by the documents at the DL perspective. This fits TagLDA better to the dataset (see Section 4.4.4) at the cost of either ignoring abstract coherence markers altogether or discarding WL perspective and choosing only one coherence marker per word at WL annotation. However, if the document level GSRts are in the

form “word:GSR→GSR”, then the proposed models fit the data much better than TagLDA owing to the variance in the DL observations that are captured nicely in the topics along with the WL variations (see Chapter 5). The GSRts in the latter case cannot be considered as secondary to the WL perspective for document representation.

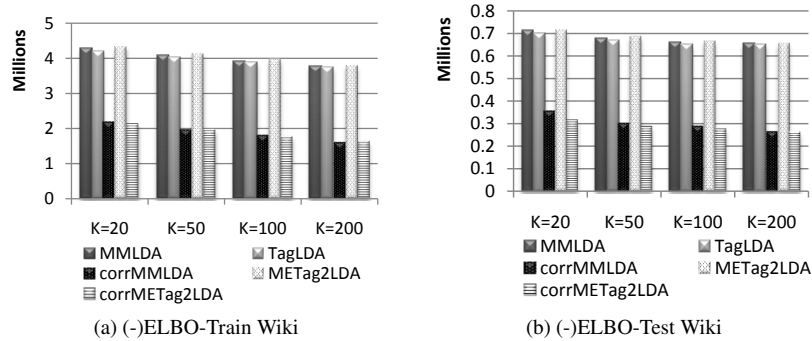


Figure 4.4: Training and test negative ELBO plots of tag topic models on the Wiki data (**Lower is better**). In each K -group, the models from left to right are MMLDA, TagLDA, Corr-MMLDA, METag²LDA and Corr-METag²LDA

Figures 4.4a, 4.4b, 4.5a and 4.5b also show that in the presence of decent variations in DL tags, the Corr-METag²LDA model performs the best in terms of both training ELBO and test ELBO. The meaning of correspondence in terms of the bag-of-words model is to find important associations where the first word comes from a document and the second from DL tags in the same document. Table 4.7 shows some word correspondences that were obtained on test documents from Wikipedia (see rows with $\lambda_{m,n}^{(.)}$).

For the Wikipedia dataset, the mixed-membership Multi-Multinomial (Exponential) class of models: MMLDA [Ramage et al., 2009b, Si and Sun, 2009] and METag²LDA (fig. 4.1e) perform worst. TagLDA [Zhu et al., 2006] performs a little better. This trend is seen on both the training and test sets. Note however that METag²LDA does a simultaneous joint and conditional modeling of DL and WL tags w.r.t. the document’s words. Thus it, along with Corr-METag²LDA, captures what MMLDA, Corr-MMLDA and TagLDA individually misses out. The ELBO trends of the correspondence class of LDAs are quite similar, with Corr-METag²LDA beating Corr-MMLDA. Again, this trend is seen on both the training and test sets. For the AR dataset, the Corr-METag²LDA model beats all other models convincingly in both the training and test set ELBOs. ELBO of MMLDA is the highest during training, followed by (supervised) sLDA [Blei and McAuliffe, 2007] and the predictive power of sLDA decreases even further on the test set. In the AR dataset, TagLDA performs much better due to less variability in DL tags. The proposed Corr-METag²LDA combines the best of TagLDA and Corr-MMLDA to achieve the best predictive power on the AR dataset consisting mostly of very short review documents.

Table 4.7 shows some topics from Wikipedia dataset corresponding to the best performing Corr-METag²LDA model. Note that “positional facets” of topics 175 and 196 have been collapsed for space limitations. The test documents for these collapsed topics were the Wikipedia articles on “galaxy” and “fog”. The learned β parameters contributing marginally to word generation are listed for the collapsed topics. Top suggested tags from image captions for the test documents also appear as the re-weighted ρ topic multinomial over all DL tags after document inference. Top correspondence tuples are listed as

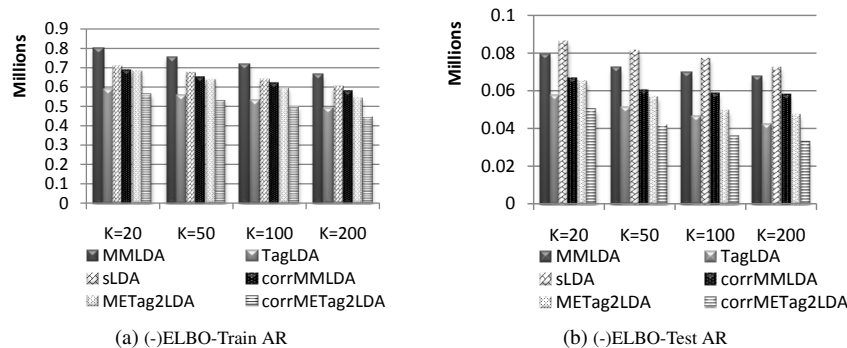


Figure 4.5: Training and test negative ELBO plots of tag topic models on the Amazon Review (AR) data (**Lower is better**). In each K -group, the models from left to right are MMLDA, TagLDA, sLDA, Corr-MMLDA, METag²LDA and Corr-METag²LDA

$\lambda_{m,n}$. For topics 54 and 76, notice how there is a “drift” from the beginning sections of the Wikipedia articles to the end sections. Words like “University Press, ISBN” have high mass on the “Middle_to_end” and the “End” facets of the topic 76. The image labels for topics are obtained from an inverted index of DL tags to thumbnail image files. From figures 4.4 and 4.5, 200 topics are good fits to both the Tag² topic models for both Wikipedia and AR datasets.

Similarly, in table 4.9 only three topics out of 200 are shown due to space constraints. Note here that for topic 35, although the review was titled “too many ads, too little substance,” the word “ads” has obtained a higher probability mass conditional on “CONTENDED.” This has happened because of the bag-of-words assumption in emotion tagging. In many reviews showing low-ratings, the number of negative affect words are outweighed by the number of positive affect words. However, taken in context, for e.g. a phrase, the positive affect is actually a part of a negative connotation like “do not like,” etc. This is a shortcoming of the simple lookup based WL tagging procedure and is outside the scope of the topic models. The next subsections mention two important uses of the proposed models for tasks that provide deeper insights into the data including their measures of validation.

The nature of WL affect tags, though, needs some mention. The assignment of affect tags to review words based on maximum valence score do indeed make them orthogonal and one might choose not to use them at all in the modeling process. This orthogonality is also the principle reason behind using the regularization term for the parameters. However, including such orthogonality do have some good uses. The conditioning of topics on the WL tags allows us to discover terms that might be related to tagged words through shared topics. For example, words like “advertisements,” “ads,” “listing,” etc. that do not appear in [Bradley and Lang, 1999] could receive higher probability mass for some topic (e.g. magazines) conditioned on “MELANCHOLIC” affect while it could receive higher probability mass for another topic (e.g. software) conditioned on “CONTENDED” affect thereby introducing a relaxation over orthogonal WL tagging constraints. This phenomenon is better reflected in the METag²LDA model which does not enforce the correspondence constraint between the WL and DL perspectives in a strong manner.



	Beginning →	Begin- ning To Middle →	Middle →	Middle To End →	End	Tag Sug- gestions	Corres- pon- dence	Image Labels
Topic54 (Artillery: Wars)	Firing, United, artillery, Army, century, guns, support, forced, targets, modern	Firing, artillery, United, guns. Army, targets, forced, century, operated, design, weapons	Firing, United, artillery, guns, targets, century, Army, modern, weapons, tradi- tional	United, Firing, Army, weapons, guns, field, power, team, History, rockets, artillery	targets, attacks, aircraft, design, com- batants, radar, modern, small, electronic	artillery, guns, howitzer, French, cen- tury, Can- non,field, trajec- tories, PzH, self- propelled	(treating, soldiers) (con- struction, German) (con- struction, forge- welded)	
$\beta_{54}^{learned}$: Firing artillery guns. targets. United fuzes Army projectile mortars ammunition weapons shells. battery cannon modern								
Topic76 (Tofu: Food)	tofu, Chinese, Japanese, waters, China, Japan, century, origins, similar	tofu, Chinese, Japanese, Soy, Western, products, Asian, meat, im- portant, milk, tradi- tional, flavor	tofu, Chinese, Japanese, food, tra- ditional, Western, cultures, meat, Asian, Korean, Dishes, Japan, soy	tofu, Press, Asian, fries, food, ancient, Western, play, Japanese, Ice, oil, cuisine, fresh, America	Press, popular, cultures, China, deep, study, research, tradi- tional, Uni- versity, America, ISBN	tofu, sliced, China, water, soy, fresh, Provinces, Kong, milked, dishes, press, Hong, solid, soft	(beans, dried) (beans, tofu) (sus- pended, solid) (palm, island) (feeling, sweet)	
$\beta_{76}^{learned}$: tofu soy production Chinese milk firm texture flavor coagulated sauces soft Japanese “dufu” fries Protein cooking fresh beans								
Topic175	$\beta_{175}^{learned}$: galaxy Star spiraled milky matter cluster Hubble gas Universe structure Formation elliptical active galactic nebula dwarf							
ρ_{175}^{inf} : Galaxy, spiral, stars, Hubble, classification, Andromeda, rings, core, Great, compared								
$\lambda_{m,n}^{(galaxy)}$: (Planet, Hubble) (Planet, object) (Planet, galaxy) (Herschel, Hubble) (ring, galaxy) (Heat, galaxy) (discoveries, Hubble)								
Topic196	$\beta_{196}^{learned}$: fog air Shadow Ice condensation light vapor Humidity layer temperature freeze particle cool waters moisture evaporation salt							
ρ_{196}^{inf} : fog, Francisco, San, visible, high, temperature, streets, photo, Bai, lake, California, bridge, air								
$\lambda_{m,n}^{(fog)}$: (dimensions, high) (beam, visible) (parallel, bridge) (droplets, fog) (combustion, temperature) (invisible, visible) (absorbed, air)								

Table 4.7: Topics and correspondences from the Corr-METag²LDA for the Wikipedia data for $K = 200$

4.4.2 Automatically Evaluating Suggested Tags From Image Captions

For each of the Wikipedia test documents, top five predicted tags (coming from image captions and article names) were chosen. Following [Pedersen et al., 2004], the method described in section 4.1.3 was chosen as a quantitative measure of tag suggestion success. Figure 4.6a shows the relative values of the proposed Tag²LDA models for macro averages of maximum of best path distance scores for all test documents.

Some concept pair evidence chains from domain ontology

soy ↔ legume ↔ herb ↔ vascular_plant ↔ plant ↔ organism ↔ person ↔ inhabitant ↔ Asian ↔ Vietnamese
spiral ↔ curve ↔ line ↔ shape ↔ attribute ↔ abstraction ↔ group ↔ collection ↔ galaxy
weapon ↔ persuasion ↔ communication ↔ act ↔ activity ↔ occupation
french ↔ sculptor ↔ artist ↔ creator ↔ person ↔ modern
bottle-nosed_whale ↔ beaked_whale ↔ toothed_whale ↔ whale ↔ cetacean

Table 4.8: Sample evidence Chains for DL Tag suggestions from image captions to ground truth category labels from the Tag² topic models

Figure 4.6a suggests that people ignore the specific contents of the documents while assigning a category label. The METag²LDA model, in spite of higher perplexity, performs a little better here because of the lack of specificity of suggested DL tags to the document contents. This shows that humans assign DL tags that belong to higher levels of abstraction. Nevertheless, the best DL tags suggested by both METag²LDA and Corr-METag²LDA are only within 1 to 2 hops away from the ground truth tags based on a chosen WordNet ontology. Thus image captions in Wikipedia articles provide powerful clues for suggesting document tags. Table 4.8 shows “*explanations*” of the suggested DL tags to the ground truth category labels which is a desirable output of this type of evaluation. One could also use cross-document evidence trails [Srihari et al., 2007] to measure semantic relatedness. It is to be noted here that the ontology chosen must be specific to the nature of the task. For example, in medical domain, WordNet is a poor choice for providing explanations to DL tag suggestions.

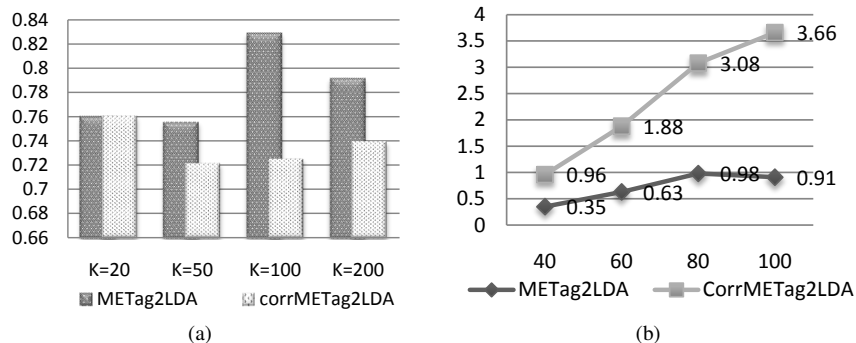


Figure 4.6: (4.6a) The best ontological inverse path length measure between suggested DL tags from image captions and ground truth Wikipedia categories for the test set in Fig. 4.4b and (4.6b) PERSON Named Entity-pair coverage ratio to baseline from DUC 2005 Newswire data

	Unsatisfactory	Melancholic	Bored	Hopeful	Contended	Pleased & Happy	Untagged
Topic35 (magazines)	terrible failed dead abused unhappy hurt like business diamond Advertised articles. home	Problem bad disappointed annoying garbage trouble waste poor bored useless horrible regret	price end dust trouble flawed. hinders bored cutting old broken slips leaking scuff. tire moronic	business issue short REPORT feet concerned distract lazy RE-MOVE fall tire dark risk shocked	ads. months want parts hand information book numbers materials, eye continuously. title activation	quality know day right save WORTH photos daughter. quick, interested wife pleased bargain	PAGE articles RETURN received Advertised find SMALL Need better little different
	$\beta_{35}^{learned}$: business diamond Advertised articles. home different people SMALL received little "base" listing RETURN find information ads						
	ρ_{35}^{inf} : ads little home_business_magazine Substance set car SCUM!! AHHHHHH!! Climates!						
	$\lambda_{m,n}^{toomanyads,little substance}$: (home, business_magazine) (business, little) (business, ads) (aimed, ads) ("base", Substance)						
Topic49 (food)	terrible, failed, hurt, stereo, sound, don't, CMT-NE3, CD, Philips, Sony, like, eat, money, unit, smelling	waste, poor, bad, crashing, disappointed, trash (...), didn't, WORTH, ship, regularly, noodles box. pasta	price, old, bother, denies, scuff, cells, damaged, stinks (...), pasta , leaving, huge, numbers inferior taste	expense, break, REMOVE, cold, wires, policy, submitted, loud..., short, passive, odd, bland , fall, onion	stereo, taste, quality, right, save, nice, know, feelings, Steaks, pasta, cooked, thin, ice, wheat, tray, grounds	like, good, money, eat, give, gift, understand, careful, food, Song, dollars, Glad, stars:, pretty, wish, Clean, impressed	sound, CD, better, SMALL, files, replacement, Sony, need, received, RETURN, box, didn't, different
	β_{49}^{inf} : stereo sound CD SMALL Philips Sony like eat money (...) ship regularly noodles box. pasta huge numbers taste expected						
	ρ_{49}^{inf} : quality, Pasta, reception, house, tofu_shirataki_8.00_oz, Cash, noodle_shaped_tofu, Don't, buy						
	$\lambda_{m,n}^{(not.quite.a.pasta)}$: (excessively, Pasta) (rinsed, Pasta) (Tofu, Pasta) (want, Pasta) (Glad, Pasta) (local, Pasta)						
Topic8 (software upgrade)	terrible MS failure Need flash users PC nightmare. REMOTE like poor buy! Quick-books frustrating uninstalled	Problem poor lose difficult wrong waste alarm horrible. disappointed frustrating bad virus	price stop discouraged tire avoid nasty confusion foolish Beware flawed. scuff. stinks damaged	expense. business office content. ruins tons REMOVE issue fall break. smudge application hanging	SYSTEM Vista, Custom office hard windows. totally button weeks tool locate paper finger zipper. want	quality able save answered adult Christmas Learn hope pretty wish home careful impressed computer	buy! version. Need don't better UP-GRADED programs MS PC users installed Pay costs uninstalled
	$\beta_8^{learned}$: MS Need flash users PC REMOTE like poor version. buy! professionally STUCK, Quickbooks them. uninstalled costs Pay						
	ρ_8^{inf} : MS, Terrible, Don't, money, Upgrade, waste, quickbooks_pro_2005, sending						
	$\lambda_{m,n}^{upgrade.from.2004}$: (tried, Upgrade) (Problem, Upgrade) (didn't, Upgrade) (version., Upgrade) (flash, Upgrade) (Quickbooks, Upgrade)						

Table 4.9: Three sample topics from the Corr-METag²LDA for the Amazon Product Review (AR) data for $K = 200$. Topic 49 highlights the problem with correspondence when there are more than a few competing topics for explaining the DL metadata

4.4.3 Automatically Evaluating Named Entity Relationship Discovery

For the DUC 2005 data, the second column in the first, third and fifth rows of table 4.10 show selected PERSON entity pairs which have been discovered to be related through some latent topics. The WL and DL tags for this purpose are named entity classes and abstract GSRts.

Nobel Prize Winners in Science & Economics	(John_Harsanyi, John_Nash) (Von_Neumann, John_Harsanyi) (Von_Neumann, John_Nash)
Last week the Nobel Prize for Economics was awarded to three 'game theorists': <u>John Harsanyi</u> , <u>John Nash</u> and <u>Rheinhard Selten</u> . What is game theory? Game theory is still a relatively young field. <u>Von Neumann</u> and Oskar Morganstern introduced many of the central ideas in a book published in 1944.	
Women in Parliaments	(Mrs_Margaret_Beckett, Ms_Ann_Taylor) (Mrs_Margaret_Beckett, Ms_Clare_Short) (Mrs_Margaret_Beckett, Ms_Harriet_Harman) (Mrs_Margaret_Beckett, Ms_Hilary_Armstrong) (Mrs_Margaret_Beckett, Ms_Jo_Richardson)
There are at present just four women occupants - Mrs Margaret Beckett, Ms Ann Clwyd, Ms Ann Taylor and Ms Jo Richardson - of the 18 shadow cabinet seats elected each year. The plan now being discussed by the group is to create a 'recommended' list of women candidates. Women would be asked to ensure that they included more than three votes for group members. Beneficiaries might include Ms Harriet Harman, Ms Clare Short, Ms Marjorie Mowlam and Ms Hilary Armstrong.	
VW/GM Industrial Espionage	(Bill_Clinton, Mr_Lopez) (Dorothea_Holland, Bill_Clinton) (Bill_Clinton, Ms_Holland)
It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that former GM director, Mr Lopez, stole industrial secrets from the US group and took them with him when he joined VW last year. This investigation was launched by US President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of Ms Holland. Dorothea Holland, until four months ago was the only prosecuting lawyer on the German case.	
Topic34 — ORG: GM Opel EC General_Motors Harvard volkswagen Justice_Department World_Bank Volkswagen the_Times FBI	
Topic34 — LOC: Germany Los_Angeles California UK german Washington Europe Brazil London Slovakia european U.S. New_York	
Topic34: GM Mr_Lopez group yesterday company german week official Mr_Piech work production charge car investigation prosecutor	

Table 4.10: DUC 2005 dataset: Related PERSON named entity pairs and evidence from documents

The first column in the first, third and fifth rows shows queries that serve as gists of the three docsets. To validate the discoveries the following experiment has been devised: For each docset in the DUC 2005 data, all entity pairs that are co-occurring in a sentence are counted and was treated to be a baseline measure of coverage for entity pairs which are related. Then a set of best topics are inferred for the documents in docsets by the Tag²LDA class of models. For each topic set, 2450 (=50x50-50) PERSON entity pairs are created out of the highly probable entities appearing in the PERSON facet of the conditional topics. Note that for all entities A and B, two entity pairs (A,B) and (B,A) are created and that entities such as John_Nash, Nash and Dr._Nash are treated as three separate entities. Each docset has on average 2449 PERSON entity pairs and hence the number 2450. The graph in fig. 4.6b shows that the correspondence model is three times better than the robust baseline at the right number of fitted topics and using the abstract GSRt DL perspective. The second, fourth and sixth rows of table 4.10 show how topical context ties two entities together even though they do not occur in the same sentence. The last three rows show ORGANIZATION facet, LOCATION facet and *marginal* topic corresponding to

the best topic for docset “VW/GM Industrial Espionage”.

4.4.4 TagLDA Revisited

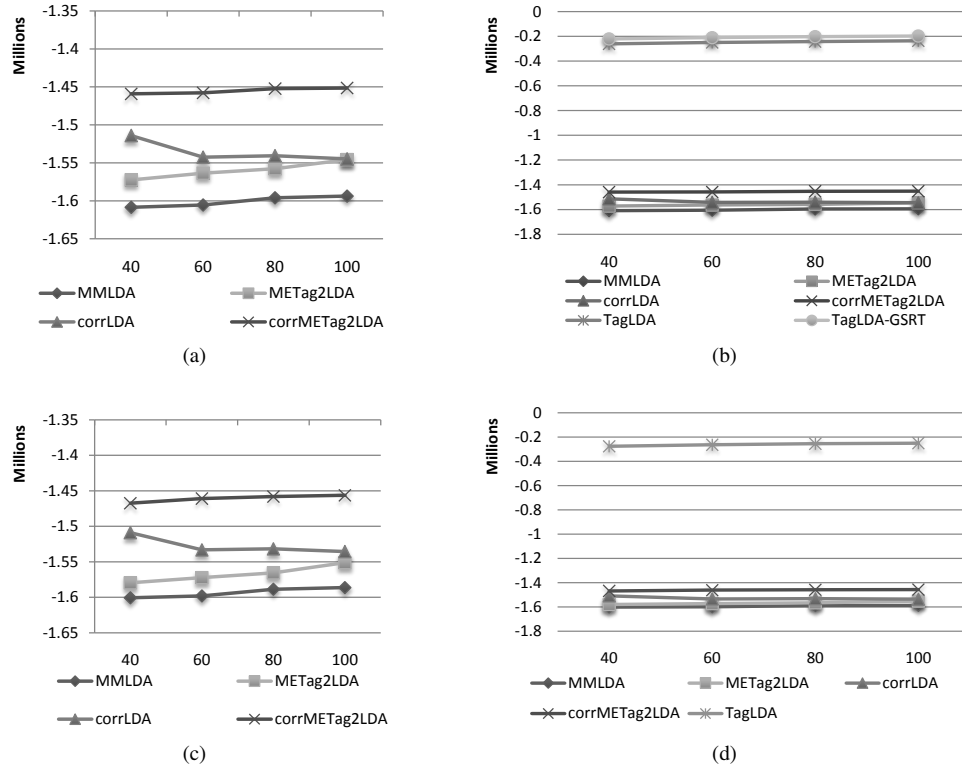


Figure 4.7: Cross-Validation results on DUC 2005 (DUC05) newswire data (**higher is better** in all figures); Fig. 4.7a: ELBO-Validation on DUC05 with GSRTNe tagging - not showing TagLDA; Fig. 4.7b: Better ELBO for TagLDA on DUC05 with Ne and GSRT tagging; Fig. 4.7c: ELBO-Validation on DUC05 with GSRTPos tagging - not showing TagLDA; Fig. 4.7d Better ELBO for TagLDA on DUC05 with Position tagging. X-axis represents the values of K

To measure predictive powers of the tag topic models, a 10-fold cross validation was performed on the DUC 2005 dataset. Fig. 4.7 shows the ELBOs (higher is better) on validation sets averaged over all folds. Fig. 4.7b shows the ELBO for TagLDA trained with only WL Named Entity (Ne) annotation classes and also the ELBO for TagLDAGSRt trained with WL prioritized GSRT tags. To annotate a word with a *single* prioritized GSRT the following is done: for a set of GSRTs $\{x \rightarrow y\}$ associated with a word $w_{d,m}$, prioritized GSRs $x_p \in \{x\}$ and $y_p \in \{y\}$ are chosen based on GSR prioritization rule given in section 4.1.1. If x_p and y_p are not the same, then these two GSRs are further prioritized. Finally the GSRT corresponding to the chosen GSR is taken to be the WL tag for the word. Fig. 4.7d shows the ELBO for TagLDA trained with only WL position (Pos) tags. It is clearly observed that when the documents are not sparsely represented over the set of all DL tags, TagLDA fits much better to the data for all choices of number of topics. However, excepting TagLDA, Corr-METag²LDA fits the data better than all other models and does so consistently across all folds in cross-validation. ELBOs for models with GSRTNe and GSRTPos tagging are comparable.

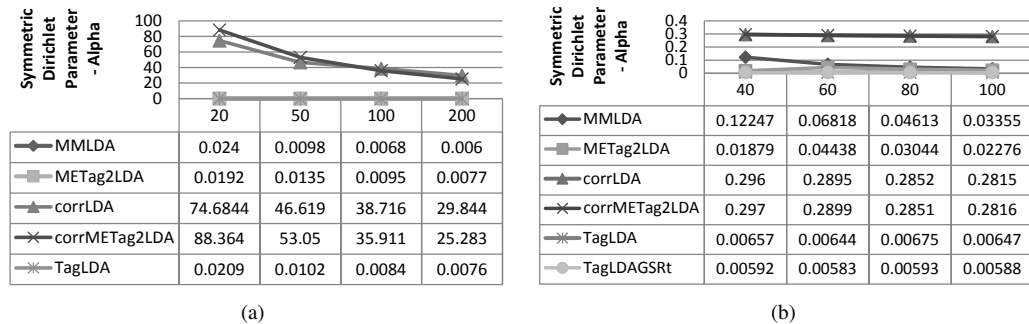


Figure 4.8: Fig. 4.8a: Optimum value of the prior parameter α for Wikipedia dataset; Fig. 4.8b: Optimum value of the prior parameter α for DUC 2005 dataset with GSRTNe/prioritized GSRT tagging (TagLDAGSRt). X-axis represents the values of K

The optimal values of α are higher in correspondence models than models where words and DL words/tags are independently generated as shown in Figs. 4.8a and 4.8b. Higher α values in correspondence models for the Wikipedia dataset is intuitive. The structure of Wikipedia documents is such that a well edited Wikipedia article often shows an orthogonal sub-topical structure among its sections. For the symmetric Dirichlet distribution, a high value of the prior α means that each document is likely to contain a mixture of a much higher number of topics, and not any single topic specifically. A low value of α puts fewer such constraints on documents and means that it is more likely that a document contain mixture of just a few of the topics. Since we are not treating each section of the Wikipedia articles as a separate document, the correspondence of the captions of the embedded multimedia to the exact subtopic is difficult particularly if the number of global topics is small. Thus in this case, the higher values of α is indicative of the fact that each Wikipedia article indeed exhibits mixed membership over many “sub-topics.” We can possibly introduce another level of hierarchy in the topical structure of the document representation (much like the Pachinko Allocation Model (PAM) [Li and McCallum, 2006]) and then correspond to the top level topics only.

4.4.5 Evaluating Tag-Topic Models through Extractive Multidocument Summarization

To quantitatively evaluate textual summarization power of the models vis-a-vis perplexities for the second task, the DUC 2005 dataset has been used. This dataset consists of 1593 newswire articles spanning several events which are binned into 50 predefined document sets (docsets) corresponding to 50 queries. Each document set also had 4 or 9 gold standard summaries written by humans. For a particular docset, a document is chosen and then a sentence along with at most two preceding and two succeeding sentences are chosen to form a short contextual document centered on a sentence. After a particular topic model is trained on the entire DUC 2005 corpus prior to context creation, the likelihoods for each such context across all documents in a particular docset are computed during the summarization procedure. These likelihoods serve as the ELBOs for the models on held out test data (sentence contexts) with no unseen out-of-vocabulary words. Finally, sentences with a minimum of n words and at least one named entity are chosen from all documents within a docset and ranked in descending order of likelihoods till a specified summary length L is reached with L being the number of whitespace separated tokens in the

summary for the docset. Further the summary sentences are chosen to contain at least 20 words including punctuations and stopwords. These summaries are then automatically scored against the model human summaries using ROUGE evaluation tool. The higher the score for a model generated summary, the better is its multidocument summarization power

The summarization power of the models are revealed from the graphs in fig. 4.9. The figures show that scores for 250 word summaries generated by TagLDA using WL tags alone is much worse compared to the summaries from the other models.

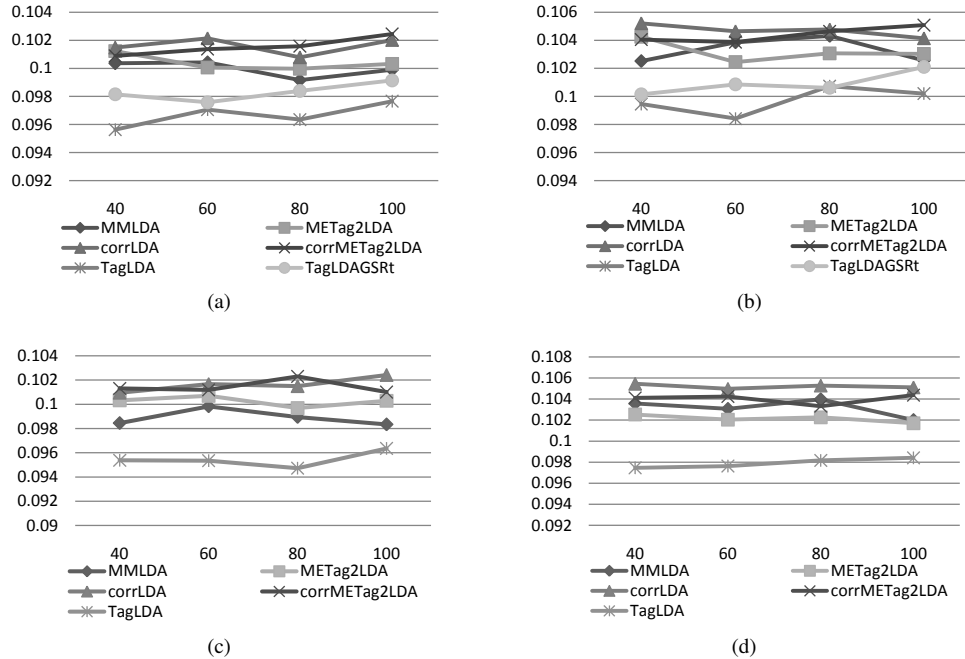


Figure 4.9: ROUGE SU4 scores (**higher is better**) of 250 word summaries for the models in fig. 4.1 - the scores are averaged over all docsets in the DUC 2005 dataset (GSRTNe/Ne/prioritized GSRT tagging); Fig. 4.9a: ROUGE SU-4 scores when each sentence in the summary is atleast 20 words (GSRTNe/Ne/prioritized GSRT tagging); Fig. 4.9b: ROUGE SU-4 scores when each sentence in the summary is atleast 30 words (GSRTNe/Ne/prioritized GSRT tagging); Fig. 4.9c: ROUGE SU-4 scores when each sentence in the summary is atleast 20 words (GSRTPos/Position tagging); Fig. 4.9d: ROUGE SU-4 scores when each sentence in the summary is atleast 30 words (GSRTPos/Position tagging). X-axis represents the values of K

ROUGE [Lin and Hovy, 2003] has been run with the official DUC 2005 command line arguments as “-e data -n 4 -w 1.2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -d”. The reason for choosing ROUGE is that it is a fully automatic scoring function using lexical matches as compared to PYRAMID [Nenkova and Passonneau, 2004] which is manual. Also ROUGE shows good correlation to scores obtained through PYRAMID evaluation. The widely adopted Skip Unigram of skip length 4 (SU-4) matching criterion of ROUGE is used to measure the quality of a summary.

It is observed that the summaries consisting of sets of most likely sentences centered around a context which are fit to the learned TagLDA model with different WL tagging (Ne, Pos and prioritized GSRT) worst reflect the information need conveyed in the document sets. This happens even if TagLDA beats all the proposed models by a wide margin in perplexity. The reason for this is intuitively clear.

When additional non-sparse features as document level GSRts are skipped, TagLDA models topics only on co-occurrence conditioned on WL tags. The more there are conditional events the higher the ELBO due to higher degrees of freedom—a fact supported by Fig. 4.7b for TagLDAGSRt for which the summarization power also increases as shown in Fig. 4.9. However, the addition of the DL assumption from a linguistic perspective that words are generated not only through co-occurrence but also from the need to continue or disrupt the propagation of attentional foci within contextual utterances [Grosz et al., 1995] improves the summarization power of the correspondence topic models. There has also been seen a general lack of coherence in the docset summaries generated using TagLDA. The lack of coherence is more pronounced when the summary consists of sentences of at least 30 words. This is understandable since very long sentences are usually self contained with more *null* GSRs in the GSRts. The frequency counts of the GSRts across the corpus also shows more $\{x \rightarrow y\}$ GSRts where either of x or y is *null* (–) because co-reference resolution has not been performed. The summarization power of the models slightly increases when summaries contain sentences of at least 30 words mostly because of slightly more lexical matches to the words in the human summaries.

Docset	Docset Query	Topic
d695c	What sentences are being imposed for financial crimes such as fraud and embezzlement?	(TopicID: 81) Sentencing crimes Landreth paying prison murder charge Harper killing tax government convicted criminal police judges attorney
<p>Query-independent Summary: A specialist firm on the New York Stock Exchange was fined \$480,350 for securities fraud by a judge who said he imposed the sentence as a deterrent. LaRouche, who has run for President four times, is known for his extreme views, including support for a quarantine of AIDS victims and allegations that Britain’s Queen Elizabeth is involved in drug trafficking. Federal prosecutors dropped conspiracy and fraud charges against Lyndon LaRouche Jr., moments before a Virginia judge sentenced the political extremist to 15 years in prison for related offenses. Former Norwalk City Administrator William Kraus was sentenced in federal court in San Diego to 5 years probation and fined \$1000 for his part in a land-fraud scheme that bilked investors out of more than \$3 million.</p>		

Table 4.11: DUC 2005 docset, latent topic and generated summary. Sentences are at least 20 words long

Table 4.11 shows a sample DUC 2005 docset with its information need, the best topic that is reflected in its documents for the Corr-METag²LDA model at $K = 100$ with GSRTNe DL+WL tagging and the central contextual sentences from the documents arranged in descending order of likelihoods to form a human readable summary. Summaries with sentences less than 20 words scored very short sentences higher due to smaller sum of log probabilities. The official maximum and baseline ROUGE SU-4 scores for the DUC 2005 dataset from NIST were 0.1316 and 0.0871 respectively for *query-dependent* summarization.

4.5 Summary

This chapter explores correspondence and mixture topic modeling of documents tagged from two different perspectives. There has been ongoing work in topic modeling of documents with tags (tag-topic models) where words and tags typically reflect a single perspective, namely document content. However, words in documents can also be tagged from different perspectives, for example, syntactic perspective as in part-of-speech tagging or an opinion perspective as in sentiment tagging. The models proposed in

this chapter are novel in: (i) the consideration of two different tag perspectives - a document level tag perspective that is relevant to the document as a whole and a word level tag perspective pertaining to each word in the document; (ii) the attribution of latent topics with word level tags and labeling latent topics with images in case of multimedia documents; and (iii) discovering the possible correspondence of the words to document level tags. The proposed correspondence tag-topic model shows better predictive power i.e. higher likelihood on held-out test data than all existing tag topic models and even a supervised topic model. To evaluate the models in practical scenarios, quantitative measures between the outputs of the proposed models and the ground truth domain knowledge have been explored. Manually assigned (gold standard) document category labels in Wikipedia pages are used to validate model-generated tag suggestions using a measure of pairwise concept similarity within an ontological hierarchy like WordNet [Fellbaum, 1998]. Using a news corpus, automatic relationship discovery between person names was performed and compared to a robust baseline.

The proposed Multinomial-Exponential Tag²LDA models capture semantics of documents with domain knowledge coming from two different and often orthogonal perspectives. The correspondence models also show impressive predictive power for inferring topics. Further, usefulness of the models have been explored with applications that provide deep insights into the data. Overall, it is possible to add domain knowledge from different perspectives, into topic models without sacrificing predictive power.

In the next chapter we explore the applicability of these models in the context of “guided” multi-document summarization.

4.6 Appendix

As is generally the case for hierarchical models belonging to the LDA family, the coupling between the hidden variables and parameters require an exponential number of state space configurations to be searched to find the best posterior distributions over the hidden variables. Since this problem is intractable, we find posterior distributions over the hidden variables by imposing tractable distributions with free variational parameters and then optimizing the expectation over the complete data log likelihood w.r.t these imposed variational distributions.

For a concrete discussion on one such scenario, we derive the expressions involving the optimization for the Corr-METag²LDA model. We begin with:

$$(\gamma^*, \phi^*, \lambda^*) = \arg \min_{(\gamma, \phi, \lambda)} KL(q(\theta, \mathbf{Z}, \mathbf{Y} | \gamma, \phi, \lambda) || p(\theta, \mathbf{Z}, \mathbf{Y} | \mathbf{W}_M, \mathbf{W}_N, \mathbf{T}_M, \alpha, \rho, \beta, \pi)) \quad (4.25)$$

The Expected Lower BOUND on the log likelihood of the data for a document d is given by:

$$\begin{aligned} \mathcal{L}(\gamma, \phi, \lambda) = & E_q[\ln p(\theta | \alpha)] + E_q[\ln p(\mathbf{Z} | \theta)] + E_q[\ln p(\mathbf{W} | \mathbf{Z}, \rho)] + E_q[\ln p(\mathbf{Y} | N)] + E_q[\ln p(\mathbf{W} | \mathbf{Y}, \mathbf{T}, \beta, \pi)] \\ & - E_q[\ln q(\theta | \gamma)] - E_q[\ln q(\mathbf{Z} | \phi)] - E_q[\ln q(\mathbf{Y} | \lambda)] \end{aligned} \quad (4.26)$$

Each of the terms in the equation (4.26) expands out to:

$$\ln \Gamma\left(\sum_{j=1}^K \alpha_j\right) - \sum_{k=1}^K \ln \Gamma(\alpha_k) + \sum_{k=1}^K (\alpha_k - 1) (\Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right)) \quad (4.27)$$

$$+ \sum_{n=1}^{N_d} \sum_{k=1}^K (\Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) \phi_{d,n,k} \quad (4.28)$$

$$+ \sum_{j=1}^{corrV} \sum_{n=1}^{N_d} \sum_{k=1}^K \phi_{d,n,k} \ln \rho_{k,w_n} \delta(w_n, j) \quad (4.29)$$

$$+ \sum_{m=1}^{M_d} \sum_{n=1}^{N_d} \lambda_{d,m,n} \frac{1}{N_d} \quad (4.30)$$

$$+ E_q[\ln p(\mathbf{w}_M | \mathbf{y}, \mathbf{t}, \boldsymbol{\beta}, \boldsymbol{\pi})] \quad (4.31)$$

$$- \ln \Gamma(\sum_{j=1}^K \gamma_{d,j}) + \sum_{k=1}^K \ln \Gamma(\gamma_{d,k}) - \sum_{k=1}^K (\gamma_{d,k} - 1) (\Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) \quad (4.32)$$

$$- \sum_{n=1}^{N_d} \sum_{k=1}^K \phi_{d,n,k} \ln \phi_{d,n,k} \quad (4.33)$$

$$- \sum_{m=1}^{M_d} \sum_{n=1}^{N_d} \lambda_{d,m,n} \ln \lambda_{d,m,n} \quad (4.34)$$

where the last three terms form the entropy of the tractable q distribution.

To find a further lower bound on $E_q[\ln p(\mathbf{w}_M | \mathbf{y}, \mathbf{t}, \boldsymbol{\beta}, \boldsymbol{\pi})]$, we use the inequality 4.4 given as $\ln(x) \leq \ln(\zeta) + \zeta^{-1}x - 1$, $\forall \zeta > 0$ to obtain the following lower bound:

$$\begin{aligned} E_q[\ln p(\mathbf{w}_M | \mathbf{y}, \mathbf{t}, \boldsymbol{\beta}, \boldsymbol{\pi})] &= E_q \left[\sum_{m=1}^{M_d} \ln \frac{\exp(\beta_{z_{y_d,m}, w_{d,m}}^\ell + \pi_{t_{d,m}, w_{d,m}}^\ell)}{\sum_{v=1}^V \exp(\beta_{z_{y_d,m}, v}^\ell + \pi_{t_{d,m}, v}^\ell)} \right] \text{ where } \beta^\ell / \pi^\ell = \ln \beta / \ln \pi \\ &= \sum_{m=1}^{M_d} E_q \left[\left(\beta_{z_{y_d,m}, w_{d,m}}^\ell + \pi_{t_{d,m}, w_{d,m}}^\ell \right) \right] - \sum_{m=1}^{M_d} E_q \left[\sum_{v=1}^V \exp(\beta_{z_{y_d,m}, v}^\ell + \pi_{t_{d,m}, v}^\ell) \right] \\ &\geq \sum_{m=1}^{M_d} \sum_{k=1}^K \left(\sum_{n=1}^{N_d} \lambda_{m,n} \phi_{n,k} \right) \left(\beta_{z_{y_d,m}, w_{d,m}}^\ell + \pi_{t_{d,m}, w_{d,m}}^\ell \right) \\ &\quad - \sum_{m=1}^{M_d} \left[\zeta_{d,m}^{-1} \sum_{v=1}^V \sum_{k=1}^K \left(\sum_{n=1}^{N_d} \lambda_{m,n} \phi_{n,k} \right) \exp(\beta_{z_{y_d,m}, v}^\ell + \pi_{t_{d,m}, v}^\ell) + \ln \zeta_{d,m} - 1 \right] \end{aligned} \quad (4.35)$$

Using Equ. 4.35, we obtain a second lower bound to the original ELBO (see Equ. 4.11) involving another free variable $\zeta_{d,m}$.

4.6.1 Inference on Variational Parameters

Here we estimate the free variational parameters for the variational model following the constraints on ϕ and λ .

For γ :

$$\mathcal{L}[\gamma] = - \ln \Gamma(\sum_{j=1}^K \gamma_{d,j}) + \sum_{k=1}^K \ln \Gamma(\gamma_{d,k}) + \sum_{k=1}^K (\alpha_k + \sum_{t=1}^T \phi_{d,t,k} - \gamma_{d,k}) (\Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) \quad (4.36)$$

$$\frac{\partial \mathcal{L}_{[\gamma]}}{\partial \gamma_{d,k}} = (\alpha_k + \sum_{t=1}^T \phi_{d,t,k} - \gamma_{d,k})(\Psi'(\gamma_{d,k}) - \Psi'(\sum_{j=1}^K \gamma_{d,j})) - (\Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) + (\Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) \quad (4.37)$$

Setting the above derivative to 0, we get,

$$\gamma_{d,k} = \alpha_k + \sum_{k=1}^K \phi_{d,n,k} \quad (4.38)$$

where $\mathcal{L}_{[\gamma]}$ denotes the expression with only those terms that depend on γ in the expression for $\mathcal{L}(\cdot)$

For λ :

$$\begin{aligned} \mathcal{L}_{[\lambda_{d,m,n}]} &= -\lambda_{d,m,n} \ln \lambda_{d,m,n} + \lambda_{d,m,n} \frac{1}{N_d} + \sum_{k=1}^K (\lambda_{d,m,n} \phi_{d,n,k}) \left(\beta_{z_{y_{d,m}}, w_{d,m}}^\ell + \pi_{t_{d,m}, w_{d,m}}^\ell \right) \\ &\quad - \zeta_{d,m}^{-1} \sum_{v=1}^V \sum_{k=1}^K (\lambda_{d,m,n} \phi_{d,n,k}) \exp \left(\beta_{z_{y_{d,m}}, w_{d,m}}^\ell + \pi_{t_{d,m}, w_{d,m}}^\ell \right) + \mu_m \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} - 1 \right) \\ \therefore \frac{\partial \mathcal{L}_{[\lambda_{d,m,n}]} }{\partial \lambda_{d,m,n}} &= 0 \implies -\ln \lambda_{d,m,n} - 1 + 1/N_d + \sum_{k=1}^K \phi_{d,n,k} \left(\beta_{z_{y_{d,m}}, w_{d,m}}^\ell + \pi_{t_{d,m}, w_{d,m}}^\ell \right) \\ &\quad - \zeta_{d,m}^{-1} \left(\sum_{v=1}^V \sum_{k=1}^K \phi_{d,n,k} \exp \left(\beta_{z_{y_{d,m}}, v}^\ell + \pi_{t_{d,m}, v}^\ell \right) \right) + \mu_m = 0 \\ \therefore \lambda_{d,m,n} &\propto \exp \left\{ \sum_{k=1}^K \phi_{d,n,k} \left(\beta_{z_{y_{d,m}}, w_{d,m}}^\ell + \pi_{t_{d,m}, w_{d,m}}^\ell \right) \right. \\ &\quad \left. - \zeta_{d,m}^{-1} \left(\sum_{v=1}^V \sum_{k=1}^K \phi_{d,n,k} \exp \left(\beta_{z_{y_{d,m}}, v}^\ell + \pi_{t_{d,m}, v}^\ell \right) \right) \right\} \quad (4.39) \end{aligned}$$

where μ_m are the m Lagrange multipliers one for each of the free N_d -dimensional multinomial parameters $\lambda_{d,m}$. The constant of proportionality is the summation of the expression in the right hand side of Equ. 4.39 over all $n \in \{1, \dots, N_d\}$ since $\sum_{n=1}^{N_d} \lambda_{d,m,n} = 1$.

For ϕ :

$$\begin{aligned} \mathcal{L}_{[\phi_{n,k}]} &= (\Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) \phi_{d,n,k} + \phi_{d,n,k} \ln \rho_{d,k, w_{d,n}} + \sum_{m=1}^{M_d} \left\{ (\lambda_{d,m,n} \phi_{d,n,k}) \left(\beta_{z_{y_{d,m}}, w_{d,m}}^\ell + \pi_{t_{d,m}, w_{d,m}}^\ell \right) \right. \\ &\quad \left. - \zeta_{d,m}^{-1} \left(\sum_{v=1}^V \lambda_{d,m,n} \phi_{d,n,k} \exp \left(\beta_{z_{y_{d,m}}, v}^\ell + \pi_{t_{d,m}, v}^\ell \right) \right) \right\} - \phi_{d,n,k} \ln \phi_{d,n,k} + \mu_n \phi_{d,n,k} \end{aligned}$$

where the last term is due to the fact that the n Lagrange multipliers are represented by $\mu_n (\sum_{k=1}^K \phi_{d,n,k} - 1)$.

$$\begin{aligned} \therefore \frac{\partial \mathcal{L}_{[\phi_{n,k}]} }{\partial \phi_{d,n,k}} &= 0 \implies -\ln \phi_{d,n,k} - 1 + (\Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) + \ln \rho_{d,k, w_{d,n}} \\ &\quad + \sum_{m=1}^{M_d} \left\{ (\lambda_{d,m,n}) \left(\beta_{z_{y_{d,m}}, w_{d,m}}^\ell + \pi_{t_{d,m}, w_{d,m}}^\ell \right) \right. \\ &\quad \left. - \zeta_{d,m}^{-1} \left(\sum_{v=1}^V \lambda_{d,m,n} \exp \left(\beta_{z_{y_{d,m}}, v}^\ell + \pi_{t_{d,m}, v}^\ell \right) \right) \right\} + \mu_n = 0 \end{aligned}$$

Using $\sum_{k=1}^K \phi_{d,n,k} = 1$, we have

$$\phi_{d,n,k} \propto \exp \left\{ \Psi(\gamma_{d,k}) - \Psi\left(\sum_{j=1}^K \gamma_{d,j}\right) + \ln \rho_{d,k,w_{d,n}} + \sum_{m=1}^{M_d} \left\{ (\lambda_{d,m,n}) \left(\beta_{z_{y_{d,m}},w_{d,m}}^\ell + \pi_{t_{d,m},w_{d,m}}^\ell \right) - \zeta_{d,m}^{-1} \left(\sum_{v=1}^V \lambda_{d,m,n} \exp \left(\beta_{z_{y_{d,m}},v}^\ell + \pi_{t_{d,m},v}^\ell \right) \right) \right\} \right\} \quad (4.40)$$

For ζ :

$$\begin{aligned} \mathcal{L}[\zeta_m] &= \zeta_{d,m}^{-1} \sum_{v=1}^V \sum_{k=1}^K \left[\left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \exp \left(\beta_{z_{y_{d,m}},v}^\ell + \pi_{t_{d,m},v}^\ell \right) \right) \right] + \ln \zeta_{d,m} \\ \therefore \frac{\partial \mathcal{L}[\zeta_m]}{\partial \zeta_{d,m}} &= 0 \implies -\frac{1}{\zeta_{d,m}^2} \sum_{v=1}^V \sum_{k=1}^K \left[\left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \exp \left(\beta_{z_{y_{d,m}},v}^\ell + \pi_{t_{d,m},v}^\ell \right) \right) \right] + \frac{1}{\zeta_{d,m}} = 0 \\ \implies \zeta_{d,m} &= \sum_{v=1}^V \sum_{k=1}^K \left[\left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \exp \left(\beta_{z_{y_{d,m}},v}^\ell + \pi_{t_{d,m},v}^\ell \right) \right) \right] \end{aligned} \quad (4.41)$$

4.6.2 Model Parameter Estimation

In this section, we calculate the maximum likelihood settings of the parameters.

For ρ :

$$\mathcal{L}[\rho] = \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{k=1}^K \sum_{v=1}^{corrV} \phi_{d,n,k} \ln \rho_{k,w_{d,n}} \delta(w_{d,n}, v) + \sum_{k=1}^K \mu_k \left(\sum_{v=1}^{corrV} \rho_{k,v} - 1 \right) \quad (4.42)$$

where the μ_k 's are the K Lagrange multipliers in (4.42).

$$\begin{aligned} \therefore \frac{\partial \mathcal{L}}{\partial \rho_{k,v}} &= \sum_{d=1}^D \sum_{v=1}^{corrV} \sum_{n=1}^{N_d} \phi_{d,n,k} \frac{1}{\rho_{k,v}} \delta(w_{d,n}, v) + \mu_k \\ \frac{\partial \mathcal{L}}{\partial \rho_{k,v}} = 0 &\implies \rho_{k,v} = -\frac{\sum_{d=1}^D \sum_{n=1}^{N_d} \phi_{d,n,k} \delta(w_{d,n}, v)}{\mu_k} \implies \mu_k = -\sum_{d=1}^D \sum_{v=1}^{corrV} \sum_{n=1}^{N_d} \phi_{d,n,k} \delta(w_{d,n}, v) \\ \therefore \frac{\partial \mathcal{L}}{\partial \rho_{k,g}} = 0 &\implies \rho_{k,v} \propto \sum_{d=1}^D \sum_{v=1}^{corrV} \sum_{n=1}^{N_d} \phi_{d,n,k} \delta(w_{d,n}, v) \end{aligned} \quad (4.43)$$

For β :

$$\begin{aligned} \mathcal{L}[\beta] &= \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{k=1}^K \sum_{v=1}^V \sum_{t=1}^T \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \right) \left(\beta_{z_{y_{d,m}},v}^\ell + \pi_{t_{d,m},v}^\ell \right) \delta(w_{d,m}, v) \delta(t_{d,m}, t) \\ &\quad - \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{k=1}^K \sum_{v=1}^V \sum_{t=1}^T \left(\zeta_{d,m}^{-1} \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \right) \exp \left(\pi_{t_{d,m},v}^\ell \right) \right) \exp \left(\beta_{z_{y_{d,m}},v}^\ell \right) \delta(w_{d,m}, v) \delta(t_{d,m}, t) \\ \therefore \frac{\partial \mathcal{L}[\beta]}{\partial \beta_{k,v}^\ell} = 0 &\implies \beta_{k,v}^\ell = \ln(\text{term}_1^\beta) - \ln(\text{term}_2^\beta) \end{aligned} \quad (4.44)$$

where,

$$term_1^\beta = \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{t=1}^T \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \right) \delta(w_{d,m}, v) \delta(t_{d,m}, t) \quad (4.45)$$

$$term_2^\beta = \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{t=1}^T \left(\zeta_{d,m}^{-1} \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \right) \exp \left(\pi_{t_{d,m},v}^\ell \right) \right) \delta(w_{d,m}, v) \delta(t_{d,m}, t) \quad (4.46)$$

For π :

$$\begin{aligned} \mathcal{L}_{[\pi]} &= \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{k=1}^K \sum_{v=1}^V \sum_{t=1}^T \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \right) \left(\beta_{zy_{d,m},v}^\ell + \pi_{t_{d,m},v}^\ell \right) \delta(w_{d,m}, v) \delta(t_{d,m}, t) \\ &\quad - \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{k=1}^K \sum_{v=1}^V \sum_{t=1}^T \left(\zeta_{d,m}^{-1} \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \right) \exp \left(\beta_{zy_{d,m},v}^\ell \right) \right) \exp \left(\pi_{t_{d,m},v}^\ell \right) \delta(w_{d,m}, v) \delta(t_{d,m}, t) \\ \therefore \frac{\partial \mathcal{L}_{[\pi]}}{\partial \pi_{t,v}^\ell} &= 0 \implies \pi_{t,v}^\ell = \ln(term_1^\pi) - \ln(term_2^\pi) \end{aligned} \quad (4.47)$$

where,

$$term_1^\pi = \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{k=1}^K \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \right) \delta(w_{d,m}, v) \delta(t_{d,m}, t) \quad (4.48)$$

$$term_2^\pi = \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{k=1}^K \left(\zeta_{d,m}^{-1} \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \right) \exp \left(\beta_{w_{d,m},v}^\ell \right) \right) \delta(w_{d,m}, v) \delta(t_{d,m}, t) \quad (4.49)$$

Note that the maximum likelihood expressions for β and π involve unconstrained optimization. If these parameters are not regularized then any improper scaling of the weights often leads to violation of the convergence criteria for fixed point iterations in the E step (see Section 2.7.5). We therefore use a 0 mean and a fixed σ standard deviation Gaussian regularizer for each component of the parameters.

In particular, we use a ‘‘fixed prior’’ for $p(w_{d,m} | y_{d,m}, t_{w_{d,m}}, \beta, \pi)$ as:

$$\exp \left\{ -\frac{1}{2\sigma_\beta^2} \left(\sum_{v=1}^V (\exp(\beta_{k,v}^\ell))^2 \right) \right\} \times \exp \left\{ -\frac{1}{2\sigma_\pi^2} \left(\sum_{v=1}^V (\exp(\pi_{t,v}^\ell))^2 \right) \right\} \quad (4.50)$$

So, the derivative w.r.t β^ℓ for the previous expression for the ELBO for Corr-METag²LDA (see Equ. 4.11) becomes:

$$\frac{\partial \mathcal{L}_{[\beta_{k,v}]}^{(R)}}{\partial \beta_{k,v}^\ell} = term_1^\beta - term_2^\beta \times \exp \left\{ \beta_{k,v}^\ell \right\} - \frac{1}{2\sigma_\beta^2} \times 2 \left(\exp \left\{ \beta_{k,v}^\ell \right\} \right)^2 \quad (4.51)$$

Setting the above derivative to 0, we obtain:

$$\exp \left\{ \beta_{k,v}^\ell \right\}^2 + \sigma_\beta^2 term_2^\beta \times \exp \left\{ \beta_{k,v}^\ell \right\} - \sigma_\beta^2 term_1^\beta = 0 \quad (4.52)$$

The expression in Equ. 4.52 is a quadratic in $\exp\{\beta_{k,v}^\ell\}$ which can be solved as:

$$\exp\{\beta_{k,v}^\ell\} = \frac{-\sigma_\beta^2 \text{term}_2^\beta \pm \sqrt{\sigma_\beta^4 (\text{term}_2^\beta)^2 + 4\sigma_\beta^2 \text{term}_1^\beta}}{2} \quad (4.53)$$

Since exponential of a real number is always positive, we select the root with the positive second term since then we have:

$$\begin{aligned} \exp\{\beta_{k,v}^\ell\} \geq 0 &\implies \frac{\sigma_\beta^2}{4} \left(\sigma_\beta^2 (\text{term}_2^\beta)^2 + 4\text{term}_1^\beta \right) \geq \frac{\sigma_\beta^4}{4} (\text{term}_2^\beta)^2 \\ &\implies \sigma_\beta^2 \text{term}_1^\beta \geq 0 \\ &\implies \sigma_\beta^2 \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{t=1}^T \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,k} \right) \delta(w_{d,m}, v) \delta(t_{d_m}, t) \geq 0 \end{aligned} \quad (4.54)$$

The last expression in Equ. 4.54 is always true. Thus we have:

$$\beta_{k,v}^\ell = \ln \left(\frac{-\sigma_\beta^2 \text{term}_2^\beta + \sqrt{\sigma_\beta^4 (\text{term}_2^\beta)^2 + 4\sigma_\beta^2 \text{term}_1^\beta}}{2} \right) \quad (4.55)$$

and

$$\pi_{t,v}^\ell = \ln \left(\frac{-\sigma_\pi^2 \text{term}_2^\pi + \sqrt{\sigma_\pi^4 (\text{term}_2^\pi)^2 + 4\sigma_\pi^2 \text{term}_1^\pi}}{2} \right) \quad (4.56)$$

For α :

$$\begin{aligned} \mathcal{L}_{[\alpha]} &= \sum_{d=1}^D \left(\ln \Gamma \left(\sum_{k=1}^K \alpha_k \right) - \sum_{k=1}^K \ln \Gamma(\alpha_k) + \sum_{k=1}^K (\alpha_k - 1) (\Psi(\gamma_{d,k}) - \Psi \left(\sum_{j=1}^K \gamma_{d,j} \right)) \right) \\ \implies \frac{\partial \mathcal{L}_{[\alpha]}}{\partial \alpha_k} &= D \left(-\Psi(\alpha_k) + \Psi \left(\sum_{j=1}^K \alpha_j \right) + \sum_{d=1}^D (\Psi(\gamma_{d,k}) - \Psi \left(\sum_{j=1}^K \gamma_{d,j} \right)) \right) \end{aligned} \quad (4.57)$$

$$\frac{\partial \mathcal{L}_{[\alpha]}}{\partial \alpha_k \alpha_j} = \partial(k, j) D \left(\Psi'(\alpha_k) - \Psi' \left(\sum_{j=1}^K \alpha_j \right) \right) \quad (4.58)$$

The derivative w.r.t. α_k depends on α_j and thus we can resort to Newton's iterative method to find out the maximal α using the gradient and Hessian vector and matrix respectively as in [Blei et al., 2003]. Ψ' is the trigamma function.

Chapter 5

Using Bi-Perspective Topic Models and Rhetorical Structure Trees to generate Bullet Lists

“... like a swan, who can take the milk from a mixture of milk and water, leaving aside the water; like an ant, who can take the sugar from a mixture of sugar and sand, leaving aside the sand” - Ramakrishna

5.1 Introduction

Recent efforts within the Text Analysis Conference (TAC) community has led to the creation of the “Guided Summarization” task to encourage research and fair comparisons of peer systems. This is the task of generating a summary of a collection of documents as an answer to the information need of a user, which is commonly expressed as a very short query. In general, solutions for automatic text summarization are approached as a combination of several factors: the importance of sentences (which can be estimated from how often they are repeated across the collection, possibly as paraphrases), the redundancy between sentences (so as not to generate redundant summaries), and the readability of the produced summary. Because of its simplicity, most summarization systems currently used are extractive, i.e. they compose the output summary by combining sentences extracted from the original documents, which are sometimes modified through sentence rewriting or compression.

Experiments on human extractive summarization [Genest et al., 2009] show that even the best content-selection mechanism (e.g., a human summarizer) which is limited to pasting together sentences cannot achieve the same quality as fully manual summaries. The clusters of documents to summarize (which we will refer to as document sets or *docsets* in this paper) fall into predefined categories with highly predictable important elements. These elements aid summary evaluation by explicitly guiding the creation of human reference summaries to contain all or most of them. The categories are usually referred to as event categories and an example of such a category is “Accidents and natural disasters” with the guiding aspects i.e. categorical elements to be *{what happened, date, location, reasons for accident/disaster}*. In order to provide information for all the elements in these templates it is usually

DocsetID / Query [event category]	Important Nouns	Important Verbs
D1105A / Plane Crash Indonesia [Accidents and Natural Disasters]	Adam, Air, Boeing, Hartono, Sulawesi, accident, board, crash, emergency, official, pain, passenger, plane, rescue, search, survivor	carry, disappear, find, kill, miss, send
D1101A / Amish Shooting [Attacks (Criminal/ Terrorist)]	Miller, Roberts, attack, child, door, dream, family, girl, man, neighbor, number, police, school, schoolhouse, victim, wife	enter, kill, leave, molest, shoot, speak, storm, tie, turn, weave
D1102A / Internet Security [Health and Safety]	Internet, VeriSign, address, attack, business, company, computer, datum, domain, investment, security, server, system, technology, traffic, user, virus	convert, grow, manage, may, operate
D1106A / Tuna Fishing [Endangered Resources]	Japan, Kobe, Ocean, catch, conference, conservation, country, fishery, fishing, management, meeting, overfishing, plan, stock, tuna	adopt, expect, include, poach, track
D1103A / Madrid Trainbombings Trial [Investigations and Trials (Criminal/ Legal/ Other)]	Ahmed, God, Italy, Madrid, Moroccan, Spain, attack, bombing, charge, evidence, face, jail, murder, police, sentence, train, trial	accuse, allege, blow, expect, injure, kill

Table 5.1: Sample Docset IDs, their corresponding information needs and categories, important nouns and important verbs from the TAC 2011 Guided Summarization dataset. The nouns and verbs are obtained using an automatic part-of-speech tagger.

desirable to find the relevant content on a sub-sentential level through the use of information extraction and other meaning-oriented techniques.

Current state-of-the-art extractive query-focused summarization systems like those mentioned in [Conroy et al., 2010, Conroy et al., 2011, Varma et al., 2010, Varma et al., 2011] are all predominantly local methods which means that their systems extract key information that is only relevant to the docset being summarized. Although some of the features that these systems use are also computed “globally” using term statistics from other docsets or even other data sources. Regarding sentence scoring, a very important aspect of all query-focused summarization systems is to model the importance of words in the sentences conditioned on the user’s information need i.e. the query.

Many systems, including CLASSY [Conroy et al., 2010, Conroy et al., 2011], derive a lexicon that best represents the category of the docset and its aspects through the use of external sources like the Internet. However, it has been recently noted in [Conroy et al., 2010] that such lexicons, if not chosen properly may lower summarization performance due to topic drift. We show in this article, how simple models that are local to the docsets can be used to derive such lexicons automatically from the data at hand. Such automatically derived lexicon is very appropriate for categorizing the document sets into the corresponding event categories and also for summarizing the documents according to the Guided Summarization task definition. A few examples of such automatically derived lexicons from our system for some of the docsets in the TAC 2011 Guided Summarization dataset are shown in Table 5.1.

Some systems like the ones in [Schilder and Kondadadi, 2008, Varma et al., 2010, Varma et al., 2011] leverage summarization performance by training supervised classifiers based on features extracted for each sentence and its bigram overlap to available human summaries for some related datasets. Although this supervised learning approach is effective for boosting summarization performance by learn-

ing the weighting of the different correlates in a principled way, it needs large amounts of available training data with human summaries - something which may not be available when one wants to port the algorithm to other genres like scientific literature, forum messages, story books and novels etc.

Alternatively, unsupervised topic models like LDA [Blei et al., 2003] are very powerful data exploration techniques which can summarize data in the form of bag-of-words summaries where each bag holds semantically related items. Recent extensions of LDA-based models that use more structure in the representation of documents have also been proposed for generating more coherent and less redundant summaries, such as those in [Daumé and Marcu, 2006, Haghighi and Vanderwende, 2009, Celikyilmaz and Hakkani-Tür, 2011, Delort and Alfonseca, 2012]. These models use the collection and target document-specific distributions in order to distinguish between the general and specific topics in documents. For example, a general topic may look like a bag of closed-class words like those listed in a standard English stopword list and a document specific topic for a particular docset may look like a bag of words comprising of the nouns and verbs as shown in Table 5.1. In the context of summarization, this distinction is very similar to identifying signature terms [Lin and Hovy, 2000] at multiple granularities in a corpus driven manner and weighting sentences accordingly for inclusion into summaries. Since many of these signature terms happen to be Named Entities, it is often useful to use supervised methods to identify them and influence the topic modeling process instead. In this chapter, one of the main reasons to choose multi-modal tag-topic models like the TagLDA model in [Zhu et al., 2006] and the ones in [Das et al., 2011] is their ability to handle the word level annotation. The aspects of the categories concerning {who, when, date, location} naturally ask for highlighting the text with Named Entities and that the discovery of latent topics should also be conditioned upon these entities in proper context. Also, in general, TagLDA [Zhu et al., 2006] shows much lower perplexity than LDA by using the word level annotations and thus we have opted for TagLDA over LDA in our experiments.

Although most of the topic model based approaches in the multi-document summarization community focus on building topic models on the documents in the individual docsets, there has been no comparison so far on how good those models are for document summarization if there is no assumption of a docset structure in the corpus for training the models. For example in the TAC 2011 Guided Summarization corpus, there are 44 docsets covering the 5 event categories mentioned before. Thus if latent topics are built per docset [Celikyilmaz and Hakkani-Tür, 2011] or indexed by docsets [Haghighi and Vanderwende, 2009] then the true power of the models w.r.t. the diversity of latent spaces are never evaluated.

The multi-document summarization power of a topic model can be understood if there is a scheme to fit sentences in a docset to the model which has been learnt across all documents in all docsets. The top ranked summary sentences respecting some fixed word limit should still be as good as a local docset specific centroid based algorithm like the popular MEAD system [Radev et al., 2004]. To this end, we perform a quantitative evaluation of the tag-topic models [Das et al., 2011] trained on the full corpus without any docset partitioning, thus opening up the possibility of evaluating topic models using a measure other than the standard measure of perplexity (or log-likelihood) on held-out test data (Sect. 5.3.7).

In general we observe that the better a topic model is able to classify a document into its corresponding event, the better its chances to compete with a local centroid based algorithm (see Section 5.4.2). One subtle aspect of the Guided Summarization problem that cannot be incorporated into the topic modeling process through word level annotations is the issue with categorical aspects dealing with

questions like “what happened?”, “reasons for accident/disaster?” etc. The analysis of the rhetorical structure of texts through the use of Rhetorical Structure Theory [Mann and Thompson, 1988] (henceforth RST) has shown promise in this regard for text summarization [Marcu, 1999]. We believe that this direction should be further explored for potentially compacting extractive full sentence summaries into bullet lists whilst conforming to the unified information model enforced in the Guided Summarization principles.

The main contributions of this chapter are three-folds:

- i. We improve upon the models in Chapter 4 using asymmetric Dirichlet priors (Sect. 5.4.2). We show that without any knowledge of docset partitioning of a newswire corpus for use by the tag-topic models during training, the summaries formed by fitting sentences within each docset easily parallel those from a very robust centroid based summarization system trained only on docset features. This allows us to evaluate the tag-topic models using ROUGE as well as topically analyze a corpora.
- ii. We use *sentence likelihood* scores from globally trained tag-topic models together with those from very granular docset specific local models to vastly improve summarization performance.
- iii. Finally, we use contiguous spans of constituents automatically obtained from rhetorical parses to create bullet lists.

We next discuss the standard datasets on which our experiments in this chapter are evaluated.

5.1.1 Datasets

The datasets we choose for our experiments in this article are two recently released newswire corpora. These corpora are released as part of the Text Analysis Conference Guided Summarization task for year 2011. We have focused primarily only on the “base” Guided Summarization task, where systems are required to generate a 100-word summary of a set of 10 documents about a single topic. This set of 10 documents is what is called a document set or docset and each docset is unique in that each one addresses a particular information need. Also all summaries are asked to address specific aspects relevant to the summary topic without actually annotating those aspects within the summaries. For TAC 2011, there are 44 such topics or docsets. TAC 2011 dataset also included a development set from 2010 which had a similar docset organization with 46 folders. A few topic titles from the event categories of the TAC 2011 dataset are as follows: 1) Category: Accidents and Natural Disasters [Topic Title: Bangladesh flood]; 2) Category: Attacks [Topic Title: Glasgow airport attack]; 3) Category: Health and Safety [Topic Title: Pet Food Recall]; 4) Category: Endangered Resources [Topic Title: Endangered turtles]; 5) Category: Trials and Investigations [Topic Title: Bernard Madoff]

A complete list of the aspects of the categories that cover subordinate information needs can be found in the TAC 2010 website¹. The reason for the TAC datasets containing a smaller number of docsets is to make sure that the automatic evaluation measures correlated well to the manual ones [Dang, 2006b] and yet the evaluation efforts were as less as possible. In the TAC datasets, the docsets are also called “topics” but these are different from the latent topics obtained from statistical topic models.

Additionally we also experiment with running our system on the update task of the TAC 2011 Summarization task. The “Update” component of the Guided Summarization task is to write a 100-word updated summary of 10 subsequent newswire articles for the topic, under the assumption that a

¹<http://www.nist.gov/tac/2010/Summarization/>

user has already read the earlier articles. We show that without actually modeling an update process of summarization explicitly, we are able to achieve state-of-the-art Update Summarization performance by utilizing the same local models, RS-trees and topic modeling over both the base and update set of documents (see Section 5.5.4).

In the TAC 2010 and 2011 datasets, the topic titles of the docsets for both the base and update collections are the same. The docset IDs for the base collection, intended for Guided Summarization only, uses a “A” suffix to differentiate itself from the docset IDs in the update collection which are appended with a “B” suffix. In this article, TAC 2011A thus refers to the “Base” collection and TAC 2011B refers to the “Update” collection for the TAC Summarization dataset for the year 2011. Similar notation holds for the dataset for the year 2010. Each docset in the Update collections also has 10 documents.

5.1.2 Global Topic Models and Local Sentential Models

Our hypothesis is that there are a number of coarser to finer latent features in documents that can be very useful for the task of summarization. These include automatically discovered latent topic clusters, dependencies within sentential words, coherence structure in documents, rhetorical structure of sentences, etc. and we want a model that captures several of these things and combines them to better meet the needs of the Guided Summarization principles.

Figure 5.1 shows two sentences from a sample document concerning “sleep deprivation.” The text appeared in docset D1127E-A in the TAC 2011 base document collection. The light blue rectangular bubble on the right contains words stylized in varying font sizes depending on their frequencies in the text. As in [Das et al., 2011], this bag-of-concepts can be looked upon as a document level perspective that provides a gist of the document in terms of salient words appearing most frequently. The frequency of words usually have considerable impact on final summaries [Nenkova et al., 2006a, Vanderwende et al., 2007].

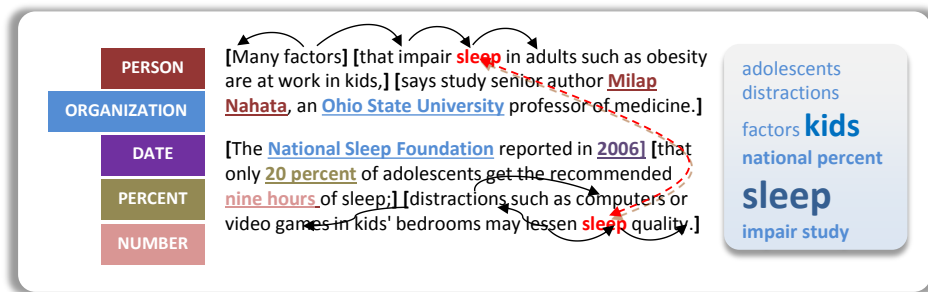


Figure 5.1: An article for the query “sleep deprivation” showing a document level and a word level perspective with some shallow and deep linguistic structures

The word “sleep” (connected by a dashed arrow) in **bold** font and colored **red** appears as a noun in the first and second sentences. The word “sleep” represented as the triplet (**sleep**, **noun**, **noun**) acts as an important center of attention that signifies an event rather than an entity. This triplet thus also helps strengthening the document level perspective focusing on “what this document is about?” In the triplet the first element is the word that appears in consecutive sentences, the second one is the role of the word in the first sentence and the third one — its role in the succeeding sentence. Using this representation,

this triplet also appears in the first sentence and is similar to the Grammatical and Semantic Roles (GSRs) mentioned in our earlier work on tag-topic models [Das et al., 2011].

In our experiments these roles belong to part-of-speech or syntactic or dependency relation classes of the words and only one role is allowed per word that we decide using some grouping and prioritization rules (see Section 5.3.1). The words and the triplets in the light blue rectangular bubble are thus indicative of a latent topic of the document over some controlled vocabulary that reinforces the topical content of the terms in the document itself. Note that we apply our tag-topic models in a **global** way since we use the entire corpus of documents irrespective of the docset partitioning.

We observe that the central ideas in a document are often conveyed in written English through syllogisms. Syllogisms are logical inference constructs that often lead to the propagation of certain important concepts similar in spirit to “centers of utterances” as in [Grosz et al., 1995]. The propagation of these centers, be they entities or otherwise, are a major contributor to the high frequency of certain open-class words in the documents. The triplets like (sleep, noun, noun) consisting of terms and their roles in adjacent sentences are thus a natural choice in the document level perspective for the tag-topic models.

The main content of the document itself can be structured in many ways. In this example, each word either belongs to a particular Named Entity class or not. In figure 5.1, we show 5 such classes with the corresponding phrases in the text appropriately color coded and underlined. The words not colored do not belong to any Named Entity class. The black arcs show extremely fine-grained semantic dependencies that exist between selected words. In our case study, an important observation is that salient high frequency verbs (i.e. verbs that do not fall into the category of standard English stopwords) across a document identify the main events to a considerable degree. In this example, we become aware that something is being discussed around the concept of “sleep deprivation”. If verbs like “impair” or “deprive” occur frequently in the documents across the docset, then we actually recover the query! The influence of models that are local to a docset and global theme generation models whose topics are influenced through word and document level annotations is thus quite apparent.

Topic models are usually trained on a part of data that is different in count proportions to the held-out test data. From the perspective of summarizing a fixed set of documents, we treat the documents with their annotations as training data. However, in our setup, the posterior inference for the latent topics is done for *sentences*, that act as held-out test set. The inference is influenced by not only the terms and the corresponding annotations in the sentence but also the document level perspective obtained from an adjacent contexts. This has the advantage of including other sentences from any new but related documents as potential summary sentences.

5.1.3 Rhetorical Structure Trees as a Local Model

Earlier we had mentioned that to satisfy the categorical aspects of unified information model as laid down by the Guided Summarization paradigm and not covered by the Named Entity classes, we need to understand how sub-sentential spans interact with each other that exhibit some meaningful relationships. By sub-sentential span we mean that if a sentence is looked upon as a sequence of words, a sub-sentential span is just a sub-sequence. By meaningful relationships we mean whether a span is related to another span through relations such as attributions, background information, cause for the event in the other span, elaboration on the event in the other span etc. Finding such relations in free text helps us identify the

potential subtle categorical aspects of the Guided Summarization principle. For example we are able to uncover facts regarding the category “Trials and Investigations” and aspects like “what is the issue? how are parties affected? why it happened?” and so on.

RST literature [Mann and Thompson, 1988, Marcu, 2000a] lays special emphasis on cue words or phrases which are sentence level connectives like “because”, “nevertheless”, “that”, “but”, “in spite of”, parenthesis etc. and certain punctuations that serve primarily to indicate document structure or flow. Elementary Discourse Units (henceforth EDUs), which are non-overlapping contiguous spans of text, can be extracted based on these cue phrases along with syntactic parse tree information, lexical rules and probabilities tuned against a training set of such annotated sentential structures [Soricut and Marcu, 2003] (or even a strongly constrained first-order logic model [Marcu, 2000a]). In figure 5.1, the square-bracketed textual extents represent such sub-sentential spans as recognized by cue words. RST emphasizes the fact that certain shallow processing of text in terms of cue phrase analysis in combination with well-constrained mathematical models can be used to create valid rhetorical structure trees (henceforth RS-trees) of unconstrained natural language text.

Rhetorical parsing allows a piece of text to be partitioned into non-overlapping spans which lend themselves into a binary tree where the leaves from left to right indicate EDUs that are related in strict rhetorical sense. Any internal node signifies a relationship between its children i.e. the text extents only spanned by the children. The root node connects all the spans in the text (possibly) through internal nodes – it is possible for a sentence to be just a root node. The spans of the text are of two types - text spans that consume subsidiary information are called **satellites** and all others are called **nuclei**. All satellites are related to their corresponding nuclei through some valid rhetorical relations. It is also possible for a text not to have any satellites as determined by the rhetorical relations and rules of RST. Figure 5.2 shows the RS-tree of the second sentence in Fig. 5.1.

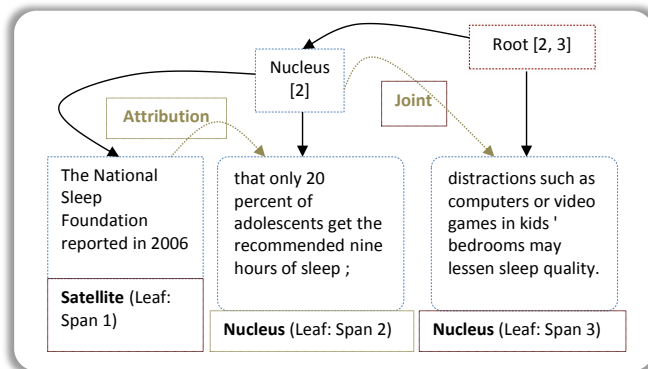


Figure 5.2: Rhetorical Structure tree of the second sentence in Fig. 5.1

We observe that the second sentence has been partitioned into three non-overlapping spans by identifying the cue phrase markers. The leftmost span (span 1) is the only satellite leaf node that is related to the middle nucleus leaf (span 2) through the “Attribution” relation. In other words the leftmost leaf is presenting information that is subordinate and can be attributed to the main information presented by the middle leaf node. The right most leaf (span 3) does not present any subsidiary information and hence is a nucleus. Spans 1 and 2 thus get connected through an internal node and span 3 can only be connected with the parent of 1 and 2 through the root node. In Fig. 5.2, lighter arcs indicate RST relations and

darker arcs indicate the directed edges of the RS-tree. The parent for spans 1 and 2 becomes a nucleus signifying that span 2 is more important. The root indicates that both spans 2 and 3 are jointly important and that they are related through the rhetorical relation of “Joint”.

In this chapter, the RS-trees of the sentences has been generated using the techniques used in Soricut and Marcu’s SPADE system [Soricut and Marcu, 2003]. We have slightly modified their accompanying software², to incorporate minor modifications and bug fixes. Better RS-tree generation is indeed helpful and is a separate direction of research [Hernault et al., 2010] which we do not pursue here. The rhetorical relations that hold between different spans of text are the same as those used in [Soricut and Marcu, 2003]. We consider only the following relations for the satellite spans to be useful for our purposes: {Background, Cause-Result, Cause, Comparison, Consequence, Contrast, Explanation and Temporal} to locally emphasize the aspects of the topic-categories that are more subtle and cannot be handled by Named Entity annotation. In fig. 5.2, we can think of spans 2 and 3 as good summary spans from the second sentence because of a global or background topic focus, presence of topically salient numeric entities, relevance to the query and the importance of the spans.

However, the problem of selecting “good summary spans” is also not trivial if we consider selecting a relevancy threshold using only unsupervised means. We look at a possible solution that works quite well for the datasets at hand using the technique of unsupervised density estimates [Kvam and Vidakovic, 2007] on the values of the cosines of the spans to the bag of selected keyterms from the docsets (see Sections 5.4.1 and 5.4.4). We are thus motivated to use both background tag-topic models that look at the corpus as a whole and local models that work at a docset level or a sentential level (see Section 5.4) for the Guided Summarization task.

Figure 5.3 emphasizes the fact that the local models collaborate with more holistic corpus based thematic models to weigh the candidacy of each sentence in a summary. The sentences (with their contexts) are fit to topic models trained over *documents* across docsets and their likelihoods are used to decide their inclusion into a summary. The intuition behind this is that if a sentence reflects a document’s topic proportions well, it satisfies the topical perspective to a greater extent.

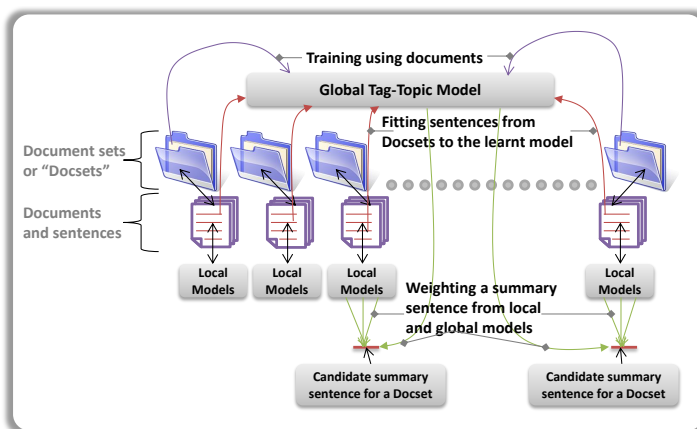


Figure 5.3: Our proposed summarization system architecture using global tag-topic models and local linguistic models.

This article is organized as follows: Section 5.2 discusses some existing multi-document summarization approaches that involve topic and non-topic models and highlights the key benefits and shortcomings of those models. Section 5.3 briefly discusses the tag-topic models as described in [Das et al., 2011]. Section 5.4 elaborates on the different local models that we use. Several summarization algorithms and sentence scoring criteria are discussed in Section 5.5. Section 5.6 concludes this article with a note on future directions.

²<http://www.isi.edu/licensed-sw/spade/>

5.2 Related Work

5.2.1 Existing Topic Model Based Summarization Approaches

Our approach to summarization has been motivated by some key studies in the area of application of topic models to multi-document summarization [Daumé and Marcu, 2006, Haghighi and Vanderwende, 2009, Tang et al., 2009, Celikyilmaz and Hakkani-Tür, 2011]. In general topic model based summarization techniques build separate models, either explicitly or implicitly, for each individual docset. Although there is no docset label bias for such models but the models are still applied to individual docsets for the purpose of identifying most *probable* n-grams (usually unigrams) within a docset is a rigorous probabilistic framework. This local modeling has a good implication in that there is no chance of drawing a word that is outside of the docset’s vocabulary. Some models also “look at” other docsets or related document collections and treat them as background distributions of n-grams [Haghighi and Vanderwende, 2009, Delort and Alfonseca, 2012].

The summarization technique in [Daumé and Marcu, 2006] cleverly implements these two strategies in a single model by using relevancy indicators of documents to queries. A reason for such a strategy is that latent spaces built out of the documents across all docsets can potentially account for more choices of words to be assigned to the same topic where the words may not be describing the event of the docset – this is particularly true of the frequently used open-class words co-occurring with more docset specific keyterms. Similar approaches are taken in [Tang et al., 2009, Chen et al., 2009] where docset specific topic models are augmented to include topic distributions over possible query terms. However it is often seen that the local topic models do not improve summarization scores by themselves and often docset specific algebraic models like weighting sentences through query term importance or constructing sentence affinity graphs [Chen et al., 2009] are needed as well.

An important thing to note in [Haghighi and Vanderwende, 2009] and [Delort and Alfonseca, 2012] is that although their “background” corpus essentially consists of all the documents *across* all docsets (or any other related document collection), there is a single latent topic that describes such a collection. Under such a constraint it is always the case that the most frequently occurring closed-class words which turn out to be stopwords for standard English collections dominate such a background topic. They similarly have a few other latent topics collapsing on the set of documents in a particular docset and on individual documents in that docset. This view is unlike the latent topic modeling view (as in [Blei et al., 2003, Celikyilmaz and Hakkani-Tür, 2011]) where the latent topics are allowed to fit to the data in a way that is governed by the data itself and the number of latent topics becomes a model parameter.

Also in models like those in [Haghighi and Vanderwende, 2009], related datasets are used for tuning the parameters and hyperparameters through training and testing on development sets and measuring summarization performance w.r.t ROUGE scores. By doing this their techniques necessitate the need for available human summaries. This makes it very hard to port the techniques to other summarization domains like biomedical and scientific literature, forums, books and novels etc. where multiple human summaries may not be available. Further, in the topic models presented in [Daumé and Marcu, 2006, Haghighi and Vanderwende, 2009, Celikyilmaz and Hakkani-Tür, 2011] is that it is unclear as to how supervised structured predictions on the text, for example Named Entities, can be used to influence the calculation of the topic distributions. Our topic models presented in [Das et al., 2011] are specifically designed to overcome this limitation.

5.2.2 Existing Linguistic and Vector Space Model Based Summarization Approaches

As a criticism to applying a topic model only on the documents of a docset is that it is equivalent to identifying the n-grams in a non-parametric way such as identifying topic signature terms as in [Lin and Hovy, 2000] which is fairly robust. Algebraic methods like finding the centroid of a document cluster [Radev et al., 2004] or using term-sentence incidence matrices can also be used in place of parametric statistical topic models and can be very successful specially in light of some supervised training [Schilder and Kondadadi, 2008]. A prominent work in that direction is the CLASSY system [Conroy et al., 2010, Conroy et al., 2011] which is continually tuned to maintain its position among the top automatic multi-document summarization systems in the TAC competitions.

In most topic model based summarization systems, the topic models are themselves part of the local models or use a single topic to capture the predominant background unigram or bigram distributions. This is very similar to the binomial log likelihood ratio model that computes topic signature terms [Lin and Hovy, 2000] using term distributions both of the current docset and that of the other docsets merged together. The success of the CLASSY system [Conroy et al., 2010] is also primarily dependent on this approach. CLASSY uses a host of hand crafted and automatically tuned linguistic and non-linguistic features to deliver a consistently good performance in multi-document summarization tasks of the TAC competitions.

Working with frequency statistics of bigrams can sometimes be very useful for multi-document summarization instead of just using unigrams [Banko and Vanderwende, 2004]. However representing all possible bigrams in a topic model is computationally expensive [Wallach, 2006] and we settle for a compromise by preprocessing the data and treating the Named Entities as a single concatenated unigram. We do this because the most frequent bigrams are usually seen to be part of Named Entity classes in our dataset. It is mentioned in [Darling, 2010] that using bigrams in local models can actually reduce ROUGE scores due to the sparsity. Although it has been seen that bigrams do actually improve the ROUGE scores for the Update Summarization task when used in a local topic modeling approach [Delort and Alfonseca, 2012].

The work in [Darling, 2010] raises a set of very basic but extremely useful questions starting from sentence weighting to sentence compression for summarizing a predefined set of documents. In their system, sentence weighting through query terms improved ROUGE scores significantly and so did compression of sentences – which has also been verified earlier in [Yih et al., 2007]. For example, consider a sentence like “At the heart of the rebuilding is the creation of a lasting memorial which will honor the memories of those we lost and help tell their story to the world, said New York Governor George Pataki.” Removal of the clause “said New York Governor George Pataki.” compresses the sentence by 6 words whilst not losing the central fact that the speaker mentioned. From the perspective of RST, this clause can be seen as a “satellite” that “elaborates” on a fact conveyed in the “nucleus.”

A different kind of summarization technique can be found in [Genest and Lapalme, 2011] where the docset specific (subject-verb-object) triplets are used in sentence generation. The selection of verbs followed some criteria that lead to better sentence generation. Sentence level syntactic parses are used to extract parts of the syntactic trees that can be provided as input to the SimpleNLG tool [Gatt and Reiter, 2009] for complex sentence generation through writing out noun phrases, prepositional phrases, verb complement and verb phrases. Interestingly, the generated sentences often resemble the EDUs obtained from a RS-tree. The rhetorical analysis of textual units for summarization has been attempted before

[Marcu, 1999] but it missed a rigorous data-driven analysis from which relevant spans can be selected. We feel that this direction need to be explored more.

In general, simultaneously satisfying the objectives of multi-document summarization i.e. relevancy to information need, non-redundancy between textual units in the summary and constraint on summary length has recently been proved to be NP-Hard [McDonald, 2007]. However, since the objective of latent structure discovery through topic models is very very different from the objective function of multi-document summarization as in [McDonald, 2007], we want to see if such latent structure discovery can indeed aid the targeted local models to generate better summaries.

To the best of our knowledge, this is the first work that uses tag-topic models as background global models as well as docset specific local models and Rhetorical Structure trees to simultaneously perform latent topic discovery and summarize multiple target documents in the form of bullet list summaries.

5.3 The Tag-Topic Models

In this section we briefly describe the multimodal tag-topic models from our previous work in [Das et al., 2011] and described fully in Chapter 4. The TagLDA model from [Zhu et al., 2006] and the the multi-modal METag²LDA model and the correspondence Corr-METag²LDA models from [Das et al., 2011] are shown in Fig. 5.4. The two main parameters of interest in the models are the ones that lead to the observation of a word w in a document d i.e. $w_{d,m}$. The word generation probabilities are obtained by the product of two distributions - β – the marginal topic distributions over the word vocabulary and π – the marginal WL tag class distributions over the same vocabulary. The “M” stands for Multinomial and the “E” stands for Exponential of log-sums of parameter components respectively. For a full symbol manifest and generative story of these models see Chapter 4. In the multi-modal tag-topic models, the parameter ρ represents the K topic multinomials over the document level (DL) tags where K is a parameter for all tag-topic models which denotes the number of topics.

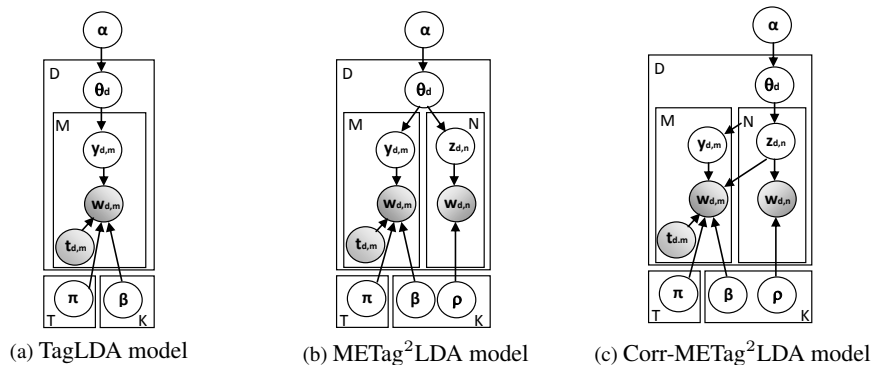


Figure 5.4: Graphical model representations of the tag-topic models used in modeling the corpus

In this chapter (unlike in Chapter 4), we augment the tag-topic models with an asymmetric Dirichlet prior over the distributions for the per document topic proportion random variates. We use the procedure described in [Blei et al., 2003]. In other words we optimize the K dimensional α in figure 5.4 such that each dimension influences the remaining ones. The use of this prior was also motivated by the work in [Wallach et al., 2009].

In our context, the models in Fig. 5.4 are *global* models in the sense that all documents form a TAC collection is used as input. So in Fig. 5.4, $d \in \{1, \dots, D\}$ can be *any* document from *any* docset in a TAC collection. Referring to the example document in figure 5.1, $w_{d,m}$ is a word in a document d such as “impair”, $w_{d,n}$ is a DL tag like the triplet “(sleep, noun, noun)” and $t_{d,m}$ is the WL tag class associated with word $w_{d,m}$ (e.g. “Normal_Word” for “impair” and “DATE” for “2006”, “ORG” for “Ohio_State_University” etc.). Note that a DL tag can also be a normal word like “sleep” or “distractions” if that word is deemed important by document frequency measurements. $z_{d,n}$ and $y_{d,m}$ are indicator variables for selecting 1-of- K topics. $y_{d,m}$ in the Corr-METag²LDA, however, is an indicator variable for selecting one of the corresponding DL tags. ρ are the K latent topic distributions over the $corrV$ DL tags in the training vocabulary; $\beta_{1:K}$ are the K latent topic distributions over the V words and $\pi_{1:T}$ are the T WL tag class distributions over the V words in the vocabulary as well. M is the number of positions in a document d where a (word,WL-tag-class) pair is observed and N is similarly the number of positions in the same document d for the DL tags.

We now digress a little to describe in brief on how the plain text documents are converted into appropriate inputs for the tag-topic models.

5.3.1 Data Preparation for the Tag-Topic Models

Our data pre-processing for the tag-topic models begin with the removal of all standard English stopwords and all words or Named Entity phrases appearing only once across the corpus are removed also. The remaining words are used in their lemmatized form. Following the example of the word “sleep” in fig. 5.1, the same lemmatized words in consecutive sentences are extended with the the part-of-speech tags or the syntactic dependency labels (i.e. a triplet like (sleep, noun, noun)). These are the primary source of the tags in the DL perspective. This is similar in spirit as [Barzilay and Lapata, 2005b] but not restricted to Named Entities only. In this paper we refer to such a triplet as *coherence triplet*. When a word’s surface form is repeated *in a sentence* we use a prioritization ordering of $\{subject > object > noun > adjective > verb > adverb > other\}$ to select only one tag to form the coherence triplet.

Since we do not rely on co-reference resolution, in order to lessen the sparsity of such DL tags for sentence selection, we also added the top 5 most frequent non-stopwords and top 10 tf-idf terms per document into the DL tag vocabulary. Although this can potentially result in additional 100 words per docset, the actual number is much less owing to numerous repetitions. The minimum number of these features are decided based less than five percent error on document event classification performances (see Section 5.4.2).

It has been quite surprising to find that the set intersection size between the 5 most frequent words and the set of all the first elements of such coherence triplets is 3.5 on average per document even without co-reference resolution. This bolsters the fact that coherence properties across sentences can indeed aid in forming a robust document level perspective.

We use the Named Entity annotation classes as WL tag classes and a “Normal_Word” tag class for all other words. All entities are automatically recognized as {Location, Misc, Number, Organization, Person} using the Stanford NLP suite³. The Date and other numeric categories were included within the Number category. These tags are not always completely orthogonal to each other and sometimes, though not often, the same surface string appears in different NE classes – particularly in the Misc class.

³<http://nlp.stanford.edu/software/corenlp.shtml>

While training, each document as a whole, with automatically annotated DL and WL perspectives, is considered without regard to any docset specific information. While performing inference on the documents within a docset, we treat each sentence in context. The context is used for DL perspective generation only. A sentence context consists of a central sentence consisting at least one Named Entity and either its immediate preceding sentence or its immediate succeeding sentence or both depending on the location of the central sentence in the document. For each such central sentence and its context we associate the DL and WL perspectives using the word level tag class vocabulary and the document level tag vocabulary obtained during training. These contexts can be created immediately after a document has been processed for input to train the models (as in our experiments) or at a later time specially for new target documents. Summary sentences for a docset topic are chosen from among the central sentences collected this way. This way of creating a “test-set” for summarization is intuitive — any sentence that reflects the document structure more under the assumptions of the generative model, should be a better candidate for the final summary.

A thorough description of the tag-topic models, optimized using the Variational Bayesian framework can be found in [Das et al., 2011] and is not repeated here in full for brevity. Rather, we briefly familiarize ourselves with the key variables and expressions of interest from the multi-modal tag-topic models in the next subsection.

5.3.2 Descriptions of the Bi-Perspective Tag-Topic Models

In this subsection, we highlight the free and model parameter updates of the multi-modal tag-topic models as those shed light on the differences between the models. To find as tight as possible an approximation to the log likelihood of the data (the conditional distribution of the observed variables given the parameters), the KL divergence of an approximate factorized mean field distribution is minimized with respect to the true posterior distribution of the latent variables given the data. A fully factorized distribution, denoted by q , with “free” variational parameters γ , ϕ and λ is imposed as

$$q(\boldsymbol{\theta}, \mathbf{z}, \mathbf{y} | \gamma, \phi, \lambda) = \prod_{d=1}^D q(\boldsymbol{\theta}_d | \gamma_d) \left[\prod_{n=1}^{N_d} q(z_{d,n} | \phi_{d,n}) \prod_{m=1}^{M_d} q(y_{d,m} | \lambda_{d,m}) \right] \quad (5.1)$$

where ϕ are the free parameters of the variational DL_tag distributions over topics and λ to be the free parameters of the variational word distributions over topics (in METag²LDA model) or word distributions over DL_tags, $w_{d,n}$, (in the Corr-METag²LDA model). These free parameters are defined for every document d . The variational parameter $\gamma_{d,i}$ which is a surrogate for $\theta_{d,k}$ reflects the posterior expectation of the number of ((word,word-tag-class), DLtag) ensembles assigned to topic k in document d .

Following [Das et al., 2011], we re-write the ELBO (Evidence Lower Bound), \mathcal{L} for the Corr-METag²LDA model as

$$\begin{aligned} \mathcal{L}_{CorrME} = & E_q[\log p(\boldsymbol{\theta} | \boldsymbol{\alpha})] + E_q[\log p(\mathbf{Z} | \boldsymbol{\theta})] + E_q[\log p(\mathbf{W} | \mathbf{Z}, \boldsymbol{\rho})] + E_q[\log p(\mathbf{Y} | \mathcal{N})] \\ & + E_q[\log p(\mathbf{W} | \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\pi}, \mathbf{t})] - E_q[\log q(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{Y}, | \gamma, \phi, \lambda)] \end{aligned} \quad (5.2)$$

where $E_q[f(\cdot)]$ is the expectation of $f(\cdot)$ over the fully factorized q distribution and \mathcal{F} is the ELBO to true likelihood. This ELBO is also directly related to measuring perplexity [Blei et al., 2003] and is

basically the log likelihood. Similarly for the METag²LDA model we have:

$$\begin{aligned} \mathcal{L}_{ME} = & E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\log p(\mathbf{Z}|\boldsymbol{\theta})] + E_q[\log p(\mathbf{W}|\mathbf{Z}, \rho)] + E_q[\log p(\mathbf{Y}|\boldsymbol{\theta})] \\ & + E_q[\log p(\mathbf{W}|\mathbf{Y}, \beta, \boldsymbol{\pi}, \mathbf{t})] - E_q[\log q(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{Y}, |\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\lambda})] \end{aligned} \quad (5.3)$$

5.3.3 Mean Field Inference

The basic idea of the mean field technique is simple: we limit the optimization of the maximum likelihood expressions of data given parameters for a probabilistic generative model \mathcal{M} such as LDA [Blei et al., 2003], TagLDA [Zhu et al., 2006], etc., to a subset of “tractable” distributions. The simplest choice is the family of product distributions such as the one in Equ. 5.1, which gives rise to the naive mean field method.

Mean field methods operate on the notion of a tractable subgraph, i.e. a subgraph \mathcal{F} of the original graph \mathcal{G} (see the original LDA model in [Blei et al., 2003]) over which it is feasible to perform exact or approximate calculations. Variational methods for optimizing parameters of a model under the mean field framework excel in cases where it is infeasible to compute expected sufficient statistics for the parameters conditioned on the given dataset. All models that are extended from LDA fall under this class. In such cases, this variational E-step thus involves replacing the exact mean parameter, under the current model \mathcal{M} , with the approximate set of mean parameters computed by a mean field algorithm. If we restrict the mean parameters of the factorized distributions to those of the distributions from the exponential family, this expectation becomes much easier to compute. The interested reader is referred to [Wainwright and Jordan, 2008] for more details on mean field theory methods. We now briefly touch upon the key expressions in the tag-topic models that will help us extract candidate summary sentences.

As in Chapter 4, following the inequality, $\log(x) \leq \zeta^{-1}(x) + \log(\zeta) - 1, \forall \zeta > 0$, the ELBO, $\mathcal{L}_{\mathcal{M}}$ must also be optimized for the ζ variable for every document d and every word $w_{d,m}$ in it. The expression for $E_q[\log p(\mathbf{w}_m|\mathbf{y}_m, \beta, \boldsymbol{\pi}, \mathbf{t})]$ can be written for the Corr-METag²LDA model as:

$$\begin{aligned} E_q[\log p(\mathbf{w}_m|\mathbf{y}_m, \beta, \boldsymbol{\pi}, \mathbf{t})] \geq & \sum_{m=1}^{M_d} \sum_{i=1}^K \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) \left(\log(\beta_{i,w_{d,m}} \times \pi_{t_{d,m},w_{d,m}}) \right) \\ & - \sum_{m=1}^{M_d} \left\{ \zeta_{d,m}^{-1} \left(\sum_{i=1}^K \sum_{v=1}^V \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) e^{\left(\log(\beta_{i,w_{d,m}} \times \pi_{t_{d,m},w_{d,m}}) \right)} + \log \zeta_{d,m}^{-1} - 1 \right) \right\} \end{aligned} \quad (5.4)$$

Using these lower bounds and the maximum likelihood estimates of the hidden variables in document d are as follows:

$$\gamma_{d,i} = \alpha_i + \sum_{n=1}^{N_d} \phi_{d,n,i} \quad (5.5)$$

$$\zeta_{d,m} = \sum_{v=1}^V \sum_{i=1}^K \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) e^{\left(\log(\beta_{i,w_{d,m}} \times \pi_{t_{d,m},w_{d,m}}) \right)} \quad (5.6)$$

$$\phi_{d,n,i} \propto \exp \left\{ \psi(\gamma_i) - \psi\left(\sum_{j=1}^K \gamma_j\right) + \log \rho_{i,w_{d,n}} + \sum_{m=1}^{M_d} \lambda_{d,m,n} (\log \beta_{i,w_{d,m}} + \log \pi_{t_{d,m},w_{d,m}}) - \sum_{m=1}^{M_d} \zeta_{d,m}^{-1} \lambda_{d,m,n} \left[\sum_{v=1}^V e^{(\log(\beta_{i,w_{d,m}} \times \pi_{t_{d,m},w_{d,m}}))} \right] \right\} \quad (5.7)$$

$$\lambda_{d,m,n} \propto \exp \left\{ \sum_{i=1}^K \phi_{d,n,i} (\log(\beta_{i,w_{d,m}} \times \pi_{t_{d,m},w_{d,m}})) - \zeta_{d,m}^{-1} \left[\sum_{v=1}^V \sum_{i=1}^K \phi_{d,n,i} e^{(\log(\beta_{i,w_{d,m}} \times \pi_{t_{d,m},w_{d,m}}))} \right] \right\} \quad (5.8)$$

5.3.4 Parameter Estimation

The expressions for the maximum likelihood (ML) of the parameters of the Corr-METag²LDA model using derivatives w.r.t the parameters of the functional \mathcal{L}_{CorrME} are obtained as follows:

$$\rho_{i,j} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{j=1}^{corrV} \phi_{d,n,i} \delta(w_{d,n}^j) \quad (5.9)$$

$$\begin{aligned} \log \beta_{i,j} &= \log \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{j=1}^V \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) \delta(w_{d,m}^j) \right) \\ &\quad - \log \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{j=1}^V \zeta_{d,m}^{-1} \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) e^{\log \pi_{t_{d,m},j}} \delta(w_{d,m}^j) \right) \\ &= \log(term_1^\beta) - \log(term_2^\beta) \end{aligned} \quad (5.10)$$

$$\begin{aligned} \log \pi_{t,j} &= \log \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{j=1}^V \sum_{t'=1}^T \sum_{i=1}^K \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) \delta(w_{d,m}^j) \delta(t_{d,m}^{t'}) \right) \\ &\quad - \log \left(\sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{j=1}^V \sum_{t'=1}^T \sum_{i=1}^K \zeta_{d,m}^{-1} \left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i} \right) e^{\log \beta_{i,j}} \delta(w_{d,m}^j) \delta(t_{d,m}^{t'}) \right) \\ &= \log(term_1^\pi) - \log(term_2^\pi) \end{aligned} \quad (5.11)$$

where $\delta(a_c^b)$ is the delta function which means $\delta(a_c^b) == 1$ iff $b == a_c$

Note that the updates for β and π are coupled. Also, since the updates for β and π are unconstrained, a Gaussian regularizer with 0 mean and a constant standard deviation, σ , is used for **every** $\beta_{i,j}$ and $\pi_{t,j}$. The expression for $\mathcal{L}_{i,t}$ is transformed to

$$\widehat{\mathcal{L}_{\mathcal{M}_{[i,t]}}} = \mathcal{L}_{\mathcal{M}_{[i,t]}} - \frac{1}{2\sigma^2} \left(\sum_{v=1}^V (\exp(\log \beta_{i,v}))^2 \right) - \frac{1}{2\sigma^2} \left(\sum_{v=1}^V (\exp(\log \pi_{t,v}))^2 \right)$$

So, taking the derivative of $\widehat{\mathcal{L}}$ w.r.t $\log \beta_{i,v}$ or $\log \pi_{i,v}$ and solving for $\exp(\mathcal{T})$ where \mathcal{T} is $\log \beta$ or $\log \pi$ we obtain, (letting $A = \exp(\mathcal{T})$),

$$2A = -\sigma^2 \text{term}_2^{(\cdot)} + \sigma \sqrt{\sigma^2 (\text{term}_2^{(\cdot)})^2 + 4 \text{term}_1^{(\cdot)}} \quad (5.12)$$

as pointwise MAP (maximum a posteriori) estimates for $\log \beta$ or $\log \pi$. Without this regularization convergence is not achieved in the TagLDA class of models [Boyd-Graber, 2010]. We next highlight the differences between METag²LDA and Corr-METag²LDA models.

Corr-METag²LDA is a strongly constrained model which becomes apparent once we observe equations (5.10) and (5.11). A topic's influence over a textual word is obtained by marginalizing out the influences of the corresponding data on it in the document. The more the corresponding data (DL tags) in a document is about a topic the more likely it is that the textual data in the document is also about that topic. This assumption is relaxed in the METag²LDA model. In METag²LDA, the relation between the DL tags (i.e. $w_{d,n}$ s) and the textual data (i.e. $w_{d,m}$ s) are somewhat loose - overall it is possible that two different topics in a document can independently be responsible for the pattern of co-occurrence of the ((word,word-tag-class),DLtag) ensembles. This is apparent by observing the β and π parameter updates for METag²LDA in [Das et al., 2011] where $\left(\sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,i}\right)$ is just replaced by $\lambda_{d,m,i}$ —the variational multinomial parameter for the i^{th} topic governing word w_m in document d in the factorized subgraph corresponding to the original graph of the model.

To optimize the α parameters, we first select out the expressions from $\mathcal{L}_{(\cdot)}$ that depend on α (as in [Blei et al., 2003]) and optimize using Newton's iterative gradient based method as in [Minka, 2009]. Optimizing α_i is dependent on the value of α_j through:

$$\begin{aligned} \mathcal{L}_{(\cdot)[\alpha]} &= \sum_{d=1}^D (\log \Gamma(\sum_{i=1}^K \alpha_i) - \sum_{i=1}^K \log \Gamma(\alpha_i) + \sum_{i=1}^K (\alpha_i - 1) (\Psi(\gamma_{d,i}) - \Psi(\sum_{k=1}^K \gamma_{d,i}))) \\ \frac{\partial \mathcal{L}_{(\cdot)}}{\partial \alpha_i} &= D(-\Psi(\alpha_i) + \Psi(\sum_{j=1}^K \alpha_j)) + \sum_{d=1}^D (\Psi(\gamma_{d,i}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) \\ \frac{\partial \mathcal{L}_{(\cdot)}}{\partial \alpha_i \alpha_j} &= \partial(i, j) D \left(\Psi'(\sum_{j=1}^K \alpha_j) - \Psi'(\alpha_i) \right) \end{aligned} \quad (5.13)$$

where γ_d is the K -dimensional parameter of the variational Dirichlet distribution corresponding to θ_d as in [Blei et al., 2003]. The gradient ascent algorithm involves finding a Newton step which is the $\text{Hessian}^{-1} \times (\text{negative gradient})$ of $\mathcal{L}_{(\cdot)[\alpha]}$. We multiply this quantity with some suitable scalar found by backtracking line search and add this to the current estimate of $\alpha_{d,i}$ and keep iterating until convergence.

5.3.5 The Need for Informative Priors for Topic Proportion Distributions

While it is quite easy to measure the relative performances of topic models for particular values of parameters, however, measuring the outputs of the topic models qualitatively by using humans in the loop is often time consuming and difficult. The most notable example in this direction is the evaluation done in [Chang et al., 2009]. It has been observed that humans *prefer* sparsity in topics by not preferring topics to influence each other. Wallach et. al. [Wallach et al., 2009] also experimented on topic sparsity by using priors at different hierarchies of the topic models and concluded that it is best to use an asymmetric Dirichlet prior over the distribution of the topic proportions θ_d . The asymmetric Dirichlet prior i.e. α in

Figs. 5.4a, 5.4b and 5.4c consists of a base measure \mathbf{m} and a concentration parameter c with $\alpha_i = c \times m_i$ [Wallach et al., 2009]. It has a great advantage over its symmetrical counterpart. It can vary the sparsity in individual topics such that the θ_{ds} can be drawn from at least two different topic modalities. One or more topics get assigned common words that frequently co-occur with other words in every document—if we do not remove the stopwords then one of the topics will solely be focusing on these. This is similar to the background topic in [Delort and Alfonseca, 2012]. The other topics will focus on the actual topical words of the documents and thus lead to sparsity.

k	Normal Word	Person	Organization	Location	Misc	Number
30	food company recall product dog pet safety die sell death cat owner test kidney failure	Henderson Sundlof [Sarah Tuite] Iams Tuite Nelson [Paul Henderson] Burnton	[Menu Foods] FDA Iams [Food and Drug Administration] WalMart Kroger PetSmart	China [United States] Mexico U.S. Canada Chinese Arizona Ontario Beijing	[Menu Foods] cat Chinese pet Tootsie Canadian [North American] Bernese room	Monday Friday Saturday Wednesday Tuesday Sunday Thursday [March 6] [two week]
	β_{30} : food company pet recall dog cat product kidney [Menu Foods] sell safety brand failure test eat death die supplier wheat poisoning					
7	β_7 : flood country Bangladesh river district northern water kill situation relief level government monsoon inundate rain northeastern					
10	β_{10} : airport police attack car security incident level terminal building London close raise Glasgow British alert arrest [Glasgow Airport]					
75	β_{75} : turtle endanger poach sea fisherman water egg species police jail group beach marine dead Sabah catch fishing fine protect Malaysia					
0	β_0 : Madoff investor money firm fund pay foundation invest son SEC charge jewish [New York] hedge business lose part electronic					
32	β_{32} : [The Associated Press] [Timberly Ross] Colo. [Coast Guard] [European Union] kill Calif. WFP Monday Tuesday government accord country					

Table 5.2: Latent topics from the TAC 2011A dataset for $K = 80$ using asymmetric Dirichlet prior α over θ_d in TagLDA. Terms within square brackets [] are Named Entity phrases that are treated as single concatenated tokens.

This intuitively makes sense if we observe the topics from TagLDA from table 5.2. Table 5.2 shows some sample latent topics (the marginal β from Fig. 5.4a) learnt from the TAC 2011A data. The first row shows topic 30 conditional on Named Entities and Normal words that receive prominent focus in the topic. The next four rows just show the marginal topics, β , randomly sampled. The last row showing topic 32 highlights the use of the asymmetric prior.

Topic 32 (and some other similar topics) have clustered Named Entities and words, like “the_Associated_Press,” “kill”, etc. that occur frequently in many documents and are not removed by stopword and low frequency word removal. Words like these do have the tendency to dominate *all* latent topics if an asymmetric α prior is not used for θ_d . The reduction of this domination may be lowered in the case of the Tag²LDA class of models depending upon the nature of the corresponding DL tags.

Topics imputed to *vocabulary words* in sentences from news documents annotated both at the word and meta-data level are shown in Fig. 5.5. We reiterate though that the topics are learnt using a document level partition of the corpus while during summarization, we fit a contextual sentence partitioned corpus to the learnt models. Three sentences are shown in Fig. 5.5a with the middle grayed one being a common context of the first and third sentences and not containing any Named Entities. We chose this example

A Western drought that began in 1999 has continued after the respite of a couple of wet years that now feel like a cruel tease.

But this time people in the driest states are not just scanning the skies and hoping for meteorological rescue.

Some \$2.5 billion in water projects are planned or under way in four states, the biggest expansion in the West's quest for water in decades.

(a) Some sample sentences

A Western drought that began in 1999 has continued after the respite of a couple of wet years that now feel like a cruel tease.

Some \$2.5 billion in water projects are planned or under way in four states, the biggest expansion in the West's quest for water in decades.

cyclone Bangladesh storm coastal
wind shelter evacuate India hit
severe government thousand
district coast disaster

coral trade grow reef species
world scientist water fish
Nedimyer wildlife marine
threaten effort jewelry

(b) Keywords tagged with TagLDA

A Western drought that began in 1999 has continued after the respite of a couple of wet years that now feel like a cruel tease. [Western drought]

Some \$2.5 billion in water projects are planned or under way in four states, the biggest expansion in the West's quest for water in decades. [plan state (state,noun,noun) water (water,noun,object)]

city superintendent bomb hospital
historic tourist explosion crowd

turtle endanger poach sea
fisherman water egg species

city superintendent bomb hospital
historic tourist explosion crowd

turtle endanger poach sea fisherman
water egg species

(c) Keywords tagged with METag2LDA

A Western drought that began in 1999 has continued after the respite of a couple of wet years that now feel like a cruel tease. [Western drought]

Some \$2.5 billion in water projects are planned or under way in four states, the biggest expansion in the West's quest for water in decades. [plan state (state,noun,noun) water (water,noun,object)]

hospital
organ
donation
care wound

bombing
blast India
attack
tourist city

water level
meter
drought
reach

drug plan
safety food
country
quality

bridge Army
story
collapse
government

water state river Colorado Mexico shortage delta flow drought California
agreement reservoir Arizona Lake_Mead change U.S. Colorado_River plan Nevada

(d) Keywords tagged with Corr-METag2LDA

Figure 5.5: Some negative examples of topic annotation on *sentences* from news documents. The topic annotations are shown as color coded text. The text within the first row of bubbles indicate the topics which annotate the sentences. These are obtained by finding $k^* = \arg \max_{k \in \{1, \dots, K\}} \lambda_{d,m,1:K}$ for the TagLDA and METag²LDA models and $\arg \max_{k \in \{1, \dots, K\}} \sum_{n=1}^{N_d} \lambda_{d,m,n} \phi_{d,n,1:K}$ for the Corr-METag²LDA model. The text in the second row of bubbles for the Tag²LDA family of models denote the topic of the sentence obtained directly from $\arg \max_{k \in \{1, \dots, K\}} (\gamma_{d,1:K} - \alpha_{1:K})$

to show how the type of document level perspective can influence the quality of topical structures purely based on the assumptions of data generation.

Figures 5.5b, 5.5c and 5.5d shows a highlighting of a set vocabulary words in the first and third sentences through TagLDA, METag²LDA and Corr-METag²LDA respectively. Of the three models, TagLDA without any influence of the syntactic coherence meta-view of the sentences, ascribes words in the sentences to topics which indeed look plausible from a corpus point of view, however, does not explain the theme of the sentences very well. This example highlights both the usefulness and the difficulty of this approach to topic inference on shorter sentences for summarization. Although the annotations reflect topics which belong to a similar coarser event categories of the news corpora, TagLDA and METag²LDA show much lower variance in topic annotation both from the free variational distributions over the observations as well as from the sentence level topic proportions leading to consistent, and in this case, wrong, topic annotations. The Corr-METag²LDA shows much higher variance in topic annotations based on free variational distributions over the observations but annotates the topic of the sentence correctly based on the topic proportions. These observations led us to believe that likelihood fits of entire

sentences are better indicators of topic fit than using topic weights of individual observations separately. Only those sentences which reflect the topic of the document very well can be good candidates for the final summary. Incidentally, none of the sentences shown in this example have been included into the final summary.

The examples in Figs. 5.5c and 5.5d are even more interesting. The “document level” perspective of the sentences considered are shown in bold black font after each sentence. Words like “Western,” “drought,” “plan,” “state,” and “water” do not carry any notion of syntactic coherence but only that of document relevancy as described in Section 5.3.1. The triplets like “(water, noun, object)” do indeed carry some notion of coherence. The olive colored bubbles in the bottom rows in in Figs. 5.5c and 5.5d show the description of the topics for the sentences based on the topic proportion random variable θ .

The METag²LDA model is loosely coupled to the document level perspective and hence the topic attribution for the second sentence mimics that of TagLDA. The topic for the first sentence has completely drifted being dominated by the topics pertaining to many cruel attacks by terrorists who blame the West. The TagLDA model however correctly indicates the correct type environmental disaster albeit in an opposite sense. However METag²LDA still partitions these two sentences into a set of two different topics.

On the other hand, Fig. 5.5d shows that the Corr-METag²LDA model has been unsuccessful in ascribing the right topics to the individual words in the sentences except for the topic shown in the green bubble. Note that the topics are ascribed to the words in an indirect manner for the Corr-METag²LDA model: each word has a distribution over the document level perspective and each datum in the document level perspective has a distribution over topics. Thus, if the document level perspective is not *naturally generated*, the correspondence might give rise to poor topic attributions. The model, however, has been able to correctly assign the same *and correct* topic to both the sentences just based on topic proportions of the sentences.

5.3.6 Model Log Likelihoods

In this section, we investigate the power of the posterior inferences of the models based on our summarization setup. To re-iterate, during training the input is at a document level and during summarization, the input is at a sentence level. Figs. 5.6a and 5.6c show that the correspondence class of models show the best log likelihoods i.e. ELBOs in both Guided Summarization datasets. Note that the “Asym” suffix in the graph legends mean that the corresponding models employ the asymmetric Dirichlet prior α .

An interesting phenomenon to note in Fig. 5.6a is that TagLDA-Asym shows slightly poorer ELBOs as number of topics increase from 60. This means that TagLDA-Asym is favoring K to be close to the actual number of clusters i.e the 46 docsets in the TAC 2010A data. This could have been due to the more orthogonal nature of the Named Entities in the TAC 2010A data and the use of a fixed regularizer for all K settings in all our tag-topic models. This does not, however, affect the performance for the end-task of summarization or even event category detection since the objectives that the tag-topic models maximize is quite different from those for either of the tasks.

Figs. 5.6b and 5.6d show that the trend of training ELBOs does not hold true when we try to fit individual sentences and their contexts to the trained models as described in Section 5.3.1. In such a case, TagLDA shows better predictive ELBO and this reflects the choice of the DL perspective on the assumptions of the model. The figures also show signs of overfitting for 100 and 120 topics. This is possible because of very short sentences which are presented for posterior inference that can lead to

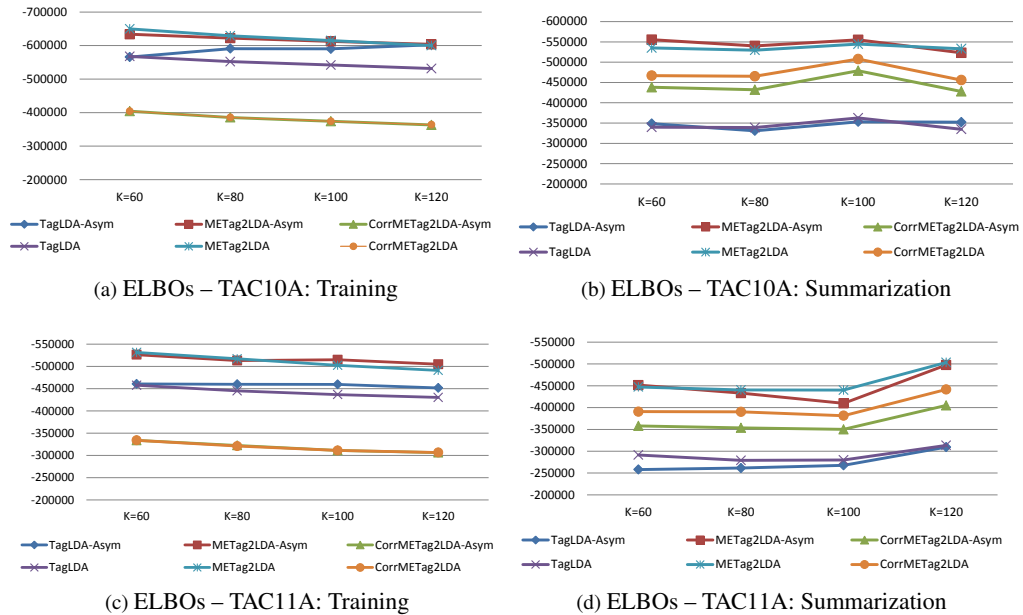


Figure 5.6: Evidence Lower BOunds (ELBO)s of the tag-topic models on the TAC 2010A and 2011A datasets – Lower is better.

poorer fits. Due to these kinds of test data points, we hold the α hyperparameters fixed while fitting sentences during summarization.

In our experiments, we created the DL perspective out of frequent words and coherence triplets to reflect the approximate attentional state that persists immediately after reading a document. While training, this indeed improves the likelihood of the Corr-METag²LDA due to the soft correspondence constraint through latent topics. However, at the sentence level, due to much lesser document level context to correspond to, TagLDA performs better by eliminating all needs for correspondence whatsoever. The models with asymmetric α also show higher ELBOs than their symmetric counterparts particularly for the correspondence class of models. Further the use of asymmetric α markedly improves event classification power (see Section 5.4.2).

Although the METag²LDA performs the worst in terms of likelihoods on sentence context fitting, we will see that the trend does not hold true for summarization performance. Evaluating the tag-topic models through summarization performance can open up another way of evaluating topic models where log likelihood measures may not shed much light on an application end task. This seems to tally with the human observations on latent topic summarization through bags of most probable words [Chang et al., 2009] where a model with higher likelihoods can actually do worse in a manual evaluation of topic interpretability.

5.3.7 Tag-Topic Model Evaluation through Multi-document Summarization

Although several different types of topic models have been used recently for the purpose of summarization that show promising results, however no studies have been conducted on how the topic models fare on the multi-document summarization task when compared to a purely local and docset specific

centroid based summarizer like the MEAD system [Radev et al., 2004], given, that the topic models are *not allowed* to use the partitioning of the corpus into document clusters. To this end, we first evaluate our extended tag-topic models through summarization performance measured against the MEAD⁴ with its default settings. We use the automatic ROUGE [Lin and Hovy, 2003] evaluation toolkit. In its default setting, MEAD do not consider any sentence lower than nine words. Also, among the models considered here, the TagLDA model does not consider any DL perspective at all and is the least complex of all the tag-topic models.

ROUGE is a recall oriented metric measuring the recall of n-grams of a system summary to that of the reference (human) summaries. To incorporate the aspect of fluency into the automatic summarization evaluation, bigrams and skip-4 bigrams are used to measure system performances in the TAC competitions and that is what we also use here to compare systems. It has been observed that for a minimum number of human summaries and docsets, the ROUGE-2 and ROUGE-SU4 scores have good correlation to the scores obtained from manual evaluation using PYRAMID [Nenkova and Passonneau, 2004].

ROUGE averages the system summary scores for all the docsets and also provides a 95% confidence intervals based on bootstrap sampling. For example, let us denote the mean ROUGE score of a system A to be m_A , the upper bound to be u_A and the lower bound to be l_A corresponding to m_A ; Although typically systems with higher mean ROUGE scores are preferred, for a system, say B , to perform significantly better than system A , m_B must be greater than u_A . It has been observed in multiple experiments that the ROUGE-2 bigram scores have very strong correlation to the ROUGE-SU4 scores and hence results using the former are not shown in this section to save space. Further ROUGE-SU4 is more robust to paraphrasing.

The MEAD system, which is a purely local docset specific summarizer, was run with the same settings as in the official TAC 2010 and 2011 Guided Summarization experiments. Full sentences are extracted without any pre or post processing. It is often seen the full summary length of 100 words is almost never satisfied when using full sentences. Since in most cases, longer sentences have higher chances of selection, some of the final summaries contained even 70 words. This is indeed a problem with fully extractive summarization systems which respect summary length as well as human readability. The sentences from our model are sorted in descending order of the sentence likelihoods from the models and the top ones are extracted as summaries as long as the total number of tokens in the summary remained within the 100 word limit. We just used an additional constraint in our summaries that no sentence in a summary can have a 50% word overlap with any other sentence in the summary. The sentence likelihoods for the Tag²LDA class of models are just the values of the expressions in Eqs. 5.2 and 5.3 for a particular central sentence. Using sentence likelihoods makes our system a *query-independent* summarizer just like MEAD.

Local Model	TAC 2010A	Conf.- Int.	TAC 2010B	Conf.- Int.	TAC 2011A	Conf.- Int.	TAC 2011B	Conf.- Int.
MEAD	0.09117	0.08676-0.09586	0.09645	0.09181-0.10104	0.11741	0.11141-0.12350	0.09147	0.08764-0.09528

Table 5.3: ROUGE-SU4 scores and confidence intervals of the summaries from the MEAD system for all base and update collections from TAC Summarization tasks.

⁴v 3.12, publicly available at <http://www.summarization.com/mead/>

We note here that since our primary focus is on the base Guided Summarization task and not on the Update Summarization task, we do not use any new techniques for understanding the update process. Instead we just assume that the features arising out of the DL perspective can provide us with enough clues for any novelty. For the Update part, we use the entire set of base and update collections as input to the tag-topic models. Also the number of topics parameter is set to 60, 80, 100 and 120 topics based on roughly twice or thrice the number of actual docsets and also some values in between.

In the official TAC evaluations, the distinction between the better and poorer systems is made based on their mean ROUGE scores w.r.t the best and the worst scoring *human summaries*. Table 5.3 shows the means of the ROUGE-SU4 scores for the summaries obtained from MEAD and their upper and lower bounds of the 95% confidence intervals for each of the Guided Summarization collections – both base and update. Tables 5.4 and 5.5 list the corresponding ROUGE SU-4 scores from the tag-topic models.

TAC 2010A								
Global Models	K=60	Conf.- Int.	K=80	Conf.- Int.	K=100	Conf.- Int.	K=120	Conf.- Int.
TagLDA	0.08885	0.08469- 0.09330	0.09353	0.08906- 0.09797	0.09261	0.08833- 0.09723	0.09231	0.08832- 0.09646
METag ² LDA	0.10793	0.10317- 0.11267	0.10627	0.10130- 0.11128	0.10494	0.10017- 0.10980	0.1071	0.10227- 0.11198
CorrME Tag ² LDA	0.09633	0.09171- 0.10074	0.09673	0.09275- 0.10096	0.09681	0.09268- 0.10084	0.0965	0.09243- 0.10033
TAC 2010B								
Models	K=60	Conf.- Int.	K=80	Conf.- Int.	K=100	Conf.- Int.	K=120	Conf.- Int.
TagLDA	0.08664	0.08308- 0.09044	0.08117	0.07775- 0.08477	0.0836	0.08023- 0.08717	0.08351	0.08008- 0.08727
METag ² LDA	0.09367	0.08959- 0.09801	0.09255	0.08869- 0.09654	0.09338	0.08933- 0.09781	0.0945	0.09046- 0.09878
CorrME Tag ² LDA	0.08909	0.08514- 0.09312	0.08837	0.08465- 0.09192	0.08962	0.08513- 0.09423	0.08963	0.08598- 0.09361

Table 5.4: ROUGE-SU4 scores for TAC 2010A/2010B datasets obtained from sentence ELBO based summarization using tag-topic models. K is the number of topics.

From Table 5.4 we observe that for both TAC 2010A and TAC 2010B, sentence likelihoods from METag²LDA perform the best i.e. at par with the local docset specific MEAD summarizer for the Update collection. METag²LDA beats the MEAD system with statistical significance on the Base collection for the Guided Summarization task on TAC 2010A. This is intuitive since METag²LDA deems a sentence to be a candidate for a summary not only because of the latent topical relevancy of sentential content words but also on the latent topical relevancy of the sentential DL features including possibly the coherence triplets. For the TagLDA model, the latter enforcement is not present. The Corr-METag²LDA model puts a strong constraint on the topic relevancy of a content word through the simultaneous satisfaction of the relevancy across the corresponding DL tags. According to our observations, this constraint usually increases likelihood on held-out test sets but is not favorable for selecting summary sentences as it can put higher weights on words that are topically relevant but ultimately redundant for a summary.

However, from Table 5.5, we see that the performance for even the METag²LDA models drops quite a bit for the TAC 2011A dataset. We believe that firstly the relevant content words in TAC 2011A dataset

TAC 2011A								
Global Models	K=60	Conf.-Int.	K=80	Conf.-Int.	K=100	Conf.-Int.	K=120	Conf.-Int.
TagLDA	0.08956	0.08494-0.09496	0.09031	0.08497-0.09603	0.08914	0.08365-0.09502	0.08685	0.08130-0.09286
MEtag ² LDA	0.10003	0.09382-0.10691	0.10094	0.09490-0.10724	0.10123	0.09556-0.10768	0.10131	0.09501-0.10806
CorrME Tag ² LDA	0.09139	0.08548-0.09783	0.09635	0.09084-0.10248	0.09688	0.09119-0.10313	0.09267	0.08702-0.09880
TAC 2011B								
Models	K=60	Conf.-Int.	K=80	Conf.-Int.	K=100	Conf.-Int.	K=120	Conf.-Int.
TagLDA	0.07979	0.07489-0.08454	0.08251	0.07814-0.08739	0.07874	0.07430-0.08335	0.08114	0.07650-0.08578
MEtag ² LDA	0.09124	0.08545-0.09708	0.09411	0.08864-0.09972	0.08942	0.08370-0.09501	0.09452	0.08867-0.10068
CorrME Tag ² LDA	0.08506	0.07988-0.09062	0.08306	0.07888-0.08717	0.08389	0.07938-0.08817	0.08433	0.08025-0.08820

Table 5.5: ROUGE-SU4 scores for TAC 2011A/2011B datasets obtained from sentence ELBO based summarization using tag-topic models. K is the number of topics.

can be infrequent i.e. multiple concepts may be equally important and it is difficult to thematically prefer one over the other. Secondly, full sentence lengths have played a major role in lowering the ROUGE scores. It is observed that the topic models have favored sentences that are long and provided better posterior fits but failed to include the final long sentences as it was violating the 100-word limit. This has resulted in very short summaries but to be fair to MEAD (run with official TAC settings) as a full sentence extraction system, this needs to be done. It also highlights the need for better sentence compression as well as the importance of local models.

We notice that the ROUGE-SU4 scores across the different values of topics are not enough for selecting the best configuration of a system. The number of topics also do not play a significant role in altering the summaries – in most cases, they just permuted the order of sentences. Thus in all forthcoming summarization experiments with the tag-topic models, we sum the sentence likelihood scores of the central sentences for a particular value of K , say 100, with those obtained from models run with all the lower settings of K (i.e. 60 and 80). This simulates non-parametricity in topic model based scoring to some extent.

5.4 The Local Models

We now briefly describe the intuitively simple and extremely effective summarization models local to the documents in each docset. We recall from Fig. 5.3 that these local models extract features from the documents in each docset and use these features for weighting sentences. Depending on the linguistic assumptions, the most time-consuming part for computing the local models is the syntactic parsing of sentences. However, all of these computations are done offline as the documents are pre-processed. Five main local models are considered for each docset to understand the discriminatory power of the feature sets measured by 5-fold cross-validation accuracies of event category classification of the newswire

documents. Note that none of the feature sets contain standard English stopwords.

5.4.1 Document Set Models—Bags of Key Terms

We first derive two of the feature sets that are used in the DL perspective in the tag-topic models in a docset independent way (see Section 5.3.1). Intuitively if these frequency based features are used in a local way i.e. collected over docsets and show sufficient event discrimination power, then their inclusion as DL tags into the tag-topic models is also justified. Intuitively if these frequency based features are used in a local way i.e. collected over docsets and show sufficient event discrimination power, then their inclusion as DL tags into the tag-topic models is also justified. These feature sets are as follows:

A) Collection of the top 20 words (**doc-corpus-tfidf**) using the $tf \times idf$ weights, where idf (inverse document frequency) has been calculated across the corpus. The set of top 10 words per document in terms of cumulative $tf \times idf$ weights are used to collect this set.

B) Top 5 most frequent words (**doc-frequency**) per document. These two feature sets become local models (as baselines) when terms are restricted only to a docset.

The other local models that we consider for summarization which are based on linguistic assumptions are the part of speech models to extract nouns and verbs; syntactic dependency tree generation and the RS-tree parsing models. All of these models are sentence based models and are used based on the illustration and intuitions mentioned in Section 5.1.2. Our intuitions are further validated through superior event classification performance using the features from the part of speech models. The next set of features are:

1) Collection of the top 20 nouns (**docset-tfif-noun**) including proper nouns using cumulative $tf \times isf$ weights. The isf (inverse sentence frequency) is calculated only for sentences in the documents within the respective docset.

2) Collection of the 5 most frequent verbs (**docset-tfif-verb**) collected across all documents in a docset using cumulative $tf \times isf$ weights.

3) Collection of the top 20 nouns including proper nouns and top 5 verbs (**docset-tfif-noun+verb**) using **(1)** and **(2)** above.

The numbers 5 and 20 are set based on a decision to search for a minimum number whilst achieving a joint event classification accuracy of at least 90% for nouns and 80% for verbs. Next we also considered sentential dependency graphs and RS-trees but these models are not used to verify any event classification performance.

Although a sentential dependency graph are independent of any docset label bias, however it becomes a local model in our scenario due to its dependence on docset specific nouns and verbs while scoring sentences. We also use an unsupervised thresholding technique to select the better sub-sentential spans from the RS-trees based on cosine similarity of the spans to the **query title** of the docset and the **doc-corpus-tfidf** set.

We next discuss event classification performance of several feature sets first and then discuss the use of the syntactic dependency and the RS-tree parses of the sentences in Section

5.4.2 Event Classification Performance

Figure 5.7 shows the predictive power of the local features for document event categorization during a 5-fold cross-validation. As expected the top 5 most frequent verbs from the documents leading upto

a collection of verbs in the docset do not have very high discriminatory power but are not bad either. On the other hand, **docset-tfif-noun+verb** performed consistently high for both datasets. **doc-corpus-tfidf** also performed remarkably well - which is mostly due to the fact that in many cases, it included words from **docset-tfif-noun+verb** as well. Many of the words in **doc-corpus-tfidf** even are the prominent Named Entities that occurred most frequently in the docset. The performance of **doc-frequency** is encouraging as well given that it is the cheapest feature to compute. Note that both **doc-corpus-tfidf** and **doc-frequency** are restricted to a docset as a feature set for event classification. The minimum number of these features that we used for extraction from each document thus proved to be very good for event classification and their inclusion into the DL perspective of our Tag²LDA models to compensate for the sparsity in the coherence triplets that were discussed earlier in Section 5.3.1, seems to be well grounded. The cross-validation graphs in Fig. 5.7 are obtained using the LibSVM Support Vector Machine library [Chang and Lin, 2011] with default settings for multiclass classification. 5.4.3

On the other hand, using features from our extended tag-topic models with the asymmetric Dirichlet prior performs better than the corresponding symmetric case for each class of tag-topic models that do not use the correspondence constraint. The different types of tag-topic models that we considered are TagLDA-Asym, METag²LDA-Asym and CorrMETag²LDA-Asym where “Asym (Asymmetric)” means that the components of α in the models in figure 5.4 can give rise to different levels of sparsity in the latent topics.

The topic model features that we use are the $\gamma_{d,i} - \alpha_i$ for each document d and each topic i . Figures 5.8a and 5.8b show that contrary to better ELBOs on training set, the features from the correspondence class of tag-topic models show very poor predictive power during 5-fold cross-validation for document event classification. We believe that the strong constraints on correspondence constricts the pattern of discriminatory modes and this leads to the poor classification performance whether asymmetric topic proportion priors are used or not.

However, it is interesting to observe that the same features from TagLDA and METag²LDA employing the symmetric priors also show similar poor accuracies. METag²LDA is loosely constrained on the DL perspective and the TagLDA does not consider that perspective at all. The best performance comes for the latter two class of models but employing an asymmetric Dirichlet prior over the topic proportions with TagLDA-As performing slightly better than MPTag²LDA-Asym for the TAC 2011A dataset. Due to these results, we always use the tag-topic models with the asymmetric priors henceforth and drop the “Asym” suffix in further illustrations.

5.4.3 Sentence Dependency Graphs and RS-trees

A dependency graph of a sentence is usually an acyclic graph whose nodes are the words in the sentence and the edges denote syntactic relations. Often the relations convey a semantic meaning – for e.g. in

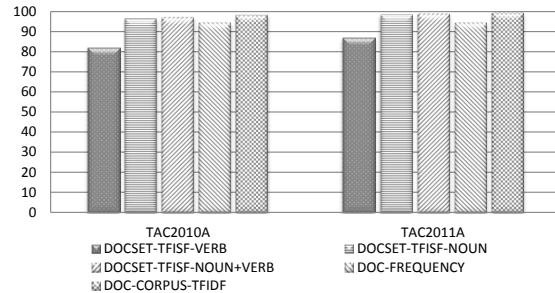


Figure 5.7: Five-fold Cross-validation accuracies of the local models (bags-of-key terms) on event category classification of TAC 2010A/2011A documents. The legend is read from left to right and from top to bottom corresponding to the bar groups for each of the TAC base collections.

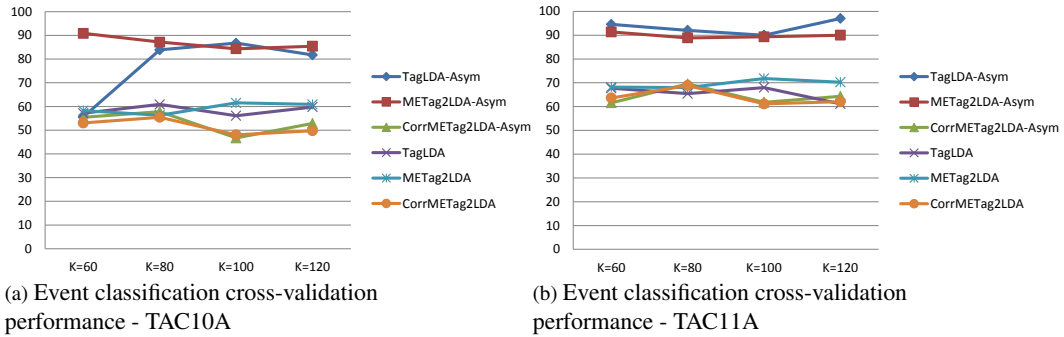


Figure 5.8: Event classification cross-validation accuracies on TAC 2010A and TAC 2011A dataset using per-document latent topic proportions from the tag-topic model as features, i.e. $\gamma_d - \alpha$, for different number of topics - Higher is better. Note that for symmetric prior over θ_{ds} , the vector α becomes the scalar α .

the second sentence shown in Fig. 5.1, “distractions” is the “Nominal Subject” of the word “lessen”; the word “sleep” is the “Open Clausal Complement” of “lessen” and the word “quality” is a “Direct Object” of sleep. In our experiments, these relations are automatically extracted through the Stanford CoreNLP Suite. While scoring sentences with this graph, we first convert this directed graph into an undirected one and find the three shortest paths between the words in the **docset-tfif-noun** feature set and the words in the sentences than contain at least a verb from the **docset-tfif-verb** feature set. For each such path, the path score is simply the number of times a verb from the **docset-tfif-verb** feature set is found as a node. The sentence score is cumulative over the path scores. This becomes our fourth local model, the **docset-dep-graph-noun+verb** model.

Finally we use our fifth local model to be the RS-trees automatically obtained through Rhetorical Structure parsing as mentioned in [Soricut and Marcu, 2003] to select relevant spans as a measure of sentence compression. This model becomes a local model due to the use of the query titles and the **doc-corpus-tfidf** features. We first follow the work in [Marcu, 1999] to score each node of the RS-trees using the propagation of the salient text spans upward to the root of the tree. The spans i.e. the leaves which are promoted up the RS-tree through internal nodes receive a score proportional to maximal heights of such nodes in the tree that contain the promoted spans. In Fig. 5.2, span 1 only gets a score of 1 while spans 2 and 3 get scores of 3. As in [Marcu, 1999], the scores are computed using the recursive scoring formulation using node heights in the RS-tree. These height scores form the “base scores” for the RS-tree spans.

5.4.4 Sentence Compression using RS-tree Spans

Using just the base scores of the RS-tree spans and using those to compress sentences for summarization resulted in very poor summaries. We have found out that many of the salient sentences do not give rise to deep trees and the better spans are thus scored lower. The selection order in this case is based on the descending order of span weights in the sentences. Thus to determine relevance, we next focus on obtaining the cosine similarities of the satellite and the nucleus spans from the RS-trees to the vector space of the docset specific **query title** and the **doc-corpus-tfidf** feature set.

However, this also raises the question of how do we best select a threshold for such cosine similarity

Full sentence: A previous study on sleep deprivation published in 1999 in the Lancet found that less sleep resulted in impaired glucose metabolism, which affects how the body stores and processes glucose for energy.			
[Spans]_{index}	Base Score	Type of Span	Cosine Value
[A previous study on sleep deprivation] ₁	5	Nucleus	0.381000381
[published in 1999 in the Lancet found] ₂	1	Satellite	0.0
[that less sleep resulted in impaired glucose metabolism,] ₃	5	Nucleus	0.207390338
[which affects] ₄	1	Satellite	0.0
[how the body stores and processes glucose for energy.] ₅	2	Nucleus	0.0
Query Title: Sleep Deprivation; doc-corpus-tfidf: Cheri_Mah, Sleep, adolescent, age, brain, cardiovascular, deprivation, disorder, heart, hour, hypertension, impair, increase, nap, night, researcher, risk, sleep, stress, study			

Table 5.6: RS-tree spans and their importance.

values. Table 5.6 shows a sample sentence and its spans obtained through RS-tree parsing, the base score of the spans (nodes), the status of the spans and the cosine similarity of the spans to the feature set. Clearly spans 2 and 4 are not very important; span 5, although having a competitive base score, is not relevant to the feature set. However in some other example sentence, if spans like this has a relevance score of, say, 0.17 then the question arises whether we choose it or not?

To this end, we use the unsupervised density estimation technique using Gaussian kernels [Kvam and Vidakovic, 2007]. Using Gaussian kernels is usually a method of choice since piecewise convolutions of Gaussians can represent functions of arbitrary complexity. The thresholds for the satellites and the nuclei are handled separately. For all density estimation techniques, we first create an array of values that contain strictly positive cosine scores.

ALGORITHM 1: `select_satellite_threshold`

```

1: input: the data array  $D$  (double), percentage of maximum density  $densityPerc$  (double) and  $isBaseCollection$  (boolean)
2: output:  $threshold$ 
3: {The array representation and function calls follow MATLAB's conventions here}
4:  $npoints \leftarrow 100$ ; % This is also the default setting in MATLAB
5:  $[f, xi] \leftarrow ksdensity(D, 'npoints', npoints)$ ; { $f$  is an array of the same size as  $xi$ ;  $f$  holds the densities evaluated at the points in the array  $xi$ }
6:  $[maxf, maxI] \leftarrow \max(f)$ ; %  $maxI$  is the index in the array where  $maxf$  occurs
7:  $index \leftarrow (f < maxf \times densityPerc)$ ; { $index$  becomes an array of booleans}
8:  $newI \leftarrow maxI$ ;
9:  $oldXi \leftarrow xi(maxI)$ ;
10:  $length \leftarrow \text{size}(index, 2)$ ;
11: for  $i1 = 1 \rightarrow length$  do
12:   if  $(index(i1) == 1) \&\& (i1 > maxI)$  then
13:      $newI \leftarrow i1$ ; break;
14:   end if
15: end for
16:  $newXi \leftarrow xi(newI)$ ;
17:  $newD \leftarrow D(D > newXi)$ ;

```



```

18:  $[f, xi] \leftarrow ksdensity(newD, 'npoints', npoints);$ 
19:  $[newmaxf, newmaxI] = \max(f);$ 
20: if isBaseCollection then
21:    $threshold \leftarrow (xi(newmaxI - 3) + xi(newmaxI - 2) + xi(newmaxI - 1))/3;$ 
22: else
23:    $threshold \leftarrow (xi(newmaxI) + xi(newmaxI + 1) + xi(newmaxI + 2) + xi(newmaxI + 3))/4;$ 
24: end if
25:  $threshold \leftarrow \text{round}(c \times threshold)/c;$  {We set  $c$  to be  $10^3$  in our experiments}

```

The density for a scalar value x is obtained as $f_h(x) = (1/(Nh)) \sum_{i=1}^N K'((x - x_i)/h)$, where K' is the kernel function and x_i is a value in the data array D of length N and indexed by i . h is often called the bandwidth and is determined automatically from the data using its standard deviation – if set manually, lower values of h results in spiky graphs and smoother for higher values. In our experiments x ranges from the lowest to the highest x_i s in 100 equally spaced intervals – we used Matlab’s *ksdensity()* function for repeatability purposes.

The blue graphs in Figs. 5.9a and 5.9b show the initial unsupervised density graphs. The blue graphs show that the modes are around 0.08, but choosing this as a threshold for satellite spans will again introduce noise in relevancy calculations as these spans typically reflect some common information. We thus need to look at the behavior of the densities for which the values of the cosine similarities are more than the mode. This amounts to executing Algorithm 1 with the *densityPerc* argument set to some value. We set this value to 0.8 for satellite spans. The value is intuitively set in spirit following the 80% in the 80-20 Pareto rule [Newman, 2005] – we look for contributing satellite spans that account for 80% of the density. We also validate this heuristic by manually inspecting a few spans with higher overlaps to feature sets in a previous newswire collection – TAC2009. However, it can also be set by using ROUGE scores of the summaries formed out of the resulting spans on a development set. If previous/validation datasets are not available, one can manually inspect the spans of some random sample of sentences within the input collection and ascertain an initial threshold. This can be very useful if RS-trees are being used for a previously unknown genre of documents. Intuitively, *densityPerc* identifies the location of the right tail where spans show more similarity to the feature sets.

The red graphs in Figs. 5.9a and 5.9b show a revised density estimate obtained using Algorithm 1. The dotted vertical lines denotes the thresholds obtained from the data arrays D for each dataset scenario. Using the mode information in the blue graph, we truncate the data from the left until we reach the appropriate x_i in the right. We then again perform a density estimate using the truncated data array and identify the new mode. However by doing this we might have over-estimated the relevance and thus we take the average of the rightmost three equidistant points immediately to the left of the new mode. This is intuitive for the “A” timeline or “Base” documents. Since the “B” timeline or “Update” documents deal with both old and new information, we choose the average of the new mode along with the three equidistant points immediately to the right of it as a bias towards novelty. The three points are heuristically chosen based on the significance of the estimated densities to the left and right of the new modes. This restriction can be avoided if we consider much more than 100 samples (say 10,000) to begin with but computational costs begin to increase.

Following our procedure, the thresholds for the cosine similarity values of the satellite spans to the **query title** and the **doc-corpus-tfidf** feature set for the different datasets are obtained as: TAC 2010A –

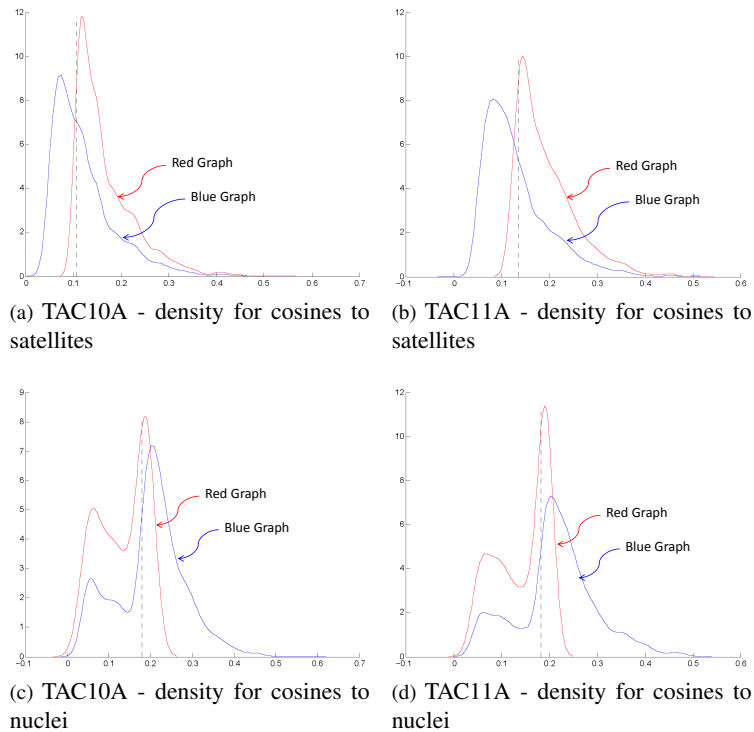


Figure 5.9: Choosing thresholds for selecting RS-tree spans using Parzen density estimates of the cosine similarity values of the words in the spans to those in the feature set.

0.11; TAC 2011A – 0.135; TAC 2010B – 0.12; TAC 2011B – 0.125. In a supervised setting, it is possible to select such thresholds using training data that contains human summaries but we do not attempt an easier supervised framework where the principles of “Guided Summarization” are difficult to incorporate without even more annotation.

To select the thresholds for the nuclei spans, we first create a data vector with values from cosine similarities with the nuclei spans as well as the Root spans – the Root spans can be justified as one long “nucleus” span with no subordinates. Using Algorithm 1 causes a problem due to very large and thick tails and much lower overall densities as compared to the ones for the satellites. The mode, and initially there is only one, turns out to be just 0.05 which leads to much noise in the selection of the nucleus spans as before. This arises due to larger lengths of the Root spans and hence their lower cosine scores. Using the knowledge of the satellite thresholds, it is safe to assume that the cosine scores for the nuclei spans must be greater than those for the satellites. We thus multiply the initial mode of 0.05 by consecutive integers until we cross the threshold for the satellites. This turns out to be 0.15 for all the datasets. Doing this also penalizes selection of longer root spans containing minimally relevant information.

We next perform a primary mode finding for all of the data points having value greater than 0.15 and further truncate the data based on the new secondary modes. The blue graphs in Figs. 5.9c and 5.9d show the density graphs of the data arrays after the secondary truncation. Observing the bi-modal density graphs (blue) of the truncated data, we postulate that a value to the left of the mode should be chosen so as to guard against possible over-estimation. Thus, using the doubly truncated data, we just change line

17 of Algorithm 1 with $D1 \leftarrow D(D < newXi)$. Using this change, we can call Algorithm 1 using a *densityPerc* value set to 1 since the data has already been doubly truncated. Using these modifications, the thresholds for the cosine values of the nuclei spans to the **query title** and the **doc-corpus-tfidf** feature set for the different datasets are obtained as follows: TAC 2010A – 0.18; TAC 2011A – 0.18; TAC 2010B – 0.19; TAC 2011B – 0.1925.

In order to compress sentences through spans, we started out with an empty sentence string and apply the following steps by iterating over the spans of a RS-tree from left to right:

- If no nucleus has been found so far but a satellite exists whose relation can even be outside of the set of chosen relations, we add that span given that its cosine score is above the threshold for satellites for the dataset being summarized. Most of these spans occurred in sentences which started off with “Background” spans and then had a nucleus in between.
- If the number of nuclei found so far is ≤ 2 and the current span is a satellite whose cosine score is above the threshold and is related to a nucleus through the relations mentioned in Section 5.1.3, then it is added to the sentence. We have observed that a sentence with more than two nuclei and satellites succeeding them is often a very long sentence.
- If the span is a nucleus and its cosine score is above the threshold for nuclei for the dataset being summarized, it is added. Often this is the first span being added in the sentence.
- If a span is neither a nucleus or a satellite but just a single root node acting as a surrogate for a nucleus span, then it is added irrespective of its cosine score. Generally it has been seen that shorter root spans are summary worthy.

Cue words at the beginning of the spans can be removed but doing so in a principled way and combining different spans within a sentence or across sentences lead to the field of paraphrase generation and we intend to explore that direction as part of future research. In our case, it is mostly the appositives which are eliminated from the best summary sentences — a beneficial side-effect of using RS-tree based sentence compaction.

5.5 Summarization Experiments

In this section, we briefly describe the methods we use to obtain the summaries for each docset based on the sentence weights obtained from both the tag-topic models and/or the local models. We also lay down several baselines and compare results.

5.5.1 Basic Summarization Algorithms

Here we highlight the sentence scoring strategies based on the global and local models. We first discuss the strategies for the global tag-topic models:

G1) We use the likelihood values corresponding to each central sentence that is fit to the trained models. These values are obtained following the equations for \mathcal{L}_{ME} and \mathcal{L}_{CorrME} in section 5.3.2. The equivalent \mathcal{L}_{TagLDA} can be found in [Zhu et al., 2006] and is not repeated here.

G2) The sentence weights are obtained by using the expression $\sum_{w_{Q,m}} E_q[\log p(w_{Q,m}|z_{Q,m}, \theta_s)]$ where $w_{Q,m}$ is a word in the query title (average 2-3 words) that is also in the vocabulary V which is input to the models; s is the current sentence whose DL perspective depends on adjacent sentences.

G3) Same as **(G2)** above, except that the summation is over all words in the sentence s that are also in V .

The first type of weighting has a purely probabilistic interpretation, but the second and third follows from [Nenkova et al., 2006a] and is less intuitive probabilistically.

Next we highlight sentence scoring strategies based on the local models: (Note that when the RS-tree spans are used to generate compressed sentences, we use the RS-tree span selection criteria (see Section 5.4.4) only to concatenate the relevant spans into items of a bullet list summary.)

L1) The score of a sentence is determined by the base score and the cosine scores of the spans (leaf nodes) of the RS-tree to the feature set but respecting the cut-offs. We name this method **docset-tfidf-RSTree**

L2) The lemmatized words in a sentence with their counts are treated as a list and its cosine is measured against the **docset-tfist-noun** feature set, again treated as a list. This method is named as **docset-noun-RSTree**

L4) The lemmatized words in a sentence with their counts are treated as a list and its cosine is measured against the **docset-tfist-verb** feature set, again treated as a list. This method is named as **docset-verb-RSTree**

L4) The lemmatized words in a sentence with their counts are treated as a list and its cosine is measured against the **docset-tfist-noun+verb** feature set, again treated as a list. This method is named as **docset-noun+verb-RSTree**

We also used the version of these baseline methods but without considering any RS-tree based compression and use full sentences instead. For those cases, we replace the “RSTree” suffix with the suffix “FS” denoting “full sentence”. For **docset-tfidf-FS**, we score the sentences based on the weights of the spans as in **docset-tfidf-RSTree**, but we use the full sentences while generating a summary.

Redundancy is handled by adding new sentences or RS-tree span sequences that do not share some percentage of the unigrams or/and bigrams in the set of summary sentences previously added. The percentage was set to 50% for the base collection and 40% for the update collection based on the means and variances of the sentence token overlap percentages in the model summaries of previous years’ datasets. This is the only instance where we obtain statistics from the human summaries. We have tried the MMR strategy [Carbonell and Goldstein, 1998] of ordering sentences with score $s_{score} = similarity(q, s_i) - redundancy(s_i, s_j) \forall \{s_i, s_j\} \in \{s_in_docset\}$, where q is the query and s is a central sentence, but it proved worse since the similarity and redundancy scores are not calculated in a homogeneous way. In all of our experiments we eliminated all sentences from the summary having ≤ 10 tokens and ≥ 25 tokens. These limits are also set using the first and second order statistics of the human summaries from previous years’ datasets. Subjective sentences that start with pronouns enclosed in quotes are not considered. Sentences with more than 4 numbers - suggestive of a table row or a list of results are eliminated too. These heuristics apply to the sentences created from spans as well.

Finally we consider our proposed method of using sentence weights from both the global tag-topic models and some selected local models which is aided by the use of sentence compression using RS-trees. We discuss them in Sections 5.5.3 and 5.5.4.

In all our experiments we order the full sentences (or a sequence of RS-tree spans) in the descending order of the weights assigned to them. All scores from individual models are normalized between [0,1]. When used in combination, the normalized scores from each model are combined and the final scores are normalized again.

5.5.2 Evaluation Settings

Due to the lack of resources for manual evaluation, we only use ROUGE as the standard automatic summary evaluation toolkit. ROUGE uses a Wilcoxon test to establish confidence intervals from the mean scores obtained through bootstrap sampling. The tools that we use are the RS-tree parser implementation that is used in [Soricut and Marcu, 2003], the Stanford CoreNLP toolkit and our own implementations of the tag-topic models in [Das et al., 2011].

Apart from the local models acting as baselines, two other standard baselines are chosen. The Baseline-Naive simply returns all the leading sentences (up to 100 words) in the most recent newswire documents – this is a very strong baseline particularly for the update collections. The Centroid [Radev et al., 2004] baseline is output of MEAD automatic summarizer. Official summarization scores from a very competitive peer system named CLASSY [Conroy et al., 2010, Conroy et al., 2011] are also chosen for comparison. Over the years at the TAC summarization competitions, CLASSY has had been continuously updated and fine tuned based on training data from previous year’s. For e.g., for the TAC 2011 dataset, it uses a very finely crafted vocabulary reflecting the categorical aspects of the Guided Summarization task. The TopicMarks baseline is obtained from a recent commercial summarization service⁵. Topicmarks summarizes multiple documents by treating all documents as one large document. It does not depend on any query and tries to generate key concepts which are fairly close to the topic titles.

5.5.3 Results

In this section we compare and analyze the summarization performances of the different models.

5.5.3.1 PERFORMANCE OF BASELINE MODELS ON TAC 2010A AND 2011A DATASETS

Table 5.7 shows the ROUGE skip-4-bigram (ROUGE-SU4) and the ROUGE-bigram (ROUGE-2) scores of the summaries from the local baseline models which do not employ any topic model. The scores are reported for both TAC 2010A (left columns of Table 5.7) and the TAC 2011A (right columns of Table 5.7) datasets. Clearly we see a significant increase in ROUGE scores when sentence compression is achieved through salient spans from RS-trees. For the “full sentence” settings for our baseline models, we allow summaries to be longer than 100 words but truncate the summary at the 100 word limit without regards to sentence completion. This does not affect ROUGE evaluation but incomplete final sentences may easily affect readability – human evaluations w.r.t. responsiveness are not considered in this article. Owing to the success in using RS-trees for sentence compression, we always perform sentence compression in the summaries obtained using the tag-topic models in the forthcoming comparisons.

We observe from Table 5.7 that the summaries obtained from the official CLASSY system for TAC 2010A [Conroy et al., 2010] are at par with those obtained from the local models **Docset-Tfidf-RSTree** and **Docset-Noun+Verb-RSTree**. Hence we can safely assume that our baseline local models are quite competitive given that the relevance determination of the RS-tree spans is unsupervised. However, as mentioned in [Conroy et al., 2011], the CLASSY system has been updated with many more adjustments, both automatic and manual, and thus performed better than our baseline models with compression for

⁵topicmarks has been acquired by <http://www.tagged.com>

<i>Baseline Models</i>	TAC 2010A				TAC 2011A			
	ROUGE-SU4	Conf.-Int.	ROUGE-2	Conf.-Int.	ROUGE-SU4	Conf.-Int.	ROUGE-2	Conf.-Int.
Docset-Tfidf-RSTree	0.12556	0.12234-0.13258	0.08804	0.08620-0.09233	0.14146	0.13620-0.14635	0.1034	0.09743-0.10925
Docset-Noun-RSTree	0.12304	0.11861-0.12728	0.08361	0.07848-0.08901	0.14342	0.13713-0.14954	0.10499	0.09846-0.11143
Docset-Verb-RSTree	0.12457	0.11897-0.13013	0.08792	0.08149-0.09444	0.13905	0.13417-0.14375	0.09973	0.09441-0.10537
Docset-Noun+Verb-RSTree	0.12645	0.12170-0.13086	0.08799	0.08343-0.09102	0.14584	0.14020-0.15169	0.1090	0.10064-0.11406
Docset-Tfidf-FS	0.11201	0.10698-0.11669	0.07549	0.07003-0.08064	0.1343	0.12889-0.13950	0.09303	0.08685-0.09908
Docset-Noun-FS	0.11478	0.11012-0.11956	0.07456	0.06951-0.07993	0.1427	0.13702-0.14874	0.10095	0.09442-0.10828
Docset-Verb-FS	0.10393	0.09945-0.10854	0.06254	0.05727-0.06766	0.12203	0.11719-0.12710	0.08095	0.07538-0.08724
Docset-Noun+Verb-FS	0.11752	0.11287-0.12239	0.07783	0.07227-0.08336	0.14616	0.14059-0.15217	0.10518	0.09882-0.11210
Centroid	0.09117	0.08676-0.09586	0.05929	0.05453-0.06417	0.11741	0.11141-0.12350	0.08672	0.08013-0.09347
Baseline-Naive	0.08565	0.08071-0.09033	0.05386	0.04846-0.05920	0.09927	0.09368-0.10517	0.06399	0.05780-0.07080
Topic-marks	0.11524	0.11050-0.11994	0.07831	0.07260-0.08380	0.11976	0.10818-0.13166	0.08351	0.07079-0.09696
CLASSY	0.12258	0.11783-0.12749	0.08554	0.07956-0.09160	0.15812	0.15089-0.16513	0.1278	0.11947-0.13637
Human-Highest	0.16294	0.14759-0.17574	0.12862	0.11087-0.14681	0.16412	0.14974-0.17767	0.1282	0.12686-0.14216
Human-Mid1	0.15289	0.13691-0.16798	0.11695	0.09822-0.13447	0.15731	0.14284-0.17129	0.11502	0.09672-0.13220
Human-Mid2	0.14637	0.13212-0.16064	0.11313	0.09197-0.13365	0.1492	0.13697-0.16171	0.11146	0.09902-0.12486
Human-Lowest	0.13805	0.12685-0.14972	0.09623	0.08074-0.11167	0.1462	0.13294-0.16057	0.10944	0.09014-0.12833

Table 5.7: ROUGE-SU4 and ROUGE-2 scores for summaries from baselines and human summarizers for TAC 2010A and TAC 2011A base collections.

the TAC 2011A dataset. CLASSY also employs a supervised approach by incorporating human summaries from developments sets to fine tune the system. On the other hand, we are truly surprised as to how an extremely simple local model, **Docset-Verb-RSTree** can perform so well with RS-tree based compression on the TAC 2010A dataset.

The TAC 2010 and the TAC 2011 datasets also have eight 100-word human summaries for each docset for the systems to compare with. The human summaries are measured against each other by using subsets of other human summaries and averaging over them. We thus also report the scores of the highest scoring and the lowest scoring human summaries along with the two medians. As is the case with previous years' datasets, the human summaries have always performed better. This is true for the 2010 dataset as well. However, in the 2011 dataset, this trend does not hold good. While the human annotators are free to choose any word to construct their summaries and that the words did not have to belong to the input documents, it seems that many of the TAC 2011A docsets have many relevant concepts and it is difficult for even some of the human annotators to generate sentences that cover all the right concepts within the 100 word limit.

Table 5.8 shows the ROUGE-SU4 and the ROUGE-2 scores of the summaries obtained from just the tag-topic models but using RS-tree span relevancy based sentence compression. The confidence intervals are suppressed here to save space. Clearly for the TAC 2010A dataset, the weighting of sentences using sentence likelihoods (**G1**) and using the cumulative probability mass of the query words (**G2**) show best results and even parallel some of the best baselines. Using (**G3**), however, leads to topic drift. In general the mean ROUGE scores of the summaries from the TagLDA and Tag²LDA class of models that do not use the correspondence constraint are higher. For TAC 2011A, METag²LDA shows slightly superior performance as far as sentence likelihood weights are concerned. With individual word weighting, the TagLDA model shows slightly better performance due to better sparsity in topic inference owing to the complete lack of correspondence. However, even TagLDA with (**G2**) under-performs the **Docset-Tfidf-RSTree** local model for the TAC 2011A dataset – though this is understandable since the tag-topic models are not docset specific and event correlations through latent topics weighs down the specificity of keyterms for a single docset. Nevertheless, with RS-tree based sentence compression, global tag-topic models can produce summaries that are competitive with local models using the same compression.

5.5.3.2 PROPOSED MODEL PERFORMANCE ON TAC 2010A DATASET

In this section and the next, we discuss summarization results from our proposed models. We use the simplest possible aggregation technique to score a central sentence which has been assigned different weights by different models – both local and global. The final score of a sentence is just the sum of the scores of that sentence assigned by the different models we choose. This scheme is simple and intuitive and reflects a weighted voting mechanism similar in spirit to the algorithm in [Lee, 1997], which is quite effective in practice. Rank fusion is a separate research area by itself and can indeed be a direction for future research within the summarization community.

Our proposed model consists of aggregating the scores of a sentence obtained from 1) sentence likelihoods from the global tag-topic models, 2) the cosine similarity of the sentence to the **docset-tfidf-noun+verb** feature set, 3) the cosine similarity of the sentence to the **doc-corpus-tfidf** feature set and 4) the **docset-dep-graph-noun+verb** model (Section 5.4.3). RS-tree based sentence compression is used by default.

<i>Topic-Models (Sentence Scoring)</i>	TAC 2010A							
	ROUGE-SU4				ROUGE-2			
	K=60	K=80	K=100	K=120	K=60	K=80	K=100	K=120
TagLDA(G1)	0.12482	0.12401	0.12359	0.1234	0.08866	0.0873	0.08698	0.08764
ME ² Tag ² LDA(G1)	0.12331	0.12303	0.12149	0.1247	0.08661	0.08653	0.08642	0.08843
CorrME Tag ² LDA(G1)	0.12368	0.12161	0.1223	0.12359	0.08845	0.08535	0.08793	0.0858
TagLDA(G2)	0.12697	0.12848	0.12812	0.12751	0.09335	0.09444	0.094	0.09351
ME ² Tag ² LDA(G2)	0.12717	0.12461	0.12687	0.1278	0.09197	0.08814	0.09174	0.09355
CorrME Tag ² LDA(G2)	0.12552	0.1214	0.12391	0.11966	0.08881	0.08375	0.08685	0.0832
TagLDA(G3)	0.12524	0.12591	0.12581	0.12541	0.09022	0.09044	0.09136	0.09024
ME ² Tag ² LDA(G3)	0.12553	0.12623	0.12426	0.12655	0.08891	0.09034	0.08844	0.09057
CorrME Tag ² LDA(G3)	0.11942	0.11915	0.12038	0.11916	0.08307	0.08407	0.0844	0.08343
<i>Topic-Models (Sentence Scoring)</i>	TAC 2011A							
	ROUGE-SU4				ROUGE-2			
	K=60	K=80	K=100	K=120	K=60	K=80	K=100	K=120
TagLDA(G1)	0.10976	0.11082	0.11143	0.11374	0.06978	0.07043	0.07191	0.07289
ME ² Tag ² LDA(G1)	0.11985	0.12004	0.12425	0.12303	0.07674	0.07845	0.08242	0.08051
CorrME Tag ² LDA(G1)	0.11358	0.11507	0.11493	0.11829	0.07044	0.07164	0.07271	0.07551
TagLDA(G2)	0.14109	0.14108	0.14053	0.14111	0.1002	0.09989	0.09956	0.10044
ME ² Tag ² LDA(G2)	0.1392	0.13437	0.12926	0.13161	0.09748	0.09335	0.08662	0.09047
CorrME Tag ² LDA(G2)	0.126	0.12888	0.1245	0.12491	0.08213	0.08675	0.08241	0.08258
TagLDA(G3)	0.12469	0.12309	0.12272	0.12211	0.08476	0.08288	0.08174	0.08275
ME ² Tag ² LDA(G3)	0.11544	0.11572	0.11247	0.11709	0.07286	0.07471	0.06904	0.07576
CorrME Tag ² LDA(G3)	0.11399	0.11002	0.10828	0.10983	0.06854	0.06859	0.06513	0.06504

Table 5.8: ROUGE-SU4 and ROUGE-2 scores for summaries from topic model baselines for TAC 2010A and TAC 2011A base collections.

TAC 2010A – ROUGE-SU4 scores									
<i>Human Summaries</i>	Human-Highest	Conf.-Int.	Human-Mid1	Conf.-Int.	Human-Mid2	Conf.-Int.	Human-Lowest	Conf.-Int.	
		0.16294	0.14759-0.17574	0.15289	0.13691-0.16798	0.14637	0.13212-0.16064	0.13805	0.12685-0.14972
<i>Proposed Models</i>	K={60}	Conf.-Int.	K={60,80}	Conf.-Int.	K={60,80,100}	Conf.-Int.	K={60,80,100,120}	Conf.-Int.	
TagLDA+ local models	0.13182	0.12625-0.13739	0.13176	0.12621-0.13737	0.1318	0.12634-0.13729	0.13167	0.12622-0.13711	
METag ² LDA+ local models	0.13118	0.12578-0.13664	0.13114	0.12571-0.13668	0.13122	0.12581-0.13672	0.13122	0.12581-0.13672	
CorrME Tag ² LDA+ local models	0.1312	0.12573-0.13685	0.13092	0.12543-0.13647	0.13093	0.12545-0.13648	0.13122	0.12574-0.13686	
TAC 2010A – ROUGE-2 scores									
<i>Human Summaries</i>	Human-Highest	Conf.-Int.	Human-Mid1	Conf.-Int.	Human-Mid2	Conf.-Int.	Human-Lowest	Conf.-Int.	
		0.12862	0.11087-0.14681	0.11695	0.09822-0.13447	0.11313	0.09197-0.13365	0.09623	0.08074-0.11167
<i>Proposed Models</i>	K={60}	Conf.-Int.	K={60,80}	Conf.-Int.	K={60,80,100}	Conf.-Int.	K={60,80,100,120}	Conf.-Int.	
TagLDA+ local models	0.0952	0.08856-0.10201	0.09509	0.08841-0.10193	0.09514	0.08848-0.10199	0.09525	0.08848-0.10212	
METag ² LDA+ local models	0.09451	0.08790-0.10106	0.09457	0.08802-0.10112	0.09463	0.08806-0.10115	0.09463	0.08806-0.10115	
CorrME Tag ² LDA+ local models	0.09468	0.08817-0.10128	0.09463	0.08811-0.10123	0.09463	0.08811-0.10123	0.09468	0.08817-0.10128	

Table 5.9: ROUGE-SU4 and ROUGE-2 scores from our proposed models for TAC 2010A.

The choices of most of the local models are intuitive. The **docset-noun+verb** model is being reinforced by more linguistic assumptions through the **docset-dep-graph-noun+verb** model. The **doc-corpus-tfidf** model is similarly reinforced by the span selection RS-tree based sentence compression model. The only thing that makes our proposed summarization query dependent is the inclusion of the **query titles** in the feature set for determining the threshold for RS-tree span relevancy. If we remove this constraint, our summarizer becomes query independent but the thresholds need to be determined accordingly based on the techniques mentioned in Section 5.4.4.

In Tables 5.9, 5.10, 5.13 and 5.14 the headings for the topic models that read as $K = \{60, 80, 100, 120\}$ mean that the current model has been run for $K = 120$ topics but sentence weights from the tag-topic models run for $K = 60, 80$ and 100 has also been summed up.

Table 5.9 shows that our proposed method of combining the summary-worthiness of the sentences from both global tag-topic models and local docset-specific feature selection models significantly outperform the baselines and is within the lower bound ROUGE scores of the lowest scoring human summarizer and very close to the lower bound of the lower median as well.

5.5.3.3 PROPOSED MODEL PERFORMANCE ON TAC 2011A DATASET

Table 5.10 also shows that our proposed method of combining sentence evidences works best. Further, the ROUGE-SU4 scores from our proposed method are very close to the upper median scores of the human summaries and is statistically not different than the improved CLASSY system for TAC 2011 [Conroy et al., 2011]. Unlike TAC 2010A, the METag²LDA model does slightly better for the TAC 2011A dataset. For TAC 2011, the CLASSY system is modified to use bigrams, more categorical aspect specific vocabulary and the feature weights tuned against human summaries from previous collections. Our system do not use any hand crafted vocabulary for aspect matching and is based on the intuitions of a reader’s behavior.

TAC 2011A – ROUGE-SU4 scores									
<i>Human Summaries</i>	Human-Highest	Conf.-Int.	Human-Mid1	Conf.-Int.	Human-Mid2	Conf.-Int.	Human-Lowest	Conf.-Int.	
	0.16412	0.14974-0.17767	0.15731	0.14284-0.17129	0.1492	0.13697-0.16171	0.1462	0.13294-0.16057	
<i>Proposed Models</i>	K={60}	Conf.-Int.	K={60, 80}	Conf.-Int.	K={60, 80,100}	Conf.-Int.	K={60, 80,100, 120}	Conf.-Int.	
TagLDA+ local models	0.15319	0.14706-0.15947	0.1529	0.14692-0.15916	0.15289	0.14690-0.15918	0.15288	0.14689-0.15916	
METag ² LDA+ local models	0.153	0.14695-0.15913	0.15382	0.14758-0.15991	0.1537	0.14744-0.15977	0.15367	0.14742-0.15976	
CorrME Tag ² LDA+ local models	0.15342	0.14738-0.15962	0.15328	0.14723-0.15947	0.15317	0.14718-0.15937	0.15312	0.14714-0.15934	
TAC 2011A – ROUGE-2 scores									
<i>Human Summaries</i>	Human-Highest	Conf.-Int.	Human-Mid1	Conf.-Int.	Human-Mid2	Conf.-Int.	Human-Lowest	Conf.-Int.	
	0.1282	0.12686-0.14216	0.11502	0.09672-0.13220	0.11146	0.09902-0.12486	0.10944	0.09014-0.12833	
<i>Proposed Models</i>	K={60}	Conf.-Int.	K={60, 80}	Conf.-Int.	K={60, 80,100}	Conf.-Int.	K={60, 80,100, 120}	Conf.-Int.	
TagLDA+ local models	0.11646	0.10952-0.12352	0.11642	0.10951-0.12347	0.11642	0.10951-0.12347	0.11642	0.10951-0.12347	
METag ² LDA+ local models	0.11647	0.10959-0.12355	0.11764	0.11063-0.12488	0.11753	0.11042-0.12482	0.11753	0.11042-0.12482	
CorrME Tag ² LDA+ local models	0.11669	0.10979-0.12360	0.11664	0.10973-0.12353	0.11652	0.10954-0.12345	0.11658	0.10960-0.12347	

Table 5.10: ROUGE-SU4 and ROUGE-2 scores from our proposed models for TAC 2011A.

Careful observation of the CLASSY system scores from Table 5.7 show the larger variance of the ROUGE-SU4 scores for the TAC 2011A dataset. This means that for some docsets, the system did exceedingly well and for some others it did sufficiently worse. We believe, too much tuning increases unpredictability of summarization performance for unknown test sets. Also another interesting observation is that the ROUGE scores for both the Centroid and the CLASSY systems increase by at least 2% for

the TAC 2011A datasets over TAC 2010A. This has been observed for all systems that competed in both TAC 2010 and 2011 Guided Summarization tasks. We believe that the term distributions, particularly those that give rise to subordinate information in sentences are deemed important by some of the human summarizers. This is also validated by the thicker tails of the cosine similarity scores of the satellite spans to the feature sets as mentioned in Section 5.4.4. Indeed, the cut-offs obtained from Algorithm 1 using the red graphs as shown in Figs. 5.9a and 5.9b can possibly shed some light on the 2% increase in the ROUGE scores of the system summaries.

In our experiments we have observed that the sentence weighting schemes (G2) and (G3) from the tag-topic models do not help in the fusion of sentence scores from the local models – it is possible that such fusions are competitive rather than collaborative. A language independent version of our system can be built using **doc-corpus-tfidf** feature set and using positional information at word level (or other markup tags) and (optionally) a DL perspective as well. The differences in the likelihood contributions of the tag-topic models (c.f Table 5.8) are compensated by relative ordering.

However, given a choice on multi-document summarization, it is better to use TagLDA or METag²LDA using asymmetric Dirichlet priors on topic proportions based on our construction of DL perspective. The event classification power of the tag-topic models is a major indicator for the choice of global topic models to use in extractive multi-document summarization. The correspondence class of models show superior topic inference, however, if the same inference draw a topically relevant but less summary worthy term, then more such terms are selected due to the soft probabilistic constraints between the content words and DL perspective. This results in the selection of more topically relevant but redundant sentences from the correspondence models.

The decision to aggregate the summary worthiness of the sentences from topic models with lower settings of the number of topics also does not hinder summarization performance. In general, it is difficult to identify a correct value of K based on perplexity or likelihood measurements that can guarantee the best multi-document summarization performance also.

The use of RS-tree spans using a thresholding criteria allows us to use the spans as bullet lists and pack more information in less words. An example summary from our proposed system is shown in table 5.11 for the “Sleep Deprivation” topic in TAC 2011A. The system summaries in 5.11 actually have shown low ROUGE scores since it is much harder to assess which fact could really be given more weight in the summary. This has been seen to be true in general for all system summaries in the TAC 2011A dataset for the “Health and Safety” event category.

If we observe the human summary in Table 5.11, we can easily make out the level of artistry involved in intelligently “cut-pasting” the information in the input documents to create a seamless 100 word summary packed with the right information need. Although RS-tree spans compress sentences into items of a bullet list, however, it cannot merge two or more items into one without loss of readability. We want to pursue this direction as a future research by using tools like SimpleNLG [Gatt and Reiter, 2009] as used in [Genest and Lapalme, 2011]. In this article we do post-process the spans to correct the parse structure of the sub-sequences when an intermediate spans fails to get included - for e.g., the second bullet in the middle column of Table 5.11 show a compression with poor grammatical quality – “who slept less than or equal to 5 hours a night were twice as likely to suffer from hypertension than women.” The actual sentence is “The researchers found that those women in the study group who slept less than or equal to 5 hours a night were twice as likely to suffer from hypertension than women who slept for the recommended seven hours or more a night.” Clearly the initial phrase “who slept” and the

Best full sentence human summary for docset D1127E-A w.r.t. ROUGE-SU4 scores	bullet list summary from METag ² LDA for {60, 80, 100}topics with sentence ELBO aggregation + local models	Full sentence summary from METag ² LDA for {60, 80, 100}topics with sentence ELBO aggregation + local models
<p>[Research has found that sleep deprivation is associated with serious health problems such as depression, obesity, cardiovascular disease and diabetes.] [Lack of sleep adversely affects memory function and athletic performance.] [Sleep disorders are common in people 60 and over.] [Women’s health is at much greater risk than men’s.] [Sleep-deprived adolescents are more likely to use alcohol and tobacco.] [Sleep-deprived children can exhibit ADHD-like behavioral problems.] [Sleep medications are increasingly prescribed for children, but their safety and effectiveness are unknown.] [For adults, napping has rejuvenating effects and boosts alertness, performance & productivity.] [Other treatment options include meditation, exercise or evening activity.]</p>	<ul style="list-style-type: none"> • A previous study on sleep deprivation that less sleep resulted in impaired glucose metabolism. • who slept less than or equal to 5 hours a night were twice as likely to suffer from hypertension than women. (*) • children ages 3 to 5 years get 11-13 hours of sleep per night. • Chronic sleep deprivation can do more it can also stress your heart. • sleeping less than eight hours at night, frequent nightmares and difficulty initiating sleep were significantly associated with drinking. • A single night of sleep deprivation can limit the consolidation of memory the next day. • women’s health is much more at risk. (*) 	<p>[Naps that are too long or taken too late in the day, however, affect the quality of nighttime sleep, so proper planning is important. (*)]</p> <p>[The study found that a sleep deficit built up over just five nights can significantly impair heart function.]</p> <p>[Generally, a 20-to-30-minute nap is enough time to reap the benefits of increased alertness and performance and improved mood. (*)]</p> <p>[A study at NASA on sleep-deprived military pilots and astronauts showed that taking a 40-minute nap improved performance by 34 percent and alertness 100 percent]</p> <p>[The researchers found no difference between men sleeping less than 5 hours and those sleeping 7 hours.]</p>
<p>Summary from the CLASSY system: [Chronic sleep deprivation can do more than leave you short-tempered: it can stress your heart and raise your risk of cardiovascular disease and death. (*)] [A separate study released in June by researchers at the University of Pennsylvania found that chronic sleep deprivation adds stress to the heart, putting a person at greater risk of cardiovascular disease and death. (*)] [Women’s health is much more at risk from sleep deprivation than men’s.] [Sleep loss or disturbed sleep can heighten the risk for adolescents to take up smoking and drinking.] [Neither their safety nor effectiveness has been studied in young people.]</p>		

Table 5.11: 100-word summaries for the harder information need on “Sleep Deprivation” in TAC 2011A dataset. Individual sentences are square bracketed. A (*) indicates that the bullets or sentences belong to the same document. Notice how the CLASSY summary is drawn towards a “cardiovascular” bias while our full sentence summary is drawn towards a “napping” bias. Incidentally, “nap” has a strong focus in D1127E-A.

final phrase “than women” could have been cleverly combined into “women sleeping.” By doing several of these in a principled way it might be possible to achieve the ultimate level of human compression as can be seen in the first column of Table 5.11. However, this does not pose much concern in readability issues when summaries are presented in an interactive application since important parts of a sentence can be highlighted and the full sentence expanded based on user input.

<i>Baseline Models</i>	TAC 2010B				TAC 2011B			
	ROUGE-SU4	Conf.-Int.	ROUGE-2	Conf.-Int.	ROUGE-SU4	Conf.-Int.	ROUGE-2	Conf.-Int.
Docset-Tfidf-RSTree	0.1127	0.10845-0.11730	0.07254	0.06772-0.07776	0.12222	0.11857-0.12612	0.08053	0.07573-0.08513
Docset-Noun-RSTree	0.11008	0.10631-0.11403	0.06996	0.06521-0.07439	0.12127	0.11701-0.12562	0.0789	0.07401-0.08348
Docset-Verb-RSTree	0.10412	0.09956-0.10862	0.06146	0.05705-0.06615	0.11874	0.11352-0.12423	0.07628	0.07051-0.08279
Docset-Noun+Verb-RSTree	0.11006	0.10605-0.11431	0.06976	0.06502-0.07461	0.12738	0.12281-0.132	0.08672	0.08077-0.09265
Centroid	0.09645	0.09181-0.10104	0.06238	0.05761-0.06732	0.09147	0.08764-0.09528	0.05925	0.05482-0.06380
Baseline-Naive	0.08817	0.08383-0.09243	0.05313	0.04852-0.05757	0.09479	0.09027-0.09986	0.05718	0.05204-0.06291
Peer-1st	0.11979	0.11540-0.12440	0.07993	0.07473-0.08512	0.13086	0.12505-0.13663	0.09589	0.08942-0.10290
Peer-2nd	0.11869	0.11420-0.12340	0.07902	0.07403-0.08419	0.12817	0.12236-0.13407	0.09244	0.08570-0.09926
Peer-3rd	0.11189	0.10752-0.11621	0.07292	0.06822-0.07793	0.12803	0.12285-0.13299	0.08891	0.08176-0.09590
DualSumm	–	–	–	–	0.1285	–	0.0924	–
CLASSY	same as peer-2nd				0.1274	0.12165-0.13342	0.09244	0.08570-0.09926

Table 5.12: ROUGE-SU4 and ROUGE-2 scores of summaries from local model baselines and top performing peer systems for TAC 2010B and 2011B update collections.

5.5.4 Performance on Update Summarization

Table 5.12 shows the performance of the our local baseline models on the TAC 2010B and TAC 2011B datasets i.e. the update collections. We re-iterate here that we do not tackle the actual process of the update in a strict sense and rely on the power of local docset specific features and some heuristics in the relevancy determination of RS-tree spans in sentence compression of update summaries. We also show the ROUGE scores from the top 3 Peer systems and the DualSumm system [Delort and Alfonseca, 2012] for TAC 2011B Update Summarization task.

For TAC 2010B, the **Docset-Tfidf-RSTree** baseline model scores best and is marginally ahead of Peer-3 but is statistically worse than CLASSY. For TAC 2011B, **Docset-Noun+Verb-RSTree** performs best paralleling the CLASSY system for TAC 2011B. The ROUGE-2 scores become lower possibly due to abrupt removal of intermediate spans. We will address this issue in a future work in an effort to achieve near human paraphrasing.

The official scores of the DualSumm system has been improved by a wide margin using a bi-gram vocabulary which makes it the second best system in terms of ROUGE-SU4 as reported in [Delort and Alfonseca, 2012]. It is possible that this modification will also help the local models in our case and we leave that as a direction of future research. Note that DualSumm has not been run for the TAC 2010B

dataset since its parameters are optimized using TAC 2010B and other related datasets.

TAC 2010B – ROUGE-SU4 scores									
<i>Human Summaries</i>	Human-Highest	Conf.-Int.	Human-Mid1	Conf.-Int.	Human-Mid2	Conf.-Int.	Human-Lowest	Conf.-Int.	
		0.16193	0.14371-0.18182	13502	0.11667-0.15337	0.13365	0.12095-0.14524	0.11591	0.10286-0.12953
<i>Proposed Models</i>	K={60}	Conf.-Int.	K={60, 80}	Conf.-Int.	K={60, 80,100}	Conf.-Int.	K={60, 80,100, 120}	Conf.-Int.	
TagLDA+ local models	0.11614	0.11128-0.12133	0.11645	0.11173-0.12162	0.11642	0.11169-0.12160	0.11642	0.11169-0.12160	
METag ² LDA+ local models	0.11606	0.11120-0.12131	0.11608	0.11120-0.12131	0.11621	0.11133-0.12141	0.11605	0.11120-0.12130	
CorrME Tag ² LDA+ local models	0.11532	0.11064-0.12032	0.11533	0.11064-0.12032	0.11531	0.11061-0.12030	0.11533	0.11064-0.12032	
TAC 2010B – ROUGE-2 scores									
<i>Human Summaries</i>	Human-Highest	Conf.-Int.	Human-Mid1	Conf.-Int.	Human-Mid2	Conf.-Int.	Human-Lowest	Conf.-Int.	
		0.13021	0.10972-0.15240	0.09595	0.07451-0.11730	0.09538	0.07366-0.11592	0.07663	0.06199-0.09138
<i>Proposed Models</i>	K={60}	Conf.-Int.	K={60, 80}	Conf.-Int.	K={60, 80,100}	Conf.-Int.	K={60, 80,100, 120}	Conf.-Int.	
TagLDA+ local models	0.07438	0.06928-0.07962	0.07443	0.06941-0.07969	0.07443	0.06941-0.07969	0.07443	0.06941-0.07969	
METag ² LDA+ local models	0.07411	0.06902-0.07928	0.07439	0.06930-0.07965	0.07444	0.06932-0.07971	0.07411	0.06902-0.07928	
CorrME Tag ² LDA+ local models	0.07339	0.06863-0.07821	0.07339	0.06863-0.07821	0.07339	0.06863-0.07821	0.07339	0.06863-0.07821	

Table 5.13: ROUGE-SU4 and ROUGE-2 scores from our proposed models for TAC 2010B.

Table 5.13 shows the ROUGE-SU4 and ROUGE-2 scores of the summaries from our proposed models for Guided Summarization as applied to the update collections in the TAC 2010 summarization task dataset. We observe from the table that adding the global tag-topic models into the summarization process improves the mean ROUGE scores here as well and makes our summaries from (ME)Tag²LDA models statistically no different than CLASSY’s for TAC 2010B.

Table 5.14 similarly shows the ROUGE scores from our proposed models for TAC 2011B. While the summaries from the combination of TagLDA and the METag²LDA models with the local models remain parallel to the CLASSY system, they do not outperform the simple **Docset-Noun+Verb-RSTree** local model. However the lower bounds of the ROUGE-SU4 scores of the top peers (including the upper median of the human summarizers) encompass the mean ROUGE-SU4 score from our model and are thus statistically no different.

The reason behind the success of our local models become apparent as we observe the **docset-tfif-noun+verb** feature set as shown in Table 5.15 for the update collections for some docsets in the TAC 2011 collection. We see that the most of the important nouns and verbs remain the same in the update

TAC 2011B – ROUGE-SU4 scores									
<i>Human Summaries</i>	Human-Highest	Conf.-Int.	Human-Mid1	Conf.-Int.	Human-Mid2	Conf.-Int.	Human-Lowest	Conf.-Int.	
		0.14941	0.12998-0.16971	0.13461	0.11777-0.15447	0.12961	0.11891-0.14328	0.12105	0.10973-0.13229
<i>Proposed Models</i>	K={60}	Conf.-Int.	K={60, 80}	Conf.-Int.	K={60, 80,100}	Conf.-Int.	K={60, 80,100, 120}	Conf.-Int.	
TagLDA+ local models	0.12748	0.12307-0.13179	0.1274	0.12196-0.13162	0.12724	0.12173-0.13140	0.12724	0.12173-0.13140	
METag ² LDA+ local models	0.12721	0.12268-0.13158	0.12765	0.12209-0.13191	0.12768	0.12212-0.13194	0.12766	0.12210-0.13195	
CorrME Tag ² LDA+ local models	0.12706	0.12258-0.13137	0.12692	0.12139-0.13101	0.12697	0.12142-0.13102	0.12699	0.12142-0.13105	
TAC 2011B – ROUGE-2 scores									
<i>Human Summaries</i>	Human-Highest	Conf.-Int.	Human-Mid1	Conf.-Int.	Human-Mid2	Conf.-Int.	Human-Lowest	Conf.-Int.	
		0.11474	0.09242-0.13859	0.10069	0.08521-0.11725	0.09079	0.07634-0.10847	0.07937	0.07242-0.09742
<i>Proposed Models</i>	K={60}	Conf.-Int.	K={60, 80}	Conf.-Int.	K={60, 80,100}	Conf.-Int.	K={60, 80,100, 120}	Conf.-Int.	
TagLDA+ local models	0.08638	0.08116-0.09159	0.08511	0.07998-0.09009	0.08511	0.07983-0.09028	0.08511	0.07983-0.09028	
METag ² LDA+ local models	0.08667	0.08131-0.09191	0.08603	0.08065-0.09124	0.08603	0.08065-0.09124	0.08609	0.08068-0.09125	
CorrME Tag ² LDA+ local models	0.08621	0.08090-0.09136	0.08493	0.07960-0.08995	0.08493	0.07960-0.08995	0.08493	0.07960-0.08995	

Table 5.14: ROUGE-SU4 and ROUGE-2 scores from our proposed models for TAC 2011B.

collection but some new key terms get introduced that immediately give us an idea as to what is new. For example, for docset D1102A, one easily gets the idea that some protection is being provided to the affected parties and many organizations are being established to follow up on the Internet security threat. Similar instances can be found in the other docsets as well.

5.6 Summary

In summary, we have shown that it is possible to use unsupervised models that do latent structure discovery of (word, annotation) ensembles in text for effective extractive multi-document summarization. The benefits of exploratory data analysis with multimodal topic models is very well stated in the text mining literature and providing a multi-document summary view of the latent topics can also be extremely beneficial for searching the topic space through information needs.

From the summarization perspective, the likelihoods of the sentences with local contexts that are fit to the models together with simple docset specific models which capture relevancy in target documents show state-of-the-art multi-document summarization power in terms of automatic evaluation. The use of

Docset ID / Query [category]	Time line	Important Nouns	Important Verbs
D1105A / Plane Crash Indonesia [Accidents and Natural Disasters]	Base (D1105A-A)	Adam, Air, Boeing, Hartono, Sulawesi, accident, board, crash, emergency, official, pain, passenger, plane, rescue, search, survivor	carry, disappear, find, kill, miss, send
	Update (D1105A-B)	Adam, Air, Indonesia, Sulawesi, airline, board, crash, island, plane, rescue, search, survivor, wreckage	comb, disappear, find, fly, miss, pass, send
D1101A / Amish Shooting [Attacks (Criminal/ Terrorist)]	Base (D1101A-A)	Miller, Roberts, attack, child, door, dream, family, girl, man, neighbor, number, police, school, schoolhouse, victim, wife	enter, kill, leave, molest, shoot, speak, storm, tie, turn, weave
	Update (D1101A-B)	Roberts, burning, dispatcher, girl, lot, problem, schoolhouse, seconds, yesterday	attend, bury, expect, kill, line, molest, shoot
D1102A / Internet Security [Health and Safety]	Base (D1102A-A)	Internet, VeriSign, address, attack, business, company, computer, datum, domain, investment, security, server, system, technology, traffic, user, virus	convert, grow, manage, may, operate
	Update (D1102A-B)	China, Internet, Nelson, attack, computer, government, incident, information, official, security, space, system, user, video, vulnerability, website	establish, find, follow, protect, provide
D1106A / Tuna Fishing [Endangered Resources]	Base (D1106A-A)	Japan, Kobe, Ocean, catch, conference, conservation, country, fishery, fishing, management, meeting, overfishing, plan, stock, tuna	adopt, expect, include, poach, track
	Update (D1106A-B)	Japan, bluefin, boat, capacity, catch, conference, country, fishing, fleet, meeting, participant, plan, quota, regulator, scorecard, stock, trouble, tuna	adopt, agree, eat, poach, send

Table 5.15: Few sample docset IDs, queries and categories from the TAC 2011 dataset. The **docset-tfjsf-noun+verb** feature set is shown for base and update collections.

RS-trees for sentence compression leading to bullet list summaries also show extremely promising result within our genre-agnostic summarization framework.

As a future work we want to experiment with dependency triplets as vocabulary units to see if those can improve not only ROUGE-SU4 scores but ROUGE-2 scores as well. An alternate completely local and more efficient summarization system can be built by focusing the tag-topic models to local docsets only as in [Celikyilmaz and Hakkani-Tür, 2011]. To address the issue of readability involving coherence, we can easily apply the traveling salesman approach [Conroy et al., 2010] to order sentences using both surface similarity as well as our coherence triplets. Finally, we will want to follow up on the hard problem of summarization through natural language generation [Genest and Lapalme, 2011] but using our techniques to achieve a level of paraphrasing that is close to those by humans.

The next chapter shows a beautiful connection of the problem of summarization presented in this chapter to the problem of text summarization of videos. Video to text summarization really addresses what humans think are salient objects and actions which can be included into a concise lingual descrip-

tion of the video. We build new topic models which learn a rough translation of some *domain specific* high level concise lingual descriptions to the low level pattern of features extracted from the videos and then tries to conceptually describe a test video from the same domain given only its low level features.

5.7 Acknowledgements

We thank Lucy Vanderwende of Microsoft Research and Enrique Alfonseca of Google Research for several useful discussions on the applicability of bullet list summaries during a meeting of the Text Analysis Conference, 2011 including the latter author's permission to re-use their new scores and useful comments on the first draft.

Chapter 6

Summarizing Videos into Natural Language Text

“The translator has to do consciously what the author did instinctively. And yet it must seem instinctive.” - Richard Pevear

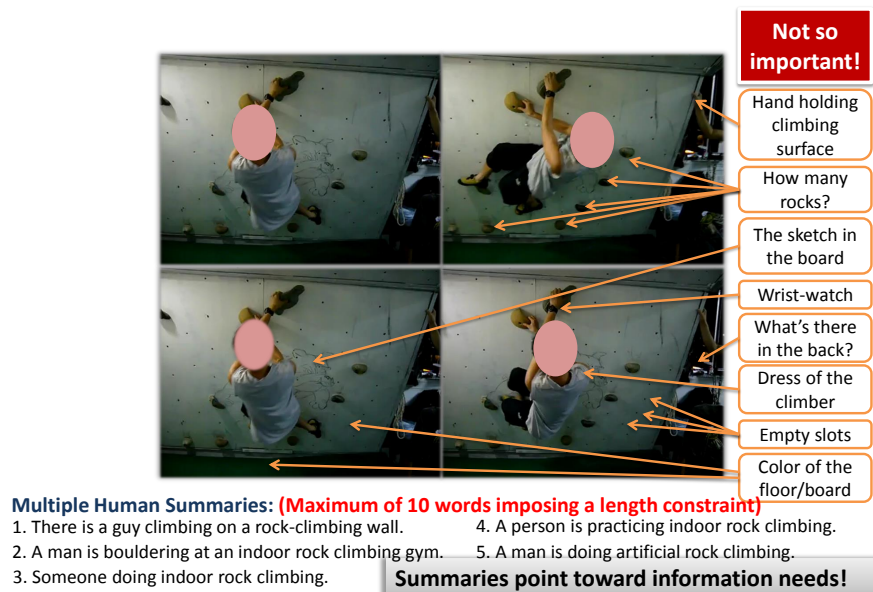
6.1 Introduction

In recent years there has been an abundance of multimedia data in the form of video contents from television networks, video uploads to websites and so on. However, organizing such data by integrating the semantic content of the videos is a very difficult problem. On the other hand, summarizing videos directly into text summaries can lead to significant improvements in many end-user applications—multimedia search experience, content based advertisements [Welch et al., 2010], helping the visually impaired and so on. In this paper we concern ourselves with two language agnostic tasks: **i)** Building a topic modeling framework to model multimedia documents consisting of videos and textual metadata and **ii)** using the topic modeling framework to predict bag-of-words summaries for a new video belonging to a previously known category. The first task helps us discover semantically related concepts in the text through latent topics and translating them to topically related videos or frames. The second task takes a test video and generates intermediate text keywords ideal for natural language generation.



Figure 6.1: An example of the task of video summarization

As a further addition we also experiment with efficient natural language sentence generation from the predicted bag of words (henceforth BoW) using language models and confidence of syntactic parse



(a) Short summaries from human annotators on an indoor rock climbing video

Figure 6.2: Do we speak all that we see? Human summaries of a short video on rock climbing

tree generation following a simple template. The first two tasks, though, are the focus of this paper. Fig. 6.1 shows some keyframes of two sample videos from our training dataset and the short summaries written by human annotators. This dataset is discussed in Section 6.1.1. *In our paper, translation from video to text is synonymous to summarizing a video with a set of textual keywords.*

Following on with the example in Chapter 1, Fig. 6.2 shows how four human annotators have summarized a rock climbing video in one sentence. Interestingly, all sentences focus on the “central” action and the objects associated with it and consider most of the background objects to be irrelevant.

It seems intuitive that a topic model which incorporates low level vision features representing objects, actions, color and scenes and correspond those to the text summaries should have a better chance of describing the multimedia data. We thus seek representations of visual data that mimic the subject-verb-object-scene quadruplet structure of English sentences in terms of subject, object and scene nouns as well as verbs. Additionally illumination gives rise to color which differentiates one object from another and is often expressed as adjectives. Objects, actions and color can be visualized using specific word concepts and can be counted over time thus lending themselves to quantization but scene represents global energy distributions which pervade the arrangement of objects and is thus better represented as real values.

In the context of this paper, our training data consists of videos and associated summaries. The training data is also available with event category labels for e.g. “boarding event” and topic modeling video documents in each event can allow us to discover “sub-topics” e.g. “skateboarding,” “snowboarding” and “surfing”. Of course, our topic models do not include any event label bias and can be applied on the overall dataset as well. However we will see shortly that doing so can be very unappealing to end users when summaries need to be generated.

Apart from the topical analysis of video documents, the problem of generating summaries directly

from videos also has significant end user appeal. We emphasize that the video summarization/translation task in this paper is to describe an entire video firstly as a bag of salient keywords i.e. BoW, and then, as a further addition, use simple Natural Language Generation (henceforth NLG) techniques to summarize the bag of words into a human readable paragraph of text wherever possible. Translating and generating summaries from a video can always be looked upon as finding the right information need which is paramount to any search problem.

Video summarization in our context is different from image annotation mentioned in early literature [Srihari, 1991] and in more modern ones [Makadia et al., 2008]. If we perfectly annotate all objects and actions in each frame correctly, we can probably use a standard topic modeling technique [Blei et al., 2003] to figure out the themes in the multimedia documents. However, detecting objects [Li et al., 2010b] and scenes [Oliva and Torralba, 2001] in images and actions [Kläser et al., 2008] and keyframes [Li et al., 2011] in videos in a reliable manner are open problems in the computer vision community [Makadia et al., 2008]. Most of these detection techniques fall in the domain of supervised learning and require large amounts of annotated data. Video annotation task is a particularly laborious process even though tools like VATIC [Vondrick et al., 2010a] have been written to ease the effort (for an interactive demo of the tool, see VATIC’s website¹). Also firing thousands of detectors to accurately label even a single keyframe of an unknown video leads to many false positives. This is particularly true of the videos *in-the-wild* i.e. videos downloaded from the Internet where there are plenty of resolution problems, severe motion blurs, camera shakes etc. These are the videos that we experiment with in this paper.

We view the video to text translation/summarization problem in the light of multidocument summarization of plain text documents which has been popularized by the Text Analysis Conference² (TAC). In the multidocument summarization track of TAC, participants are given document sets (docsets) of newswire articles typically belonging to 5 major event types like “Health and Safety,” “Accidents and Natural Disasters” etc. and are asked to generate a fixed length fluent summary of the documents in each docset. A docset in the TAC setting is unique in that it contains a set of documents that are relevant for a particular information need like “Cyclone Katrina.” The system summaries are scored in several ways including the most reliable manual way using PYRAMID [Nenkova and Passonneau, 2004] evaluation but systems usually are tuned w.r.t. the automatic ROUGE [Lin and Hovy, 2003] scoring. By analogy, we assume that each docset here corresponds to a video and contains a sequence of frames and a set of keyframes. At test time we are given unknown event specific videos without any text summary. For measuring system performance, we generate summaries of videos and evaluate them using the recall oriented ROUGE-1 score to measure the percent overlap of the words in the short ground truth summaries.



Human Summary: Montage of clips from an outdoor wedding

Predicted bag of words summary: birthday wed indoor outdoors mob dance flash cake parade ceremony fish

Figure 6.3: An example of vocabulary intrusion in the task of video summarization. Best viewed with magnification

A key concern in generating a BoW summary of a video is the vocabulary intrusion problem. Fig. 6.3 shows an example of vocabulary intrusion in the task of video summarization that arises out of topic modeling on the entire vocabulary of the corpus. If we consider a vocabulary of V words—the

¹<http://mit.edu/vondrick/vatic/>

²<http://www.nist.gov/tac/2011/Summarization/>

probability of getting the top L words correctly in the summary is $(1/V)(1/(V-1))\dots(1/(V-L+1))$. If V is large (such as 2000) then the probability is very low. Further, if the entire vocabulary is used, then intrusive words describing other but related event categories like “birthday, flash mob, dance, parade, fish” can appear with high probability (see a possible predicted BoW summary in Fig. 6.3 from a topic model (Fig. 6.4b) trained over all events with number of topics set to 200). This problem is mitigated by first classifying the test video into its corresponding event category (Section 6.4.4) and then using a topic model to predict the BoW summary. In the absence of the event labels, this direction improves readability and is much faster.

The novelty in our new approach to topic modeling video documents with textual metadata is the use of the right features for the videos and augmenting basic topic models for joint modeling with those features along with text. We represent each video in terms of objects, actions, color (represented with discrete distributions) and scenes (represented with Normal distributions with unknown means and variances) and try to find a translation space that translates the pattern of these features to a permutation in language vocabulary. Such a representation of a video is both intuitive and logical. We observe that the interplay of the full spectrum of representations (Section 6.4) indeed yield the highest likelihoods to held out test data than those using partial representations (Section 6.4.1).

6.1.1 Dataset Description

The dataset that we use for the video summarization task is released as part of NIST’s 2011 TRECVID Multimedia Event Detection (MED) evaluation set³. The dataset consists of a collection of Internet multimedia content posted to the various Internet video hosting sites. The training set is organized into 15 event categories, some of which are:

1) Attempting a board trick 2) Feeding an animal 3) Landing a fish 4) Wedding ceremony 5) Working on a woodworking project 6) Birthday party 7) Changing a vehicle tire 8) Flash mob gathering 9) Getting a vehicle unstuck 10) Grooming an animal 11) Making a sandwich 12) Parade 13) Parkour 14) Repairing an appliance and 15) Working on a sewing project.

We use the videos and their textual metadata in all the 15 events as training data. There are 2062 clips with summaries in the training set with almost equal distribution amongst the events. The test set which we use is called the Transparent Development (Dev-T) collection. The Dev-T collection includes positive instances of the first 5 training events and near positive instances for the last 10 events—a total of 630 videos labeled with event category information (and associated human synopses which are to be compared against for summarization performance). Each summary is a short and very high level description of the entire video and ranges from 2 to 40 words but on average **10** words (with stopwords). We remove standard English stopwords and retain only the word morphologies (not required) from the synopses as our training vocabularies. The proportion of videos belonging to events 6 through 15 in the Dev-T set is much low compared to the proportion for the other events since those clips are considered to be “related” instances which cover only part of the event category specifications. The performances of our topic models are evaluated on those kinds of clips as well. The numbers of videos in events 6 through 15 in the Dev-T set are {4,9,5,7,8,3,3,3,10,8} while there are around 120 videos per event for the first 5 events. All other videos in the Dev-T set neither have any event category label nor are identified as positive, negative or related videos and we do not consider these videos in our experiments.

³<http://www.nist.gov/itl/iad/mig/med11.cfm>

6.1.2 Evaluation Measures

We measure the predictive performance of the topic models using the Evidence Lower BOunds (ELBO) on held-out test set—the Dev-T collection *with* summaries, as well as the predictive ELBO for BoW summary generation on the held-out Dev-T collection *without* summaries (Section 6.4.1). ELBO is just log likelihood and is directly related to measuring average perplexity of the model per observed textual word [Blei et al., 2003, Blei and Jordan, 2003]. We also evaluate our BoW summaries using the ROUGE scorer. ROUGE measures the n-gram overlap for system generated summaries to the ones written by annotators and the scores are interpreted in terms of recall. Usually 4 gold standard summaries are needed for evaluation but here we use the base case of using only one short summary as a reference summary per video on this dataset. While summarizing, since our primary task is to evaluate only the BoW summaries generated from a video, we use the ROUGE-1 unigram measure. We evaluate 5 and 10 keywords long BoW summaries respecting the average length of the short human summaries. Since we are considering videos in the Dev-T set with event category information, we can use the ROUGE evaluation setup of multidocument summarization as used in TAC. If the categories are not known, we can multiply the ROUGE scores with the event classification accuracies to obtain lower bounds (see Section 6.4.4 for lower bounds on classification accuracies). Evaluations with higher order n-grams are not needed for unigram translations. We do not use manual evaluations since the data cannot be released for public verifications.

The task of discovering topically related words is mostly evaluated w.r.t ELBO. We use the topic models from [Blei and Jordan, 2003] as baselines. We modify the GM-LDA model in [Blei and Jordan, 2003] following [Ramage et al., 2009b] to use discrete visual data and name the model MMLDA—“MM” stands for the multinomials for text as well as the multinomials for the visual words. We implement a deterministic optimization framework for MMLDA instead of the non-deterministic sampling as in [Ramage et al., 2009b]. The Corr-LDA model in [Blei and Jordan, 2003] is also extended by using Normal-Wishart priors and named Corr-MGLDA (M for Multinomials and G for Gaussians). For evaluating video to text summarization based on ROUGE-1 scores, we use a non-topic model based automatic image annotation tool as the baseline for video labeling by using labels aggregated from keyframes. Our topic model based video summarization methods outperform the state-of-the-art image to text translation model [Li et al., 2010b] applied on video keyframes in terms of ROUGE-1 scores of the predicted keyword summaries.

6.2 Related Work

Makadia et al. [Makadia et al., 2008] uses nearest neighbor and label transfer techniques to annotate images suitable for the image retrieval task. However, we can not directly apply their methods as the individual frames/keyframes of the videos in our dataset are not annotated. Based on the size and genre of our dataset, such annotations prove very expensive and we do not follow that direction. Further, we are interested in the task of direct natural language summarization of the entire video and not specific annotation of a vast majority of possible objects, actions and scenes in every frame/keyframe of the video. The closest work to our task is by Yang et al. [Yang et al., 2011] where low level object and scene classifiers are used to obtain object and scene labels in an *image*. These are then combined using background language models and Hidden Markov Models to predict a natural language sentence that automatically includes the best possible verb i.e. action. *We will observe in Section 6.4.4 that actions,*

which are intrinsic to videos, are important event discriminators. Further, none of above mentioned methods can discover related concepts as latent topics and translate them into related frames.

In the domain of topic modeling of images with captions, the Corr-LDA model has recently been extended to handle a multinomial feature space in [Putthividhya et al., 2010] with different number of topics for visual word type and textual word type. The model learns an association from the topic proportions over image domain to those over text domain through a regression formulation. However, during prediction, this dependency needs to be marginalized out anyways. Also, if we quantize every type of real valued vision feature using some clustering algorithm such as K-means into C clusters, then each C represents a parameter of the final model and performance analysis become that much more difficult. Ahmed et al. [Ahmed et al., 2009] uses Gaussian feature vectors and mention Normal-Wishart priors but do not use them—they use uniform priors in a non-deterministic sampling framework instead. The correlated topic model in [Blei and Lafferty, 2005] is extended to capture multimodal discrete distributions in [Xu et al., 2013] for image annotation purposes. Although codebook feature extraction is standard practice in the computer vision community, the main drawback of these models is that there is no natural way to translate a set of topically related words to topically related frames from a video using only visual codebook features.

On the other hand the Continuous Relevance Model (CRM) [Lavrenko et al., 2004] and Multiple Bernoulli Relevance Model (MBRM) [Feng et al., 2004] assume different, nonparametric density representations of the joint word-image space. CRM gets rid of the latent factor representation and achieves non-parameterization. The dataset used in [Feng et al., 2004] for MBRM has hierarchical word annotations which are handled using multiple Bernoulli models rather than multinomial distributions. In our dataset, multinomial distributions are sufficient since the summaries read like very short documents with repeated word morphologies.

Detecting objects can often be seen as an important step towards identifying the main topic of a video and generating a BoW summary. To that end, Torresani et al. [Torresani et al., 2010] transform an image feature vector into a another lower dimensional feature vector whose values are the outputs of several category classifiers (which are named “classems” in their paper). We take a similar approach to convert Object Bank [Li et al., 2010b] (OB) feature vectors to high level sparse histogram of object detectors to be used in our discrete video data representation and as *baseline* for video to text translation. To extract OB features, keyframes are identified to reduce computational time. Keyframe detection is a research topic in its own right, where some recent ones include more involved techniques [Li et al., 2011] using Transfer Learning from accompanying text transcripts. However, the keyframes extracted using the change in color histogram [Zhang et al., 1995] satisfy our purposes.

In the domain of topic modeling of videos, the Hidden Topic Markov Model in [Wanke et al., 2010] does not incorporate both text and visual words in a single framework and also does not use a fuller representation of videos as we do. A Markovian assumption is also imposed in [Hospedales et al., 2009] for modeling actions and identifying behaviors (with no automatic labeling), however, we can safely ignore frame dependence because our action features are derived using temporal windows and activity tracking is not an objective in this paper. The reformulations of LDA and CTM using class labels and without any temporal dynamics in [W. and M., 2009] also target activity classification. To the best of our knowledge this is the first work to use mixed membership topic models for video to text summarization which *can* eliminate frame-wise object annotations.

The proposed models are discussed in the following section in as much depth as possible. The use

Symbol	Meaning (<i>r.v.</i> = <i>random variable</i>)
D	total number of multimedia “documents”
M	total number of discrete text features per multimedia document $d \in \{1, \dots, D\}$
H	total number of discrete visual features in a multimedia document $d \in \{1, \dots, D\}$
O	total number of real valued visual features per multimedia document $d \in \{1, \dots, D\}$
$\alpha = \{\alpha_1, \dots, \alpha_K\}$	r.v. for asymmetric Dirichlet prior for the document level topic proportions
θ_d	r.v. for document level latent topic proportions
ρ	corpus level topic multinomials over discrete video features
β	corpus level topic multinomials for textual words
μ	means of topic Gaussians for the real valued features from videos
$\Lambda = (\Sigma^{-1})$	precision (inverse covariance) matrices of topic Gaussians for the real valued features from videos
y_m in Figs. 6.4b, and 6.4d	indicator variable for a sample from θ_d for discrete text features
y_m in Figs. 6.4c and 6.4e	indicator variable for <i>document level</i> real valued datum correspondences
z_h	indicator variable for a sample from θ_d for discrete visual features
z_o	indicator variable for a sample from θ_d for real valued visual features
w_m	r.v. for textual word at position m in document d ; vocabulary size of V
w_h	r.v. for vision oriented discrete feature at position h in document d ; vocabulary size $corrV_H$
w_o	r.v. for the o^{th} Gaussian feature vector with a dimensionality of P in document d

Table 6.1: Meanings of the variables used in the models

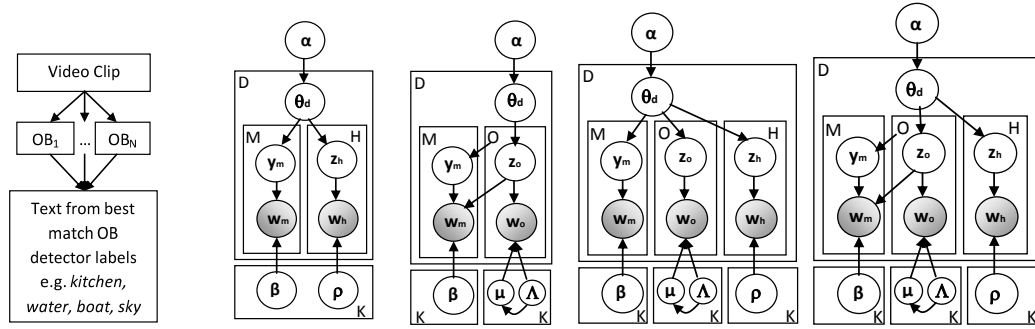
of asymmetric Dirichlet priors over the topic proportions helps us achieve better sparsity in topics. However, this also leads to singularities in precision matrices conditioned on topics when Normal-Wishart priors for real valued data are not used.

6.3 The Proposed Models

In this section we describe our proposed topic models for multimedia documents. We call the model in Fig. 6.4d MMGLDA, short for Multinomial-Multinomial-Gaussian LDA, and the model in Fig. 6.4e to Corr-MMGLDA, short for correspondence MMGLDA. In our context, the correspondence LDA model [Blei and Jordan, 2003] places a probabilistic constraint on the correspondence of summary words to Gaussian observations—a word is likely to be generated by the topic which is agreed upon by most of the Gaussian instances in the video document. Since the real valued GIST features (see Section 6.4) “summarize” a scene in an image, we want a stronger influence of the topic of the scene on the summary text. This assumption is relaxed in MMGLDA. Of course the correspondence could have been established between the discrete observations only or both discrete and real valued ones but conditioned on the current dataset, we want more flexibility in topics for sampling discrete observations. We avoid overly complicated topic models and instead go for better data representations and supporting models with just the right amount of complexity.

In MMGLDA, for a multimedia document d , it is possible to have different topics competing for each occurrence of w_m . In Corr-MMGLDA, the number of such modes is constrained to be much fewer. The asymmetric α can yield few additional modes which group co-occurring data dominant in densities or masses in separate latent topics. This phenomenon is observed for a larger number of latent topics.

Table 6.1 explains the symbols used in the two proposed topic models. Everywhere in this paper, we



(a) Object Bank object detection model [Li et al., 2010b] (b) MMLDA [Blei and Jordan, 2003, Ramage et al., 2009b] (c) Corr-MGLDA [Blei and Jordan, 2003] (extended) (d) MMGLDA (proposed) (e) Corr-MMGLDA (proposed)

Figure 6.4: Graphical model representations of existing topic models and proposed extensions— Figs. 3d and 3e. In this paper, we extend the model in Fig. 3c i.e. the Corr-LDA model in [Blei and Jordan, 2003] with Normal-Wishart priors over parameters for real valued observations as well.

assume that K is the number of topics. The generative processes for the proposed models are illustrated below:

- For each video document $d \in 1, \dots, D$
 - Choose a topic proportion $\theta | \alpha \sim Dir(\alpha)$
 - For each position h in d
 - Choose topic indicator $z_h | \theta \sim Mult(\theta)$
 - Choose a discrete video “word” $w_h | z_h = k, \rho \sim Mult(\rho_{z_h})$
 - For each real valued observation o in d
 - Choose topic indicator $z_o | \theta \sim \mathcal{N}(\mu, \Lambda^{-1})$
 - Choose a real valued $w_o | z_o = k, \mu, \Lambda^{-1} \sim \mathcal{N}(\mu_{z_o}, \Lambda_{z_o}^{-1})$
 - For each position m in video d
 - Choose $y_m \sim Uniform(1, \dots, O)$ (for Fig. 6.4e)
 - or Choose $y_m | \theta \sim Mult(\theta)$ (for Fig. 6.4d)
 - Choose a word $w_m \sim p(w_m | z_{y_m}, \beta)$ (for Fig. 6.4e)
 - or Choose a word $w_m \sim p(w_m | y_m, \beta)$ (for Fig. 6.4d)

In all further notations, \mathbf{w}_M is the ensemble of observed random variables that represent summary words in the d^{th} multimedia document. Similar notations hold for \mathbf{w}_O , \mathbf{w}_H and the indicators \mathbf{y} and \mathbf{z} . In this paper, the text vocabularies are event specific and of size 312 words on average.

Fig. 6.4a shows the Object Bank [Li et al., 2010b] (OB) baseline that we initially used to translate videos to text. The boxes labeled OB_1, \dots, OB_N are the individual object detectors in Object Bank. The positive responses of the detectors lead towards identifying the label of the objects in the keyframes and hence translating the entire video. We choose this baseline to verify the difficult nature of our dataset— there is a 10% overlap between OB’s vocabulary and the test set vocabulary, (see Section 6.8), and we should expect to see at least 2-5% recall in ROUGE-1 recall scores for most events based on a 40-50% ROUGE-1 recall achieved by the best 100-word multidocument text summarization systems in TAC competitions.

6.3.1 Inference on Latent Variables

We use the variational Bayesian Expectation Maximization [Beal, 2003, Wainwright and Jordan, 2008] algorithmic framework as the optimization framework. An advantage of VBEM is that it is deterministic.

The derivations for the MMG class of topic models become sufficiently complicated due to the need for using priors over the parameters governing the real valued observations. Since the nature of the modes for topic proportions is not known in advance, singularities arising out ill-conditioned topic covariance matrices must be handled. This problem is mitigated in a principled way by introducing independent Normal-Wishart priors governing the mean vectors and precision matrices of the Gaussians conditioned on the topics. Since both $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are unknown we cannot factorize $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ directly because the variance of the distribution over $\boldsymbol{\mu}$ is a function of $\boldsymbol{\Lambda}$. Instead we use combinations of Normal-Wishart priors on each Gaussian component as:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_0, (\kappa_0 \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0) \quad (6.1)$$

where $\Sigma_k^{-1} = \boldsymbol{\Lambda}_k$ is the precision matrix for the k^{th} factor or topic. This is similar to the mixture model used in [Nasios and Bors, 2006]. To preserve the dependence between the means and covariances, a *partially* factorized tractable q distribution with “free” variational parameters $\boldsymbol{\gamma}$, $\boldsymbol{\phi}$, $\boldsymbol{\phi}^{(O)}$, $\boldsymbol{\phi}^{(H)}$ (for every multimedia document $d \in D$) is imposed by

$$q(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}_O, \mathbf{z}_H | \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\phi}^{(O)}, \boldsymbol{\phi}^{(H)}) = \left[\prod_{d=1}^D q(\boldsymbol{\theta}_d | \boldsymbol{\gamma}_d) \left[\prod_{m=1}^{M_d} q(y_{d,m} | \boldsymbol{\phi}_{d,m}) \prod_{o=1}^{O_d} q(z_{d,o} | \boldsymbol{\phi}_{d,o}^{(O)}) \prod_{h=1}^{H_d} q(z_{d,h} | \boldsymbol{\phi}_{d,h}^{(H)}) \right] \right] \times \prod_{k=1}^K q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \quad (6.2)$$

with $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\gamma}_d)$, $z_{d,o} \sim \text{Mult}(\boldsymbol{\phi}_{d,o}^{(O)})$ and $z_{d,h} \sim \text{Mult}(\boldsymbol{\phi}_{d,h}^{(H)})$. The maximum likelihood (ML) estimates of free parameters are found by optimizing the lower bound on $\ln p(\mathbf{w}_M, \mathbf{w}_H, \mathbf{w}_O | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$.

The hyperparameters for $\boldsymbol{\alpha}$ in the asymmetric Dirichlet case (the concentration parameter and the base measure) and κ_0 , ν_0 , \mathbf{m}_0 and \mathbf{W}_0 are not shown in Figs. 6.4c, 6.4d and 6.4e and in equ. 6.2 above. Also $\boldsymbol{\phi}$ are the free parameters of the variational summary_word multinomials over Gaussian_observations in the correspondence multimodal models or summary_word multinomials over topics in the plain multimodal models; $\boldsymbol{\phi}^{(O)}$ are the free parameters of the variational Gaussian_observation multinomials over topics and similarly for $\boldsymbol{\phi}^{(H)}$ for discrete visual features. \mathbf{m}_k , κ_k , and \mathbf{W}_k , ν_k are the free parameters for the Gaussians defined for every topic. These free parameters are defined for every video document $d \in D$.

The variational posterior distribution $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ does not factorize into the product of the marginals, but we can always write it as $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = q(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k) q(\boldsymbol{\Lambda}_k)$. Then we use the result from mean field theory [Parisi, 1988, Wainwright and Jordan, 2008] that says that the \ln of the optimal solution for factor q_j is obtained by considering the \ln of the joint distribution over all hidden and observed variables and then taking the expectation with respect to all of the other factors $\{q_i\}$ for $i \neq j$ i.e. for visible and hidden variable ensembles V and H , $\ln q_j^*(H_j) = E_{i \neq j} [\ln p(V, H)] + \text{const}$. For the Gaussian parameters, the optimal solution for $q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ depends on moments evaluated with respect to the distributions of other variables, and so again the variational update equations are coupled and must be solved iteratively. This

results in a Normal-Wishart distribution and is given by:

$$q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_k, (\kappa_k \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_k, \nu_k) \quad (6.3)$$

where $\boldsymbol{\Sigma}_k^{-1} = \boldsymbol{\Lambda}_k$ is the precision matrix for the k^{th} factor or topic. The expression in Equ. 6.3 is obtained by first writing out the expression for $\ln q^*(\cdot)$ and selecting those terms that involve $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$. This yields:

$$\ln q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{i=1}^K \ln p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) + \sum_{d=1}^D \sum_{o=1}^{O_d} \sum_{i=1}^K \phi_{d,o,i}^{(O)} \ln \mathcal{N}(\mathbf{w}_{d,o} | \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i^{-1}) + const \quad (6.4)$$

Note that the variance of the distribution over $\boldsymbol{\mu}_k$ is a function of $\boldsymbol{\Lambda}_k$. The random variables \mathbf{m}_k and \mathbf{W}_k can be thought of as surrogates to \mathbf{m}_0 and \mathbf{W}_0 and that κ_k and ν_k surrogates to κ_0 and ν_0 but conditioned on latent topic k . The expressions for these variables, which are also used in the M-Step updates, can be found in Eqs. 6.29, 6.30, 6.28 and 6.31. These expressions are obtained by matching the moments of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$ to the Normal and Wishart distribution expressions. The optimal solution for $q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ depends on the moments evaluated with respect to the distributions of other variables, and so the variational update equations are coupled and must be solved iteratively. Following [Blei et al., 2003, Blei and Jordan, 2003], let us now write down the objective functional, $\mathcal{L}(\cdot)$, to be maximized which acts as the lower bound to the true data log likelihood.

For the MMGLDA model:

$$\begin{aligned} \mathcal{L}_{MMGLDA} = & E_q[\ln p(\boldsymbol{\theta} | \boldsymbol{\alpha})] + E_q[\ln p(\mathbf{y}_M | \boldsymbol{\theta})] + E_q[\ln p(\mathbf{w}_M | \mathbf{y}_M, \boldsymbol{\beta})] + E_q[\ln p(\mathbf{z}_O | \boldsymbol{\theta})] \\ & + E_q[\ln p(\mathbf{w}_O | \mathbf{z}_O, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + E_q[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \mathbf{m}_0, \mathbf{W}_0, \kappa_0, \nu_0)] + E_q[\ln p(\mathbf{z}_H | \boldsymbol{\theta})] + E_q[\ln p(\mathbf{w}_H | \mathbf{z}_H, \boldsymbol{\rho})] \\ & - E_q[\ln q(\boldsymbol{\theta}, \mathbf{y}_M, \mathbf{z}_O, \mathbf{z}_H, \boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\phi}^{(O)}, \boldsymbol{\phi}^{(H)}, \mathbf{m}, \mathbf{W}, \boldsymbol{\kappa}, \boldsymbol{\nu})] \end{aligned} \quad (6.5)$$

Thus for each video document d , the RHS in equation (6.5), with indices d suppressed where appropriate, expands out to be:

$$\ln \Gamma\left(\sum_{j=1}^K \alpha_j\right) - \sum_{i=1}^K \ln \Gamma(\alpha_i) + \sum_{i=1}^K (\alpha_i - 1) \left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^K \gamma_j\right) \right) \quad (6.6)$$

$$+ \sum_{i=1}^K \left(\sum_{m=1}^M \phi_{m,i} \left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^K \gamma_j\right) \right) + \sum_{m=1}^M \phi_{m,i} \ln \beta_{i,w_m} \right) \quad (6.7)$$

$$\begin{aligned} & + \sum_{i=1}^K \left(\sum_{o=1}^O \phi_{o,i}^{(O)} \left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^K \gamma_j\right) \right) + \sum_{o=1}^O \phi_{o,i}^{(O)} E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} \left[\frac{\ln |\boldsymbol{\Lambda}_i|}{2} - \frac{(\mathbf{w}_o - \boldsymbol{\mu}_i)' \boldsymbol{\Lambda}_i (\mathbf{w}_o - \boldsymbol{\mu}_i)}{2} \right] \right) \\ & + E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} [\ln p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)] \end{aligned} \quad (6.8)$$

$$+ \sum_{i=1}^K \left(\sum_{h=1}^H \phi_{h,i}^{(H)} \left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^K \gamma_j\right) \right) + \sum_{h=1}^H \phi_{h,i}^{(H)} \ln \rho_{i,w_h} \right) \quad (6.9)$$

$$- \ln \Gamma\left(\sum_{j=1}^K \gamma_j\right) + \sum_{i=1}^K \ln \Gamma(\gamma_i) - \sum_{i=1}^K (\gamma_i - 1) \left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^K \gamma_j\right) \right)$$

$$-\sum_{i=1}^K \left(\sum_{m=1}^M \phi_{m,i} \ln \phi_{m,i} + \sum_{o=1}^O \phi_{o,i}^{(O)} \ln \phi_{o,i}^{(O)} + \sum_{h=1}^H \phi_{h,i}^{(H)} \ln \phi_{h,i}^{(H)} \right) - \sum_{i=1}^K E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} [\ln q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)] \quad (6.10)$$

We only highlight the derivations for the expressions:

$E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} \left[\frac{\ln |\boldsymbol{\Lambda}_i|}{2} - \frac{(\mathbf{w}_o - \boldsymbol{\mu}_i)' \boldsymbol{\Lambda}_i (\mathbf{w}_o - \boldsymbol{\mu}_i)}{2} \right]$, $E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} [\ln p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)]$ and $E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} [\ln q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)]$. In the variational Bayesian setting, the expression:

$$E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} \left[(\ln |\boldsymbol{\Lambda}_i|)/2 - ((\mathbf{w}_o - \boldsymbol{\mu}_i)^\top \boldsymbol{\Lambda}_i (\mathbf{w}_o - \boldsymbol{\mu}_i))/2 \right]$$

needs to be evaluated in the log likelihood calculation for every video document d to update the free distributions given the current parameter values. The term $\left[\frac{\ln |\boldsymbol{\Lambda}_i|}{2} \right]$ is the normalization factor of the Gaussians and its expectations can cause the log likelihood to be positive. We therefore only evaluate $E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} \left[-((\mathbf{w}_o - \boldsymbol{\mu}_i)^\top \boldsymbol{\Lambda}_i (\mathbf{w}_o - \boldsymbol{\mu}_i))/2 \right]$ for the per document updates and subtract the log of the exponentials of the aggregations as an approximation. We independently derive and mention only the final expressions for the following variables are shown here:

$$E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} [\ln |\boldsymbol{\Lambda}_i|] = \sum_{p=1}^P \psi \left(\frac{\nu_i + 1 - p}{2} \right) + P \ln 2 + \ln |\mathbf{W}_i| \quad (6.11)$$

$$E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} \left[(\mathbf{w}_o - \boldsymbol{\mu}_i)' \boldsymbol{\Lambda}_i (\mathbf{w}_o - \boldsymbol{\mu}_i) \right] = P \kappa_i^{-1} + \nu_i \left((\mathbf{w}_o - \mathbf{m}_i)' \mathbf{W}_i (\mathbf{w}_o - \mathbf{m}_i) \right) \quad (6.12)$$

$$E_{q[\boldsymbol{\mu}, \boldsymbol{\Lambda}]} [\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \sum_{i=1}^K \left\{ \frac{1}{2} \ln \hat{\boldsymbol{\Lambda}}_i + \frac{P}{2} \ln \frac{\kappa_i}{2\pi} - \frac{P}{2} - H[q(\boldsymbol{\Lambda}_i)] \right\} \quad (6.13)$$

$$H[q(\boldsymbol{\Lambda}_i)] = -\ln Z(\mathbf{W}_i, \nu_i) - \frac{(\nu_i - P - 1)}{2} \ln \hat{\boldsymbol{\Lambda}}_i + \frac{\nu_i P}{2}, \text{ where} \quad (6.14)$$

$$\diamond Z(\mathbf{W}_i, \nu_i) = |\mathbf{W}_i|^{-\nu_i/2} \left(2^{\nu_i P/2} \pi^{P(P-1)/4} \prod_{p=1}^P \Gamma \left(\frac{\nu_i + 1 - p}{2} \right) \right)^{-1}$$

$$\diamond \ln \hat{\boldsymbol{\Lambda}}_i = E_q [\ln |\boldsymbol{\Lambda}_i|] = \sum_{p=1}^P \psi \left(\frac{\nu_i + 1 - p}{2} \right) + P \ln 2 + \ln |\mathbf{W}_i|$$

Note that Ψ is the digamma function. For the expression

$\sum_{i=1}^K E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} [\ln p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)]$, we have:

$$\begin{aligned} \sum_{i=1}^K E_{q[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]} [\ln p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)] &= \frac{1}{2} \sum_{i=1}^K \left\{ P \ln \left(\frac{\kappa_0}{2\pi} \right) + \ln \hat{\boldsymbol{\Lambda}}_i - \frac{\kappa_0 P}{\kappa_i} - \kappa_0 \nu_i (\mathbf{m}_i - \mathbf{m}_0)' \mathbf{W}_i (\mathbf{m}_i - \mathbf{m}_0) \right\} \quad (6.15) \\ &+ K \ln Z(\mathbf{W}_0, \nu_0) + \frac{\nu_0 - P - 1}{2} \sum_{i=1}^K \ln \hat{\boldsymbol{\Lambda}}_i - \frac{1}{2} \sum_{i=1}^K \nu_i \text{Tr}(\mathbf{W}_0^{-1} \mathbf{W}_i) \end{aligned}$$

Using the lower bound \mathcal{L}_{MMG} , the ML estimations of the hidden variables in video document d can be obtained using Lagrange Multipliers on $\phi^{(H)}$, $\phi^{(O)}$ and ϕ as follows:

$$\phi_{d,h,i}^{(H)} \propto \exp \left\{ \psi(\gamma_{d,i}) - \psi \left(\sum_{j=1}^K \gamma_{d,j} \right) + \ln \rho_{i,w_d,h} \right\} \quad (6.16)$$

$$\phi_{d,o,i}^{(O)} \propto \exp \left\{ \psi(\gamma_{d,i}) - \psi\left(\sum_{j=1}^K \gamma_{d,j}\right) + E_{q[\mu_i, \Lambda_i]} \left[(\ln |\Lambda_i|)/2 - ((\mathbf{w}_o - \boldsymbol{\mu}_i)' \Lambda_i (\mathbf{w}_o - \boldsymbol{\mu}_i))/2 \right] \right\} \quad (6.17)$$

$$\phi_{d,m,i} \propto \exp \left\{ \psi(\gamma_{d,i}) - \psi\left(\sum_{j=1}^K \gamma_{d,j}\right) + \ln \beta_{i,w_{d,m}} \right\} \quad (6.18)$$

$$\gamma_{d,i} = \alpha_i + \sum_{m=1}^{M_d} \phi_{d,m,i} + \sum_{o=1}^{O_d} \phi_{d,o,i}^{(O)} + \sum_{h=1}^{H_d} \phi_{d,h,i}^{(H)} \quad (6.19)$$

Similarly, for the Corr-MMGLDA model:

$$\begin{aligned} \mathcal{L}_{Corr-MMG} &= E_q[\ln p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\ln p(\mathbf{y}_M|O)] + E_q[\ln p(\mathbf{w}_M|\mathbf{z}_{\mathbf{y}_M}, \beta)] \\ &\quad + E_q[\ln p(\mathbf{z}_O|\boldsymbol{\theta})] + E_q[\ln p(\mathbf{w}_O|\mathbf{z}_O, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + E_q[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{m}_0, \mathbf{W}_0, \kappa_0, \nu_0)] \\ &\quad + E_q[\ln p(\mathbf{z}_H|\boldsymbol{\theta})] + E_q[\ln p(\mathbf{w}_H|\mathbf{z}_H, \boldsymbol{\rho})] - E_q[\ln q(\boldsymbol{\theta}, \mathbf{y}_M, \mathbf{z}_O, \mathbf{z}_H, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \end{aligned} \quad (6.20)$$

For the Corr-MMGLDA model, $E_{q(\mathbf{Z}, \mathbf{Y})}[\ln p(\mathbf{w}_M|\mathbf{z}_{\mathbf{y}_M}, \beta)]$ expands out to be:

$$\sum_{m=1}^M \sum_{i=1}^K \left(\sum_{o=1}^O \phi_{m,o} \phi_{o,i}^{(O)} \right) \ln \beta_{i,w_m} \quad (6.21)$$

Also,

$$E_{q(\mathbf{Y})}[\ln q(\mathbf{y}_M|\boldsymbol{\phi}_{\mathbf{y}_M})] = \sum_{m=1}^M \sum_{o=1}^O \phi_{m,o} \ln \phi_{m,o} \quad (6.22)$$

and $E_q[\ln p(\mathbf{y}_M|O)]$ is constant for all m in d . Equation (6.21) is a computational bottleneck because finding the confidence of the word w_m on topic i necessitates the elimination of uncertainties of w_m 's dependence on \mathbf{w}_o and \mathbf{w}_o 's dependence on topic i . This is also a strong point since the marginalization suggests a stronger influence of a topic on a summary word if that influence is justified by most \mathbf{w}_o s.

Using a similar lower bound $\mathcal{L}_{Corr-MMG}$ for Corr-MMGLDA, the ML estimations of the hidden variables in video d can be obtained as follows (using Lagrange Multipliers on $\phi^{(H)}$, $\phi^{(O)}$ and ϕ):

$$\phi_{d,h,i}^{(H)} \propto \exp \left\{ \psi(\gamma_{d,i}) - \psi\left(\sum_{j=1}^K \gamma_{d,j}\right) + \ln \rho_{i,w_{d,h}} \right\} \quad (6.23)$$

$$\begin{aligned} \phi_{d,o,i}^{(O)} \propto \exp \left\{ \psi(\gamma_{d,i}) - \psi\left(\sum_{j=1}^K \gamma_{d,j}\right) + E_{q[\mu_i, \Lambda_i]} \left[\frac{\ln |\Lambda_i|}{2} - \frac{(\mathbf{w}_o - \boldsymbol{\mu}_i)' \Lambda_i (\mathbf{w}_o - \boldsymbol{\mu}_i)}{2} \right] \right. \\ \left. + \sum_{m=1}^{M_d} \phi_{d,m,o} \ln \beta_{z_{y_m}, w_{d,m}} \right\} \end{aligned} \quad (6.24)$$

$$\phi_{d,m,o} \propto \exp \left\{ \sum_{i=1}^K \phi_{d,o,i}^{(O)} \ln \beta_{i,w_{d,m}} \right\} \quad (6.25)$$

$$\gamma_{d,i} = \alpha_i + \sum_{o=1}^{O_d} \phi_{d,o,i}^{(O)} + \sum_{h=1}^{H_d} \phi_{d,h,i}^{(H)} \quad (6.26)$$

The algorithms for updating the latent variables and the parameters (Section 6.3.2) are given in Section 6.8.4.

6.3.2 Model Parameter Estimation

Before deriving the expressions for the maximum a posteriori and maximum likelihood estimates of the parameters of the proposed models using moment matching (Section 6.3.1) and derivatives w.r.t the parameters of the functional $\mathcal{L}(\cdot)$ let us define the following quantities for each topic i :

$$\begin{aligned} N_i &= \sum_{d=1}^D \sum_{o=1}^{O_d} \phi_{d,o,i}^{(O)}; & \bar{\mathbf{x}}_i &= \frac{1}{N_i} \sum_{d=1}^D \sum_{o=1}^{O_d} \phi_{d,o,i}^{(O)} \mathbf{w}_{d,o} \\ \mathbf{S}_i &= \frac{\sum_{d=1}^D \sum_{o=1}^{O_d} \phi_{d,o,i}^{(O)} (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i)(\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i)'}{N_i} \end{aligned} \quad (6.27)$$

Through a fuller Bayesian treatment and using moment matching techniques on Equ. 6.3, we obtain the following parameter updates of the prior distributions over the $2K$ Gaussian parameters of the model for each topic i :

$$\kappa_i = \kappa_0 + N_i \quad (6.28)$$

$$\mathbf{m}_i = \frac{1}{\kappa_i} (\kappa_0 \mathbf{m}_0 + N_i \bar{\mathbf{x}}_i) \quad (6.29)$$

$$\mathbf{W}_i^{-1} = \mathbf{W}_0^{-1} + N_i \mathbf{S}_i + \frac{\kappa_0 N_i}{\kappa_0 + N_i} (\bar{\mathbf{x}}_i - \mathbf{m}_0)(\bar{\mathbf{x}}_i - \mathbf{m}_0)' \quad (6.30)$$

$$\nu_i = \nu_0 + N_i \quad (6.31)$$

Further, using some algebraic manipulations and utilizing Lagrange Multipliers for β and ρ for each topic i , we obtain:

For the MMGLDA model:

$$\rho_{i,j} \propto \sum_{d=1}^D \sum_{h=1}^{H_d} \sum_{j=1}^{corrV_H} \phi_{d,h,i}^{(H)} \delta(w_{d,h}, j) \quad (6.32)$$

$$\beta_{i,j} \propto \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{j=1}^V \phi_{d,m,i} \delta(w_{d,m}, j) \quad (6.33)$$

For the Corr-MMGLDA model:

$$\rho_{i,j} \propto \sum_{d=1}^D \sum_{h=1}^{H_d} \sum_{j=1}^{corrV_H} \phi_{d,h,i}^{(H)} \delta(w_{d,h}, j) \quad [\text{same as MMGLDA}] \quad (6.34)$$

$$\beta_{i,j} \propto \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{j=1}^V \left(\sum_{o=1}^O \phi_{d,m,o} \phi_{d,o,i}^{(O)} \right) \delta(w_{d,m}, j) \quad (6.35)$$

To optimize the α parameters, we follow the corresponding expressions in [Blei et al., 2003] and optimize using Newton’s iterative gradient based method. Optimizing α_i is dependent on the value of α_j through: .

$$\begin{aligned}\frac{\partial \mathcal{L}_{(\cdot)}}{\partial \alpha_i} &= D(-\Psi(\alpha_i) + \Psi(\sum_{j=1}^K \alpha_j)) + \sum_{d=1}^D (\Psi(\gamma_{d,i}) - \Psi(\sum_{j=1}^K \gamma_{d,j})) \\ \frac{\partial \mathcal{L}_{(\cdot)}}{\partial \alpha_i \alpha_j} &= \partial(i, j) D \left(\Psi'(\sum_{j=1}^K \alpha_j) - \Psi'(\alpha_i) \right)\end{aligned}\quad (6.36)$$

When α is symmetric i.e., all of the K components of α are same then there is just a single value to be optimized that has been accumulated over all topics

Keyword prediction: For predicting a bag of words summary from an ensemble of low level features of video document d and the learnt $p(w_v | z_v = k, \beta)$, we permute the vocabulary V for the new test video as:

$$p(w_v | \mathbf{w}_O, \mathbf{w}_H) \approx \sum_{o=1}^O \sum_{i=1}^K \phi_{d,o,i}^{(O)} p(w_v | \beta_i) + \sum_{h=1}^H \sum_{i=1}^K \phi_{d,h,i}^{(H)} p(w_v | \beta_i) \quad (6.37)$$

6.4 Experimental Setup and Results

Here we briefly mention the descriptors that we use to represent the videos. To represent actions, we use features known as Histogram of Oriented Gradients in 3D (HOG3D) [Kläser et al., 2008]. The gradient directions are binned by mapping them to 10 polar meridian and 6 polar parallel planes and then treating half spaces to be equivalent. We resized the video frames such that the largest dimension (height or width) was 160 pixels, and extracted HOG3D features from a dense sampling of frames. Our HOG3D parameters resulted in a 300-dimensional feature vector using support volumes of dimension $2 \times 2 \times 5$ and 5×3 polar co-ordinate bins. We then use K-means clustering to create a 1000-word codebook following [Bilinski and Bremond, 2011] from a random sampling of the training data.

Color histogram features are also used as part of the discrete visual data. We use 512 RGB color bins and histograms are computed on densely sampled frames. Due to large deviations in the extremities of the color spectrum, we use the histogram between the 15th and 85th percentiles averaged across a video and counts normalized to lie in [1,100].

Finally we use Object Banks [Li et al., 2010b] for a histogram pattern of positive object detections. OB transforms an image into a 44604 dimensional concatenated feature vector for each of the 177 *off-the-shelf* object detectors that are currently used. Each entry within a 252 dimensional detection feature vector represents the distance from the decision hyperplane midway within the margins for different scale-space transformations of the image. The object labels in OB cover only about 10% of the summary words (246 out of 2687 for the training set and 166 out of 1219 for the Dev-T set). Keyframes used for these features are extracted using the change in color histogram method [Zhang et al., 1995] and the positive OB responses are quantized following classes in [Torresani et al., 2010]. Thus \mathbf{w}_H in Figs. 6.4b, 6.4d and 6.4e consists of codebook histograms from HOG3D, color and OB. Needless to say, the contributions of these *off-the-shelf* object detectors are not significant at all.

The real valued features we use in our video representation are those representing scenes as mentioned in [Oliva and Torralba, 2001]. The scene property by itself induces image summarization in a

way that is consistent with human perception of vision [Oliva and Torralba, 2006]. A set of perceptual dimensions is proposed along the boundary viewpoint (e.g. depth, openness, expansion, perspective) and along the content viewpoint (e.g. naturalness, roughness, ruggedness, etc.) which represent the dominant *spatial structure* of a scene. These features are named GIST features as is common parlance in computer vision literature. To compute these features, we have used the setup in [Douze et al., 2009] leading to a 960-dimensional descriptor for each frame. We calculate GIST features for every 10^{th} frame.

To save computational time, the GIST features are projected into lower dimensions using Principle Component Analysis (PCA). PCA is done on the training data across all event categories to remove the dependence of the visual descriptors on specific events. We first visualize the lower (15, 30 and 60) dimensional GIST features in two dimensions using t-statistic based stochastic neighbor embedding (t-SNE) [Maaten and Hinton, 2008], however, the separations look more or less the same and do not yield conclusive evidence of choosing the right number of dimensions (Fig. 6.5).

By inspecting the plots from t-SNE, we choose 15 dimensions and validate the choice by both manually inspect-

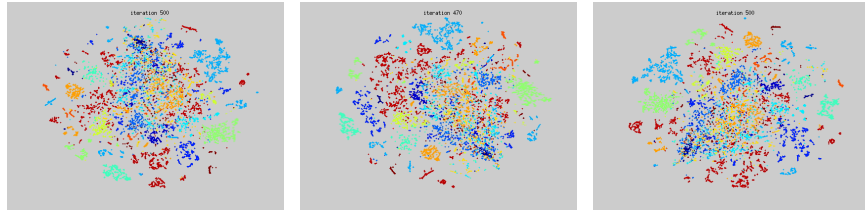


Figure 6.5: GIST features projected on to 15 (left), 30 (middle) and 60 (right) dimensions & visualized in two dimensions using t-SNE [Maaten and Hinton, 2008]

ing the eigenvalues and experimentally cross-validating with the baseline Corr-MGLDA topic model (Fig. 6.4c). 30 or 60 dimensional features decreased the ELBO of the model. We do not select further lower dimensions based on significance of the eigenvalues. Each w_o in Figs. 6.4c, 6.4d and 6.4e represents a frame in 15 dimensions corresponding to a GIST feature vector.

6.4.1 Held-out Log Likelihoods and Topics

In this section, we evaluate the topic models in terms of ELBO on the held-out Dev-T set acting as a test set (with the human summaries) for posterior inference and as a prediction set (without human summaries) for BoW summary generation. Multinomial parameters are seeded and Gaussian parameters are randomly initialized.

The base measures of α are initialized to 0.1 and normalized while its concentration parameter is set to 10. An issue with the real valued features is the influence of data normalization on the ELBOs from the topic models. We have observed that when the data is not normalized to lie within $[0,1]^P$, the sequence of ELBOs from Corr-MGLDA during EM often indicate suboptimality even during training. The ‘‘PDS’’ suffix (in the table and all other figures) means ‘‘Positive Data Scaling’’ i.e. each real valued vector is sum-normalized to $[0,1]^P$ independently.

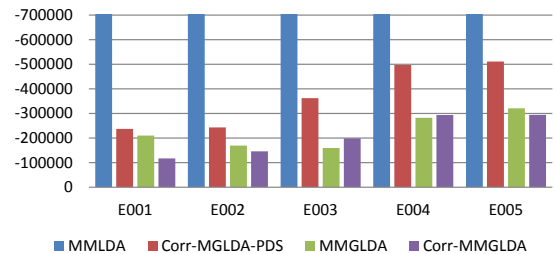


Figure 6.6: Test ELBOs on events E001-E005 in the Dev-T set. Lower is better.

Model	Topic 1	Topic 2	Topic 3	
Corr-MMGLDA	wed couple ceremony church ring footage exchange bride groom helicopter	wed ceremony bride groom church flower exchange ring walk kiss outdoors	wed Hawai US beach guest place footage scene ring minister lei Kailua	Wedding ceremony
Corr-MGLDA-PDS	wed ceremony couple bride groom church flower exchange vow	wed ceremony bride groom couple flower church man outdoors vow	wed ceremony bride groom couple flower church man outdoors vow	
Corr-MMGLDA	skateboard trick jump guy outdoors park skate per- form rail attempt ramp ol- lie	snowboard people per- form hill jump group footage trick camera helmetmount	surf surfboard jump fall water outdoors man boat wave ride tow waterboard	Boarding event
Corr-MGLDA-PDS	skateboard trick jump guy outdoors park skate per- form attempt boy rail	skateboard trick jump guy outdoors park skate per- form attempt boy rail	skateboard trick jump out- doors guy park skate per- form boy attempt rail	

Table 6.2: Three latent topics for two events from proposed five-topic Corr-MMGLDA and Corr-MGLDA-PDS models. The topics from Corr-MGLDA-PDS are similar as a result of high values of α_k obtained after running Corr-MGLDA-PDS on scaled i.e. normalized data. The topics from Corr-MMGLDA are qualitatively far superior and indicates sub-events of the “Wedding ceremony” and the “Boarding” events

The PDS normalization fixes this problem and raises ELBOs for Corr-MGLDA significantly but convergence is slower. However in the latter setting, the values of α_k become very large which destroys sparsity in topics. This is possibly due to strong overlap of modes within the $[0, 1]^P$ hypercube where one dimension is severely correlated with the others. Examples of such topics on the “Wedding Ceremony” and the “Boarding” events are given in Table 6.2 where all topics are almost alike and lose subjective interpretability.

The new topic models with both Multinomial and Gaussian distributions on the video features do not suffer from the data scaling problem. It is possible that the mean parameter space for the tractable distributions over both discrete and real valued observations prevents co-ordinate ascent steps to dwell in suboptimal regions that could arise out of extreme values in the real valued data alone. Although the Normal-Wishart priors act as regularizers, automatically tuning \mathbf{W}_0 using another level of priors or from the data itself is not used here. In general optimization with tractable distributions and parameter constraints (e.g. non-negativity, boundedness and positive definiteness) can be non-convex [Wainwright and Jordan, 2008].

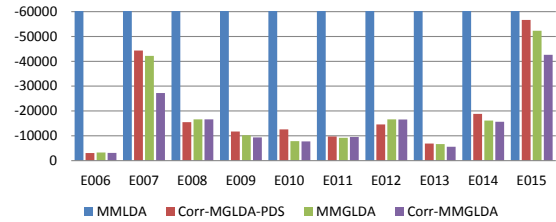


Figure 6.7: Test ELBOs on events E006-E015 from Dev-T set. Lower is better

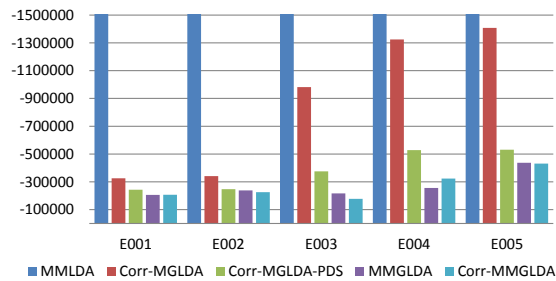


Figure 6.8: Prediction ELBOs on first 5 event for Dev-T set. Lower is better.

Figures 6.6 and 6.7 show the test ELBOs of MMGLDA and Corr-MMGLDA versus the MMLDA

model and the Corr-MGLDA model with PDS. The ELBOs for MMLDA are off the charts (at least three to four times the cut-off shown in the graphs). For the first 5 events, the videos contain positive instances of the events in Dev-T set. For this subset of events, the MMGLDA family of models outperform the best version of Corr-MGLDA in terms of ELBOs (i.e. with PDS). Figures 6.6 and 6.7 are obtained using $K=20$ topics— K being set through 5-fold cross-validation. For the last 10 events (Fig. 6.7), the videos contain only related instances of the events in Dev-T set—dissimilar to the training configuration i.e. the annotators are unsure about the relevance of the videos to the event category. In this case, Corr-MGLDA-PDS do not perform worse in general since the GIST features are *global* features [Oliva and Torralba, 2001].

The prediction performance on the first 5 events is shown in terms of ELBO in Fig. 6.8 for the same value of K . Fig. 6.8 shows that MMLDA does not perform well in terms of word prediction ELBO measure. We can also see the effects of sub-optimality when PDS is suppressed for Corr-MGLDA (Corr-MGLDA in Fig. 6.8). MMGLDA and Corr-MMGLDA again perform comparably and outperforms Corr-MGLDA-PDS on the first 5 events except E002—“feeding an animal”—a very complex event for computer vision.

For events 6 through 15, the prediction ELBO graphs also look very similar to that in Fig. 6.7 as shown in Fig. 6.9. PDS on our proposed MMGLDA family shows even better ELBOs, but topic sparsity problems mitigate only a little and we do not report those here. All these experiments are run using \mathbf{m}_0 set to $\mathbf{0}$, \mathbf{W}_0 set to a broader prior \mathbf{I} , the identity matrix, ν_0 set to P and κ_0 set to 1. Normalizing the data to lie in $[0,1]^P$ with \mathbf{I} as priors for Λ_k s leads to sharing of topic responsibilities of the real valued data by only a few

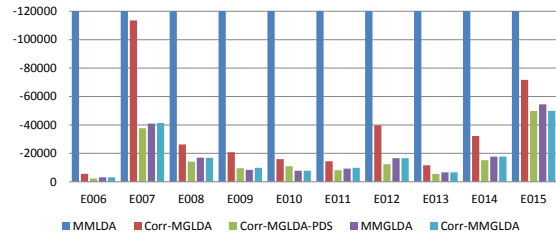


Figure 6.9: Prediction ELBOs on events E006-E015 on Dev-T set—lower is better. Best viewed with magnification

Gaussians thereby contributing much less to the overall log-likelihood. It is also observed that the means of the ELBOs of our proposed models are significantly less negative (i.e. better) at 95% confidence level (using paired t-test) than the existing topic models during cross-validation on the training set. For most events, ELBOs for proposed models with $K=10$ are not statistically worse either and show slightly higher ROUGE-1 scores for some events.

Figure 6.10 shows the macro average of test ELBOs across all the 15 events in the Dev-T set. We omit the line graph for MMLDA as it is out of axis limits. The graphs confirm the superior fit of our proposed models to a natural representation of multimedia (test) data.

6.4.2 Translating Related Words to Videos

Figs. 6.11, 6.12 and 6.13 show how latent topics can first be used to discover most probable related words from unstructured text which can then be translated to most probable frames from one or more videos (and hence the videos themselves). The frames correspond to w_o s in Fig. 6.4 and Table 6.1. We observe from Figs. 6.12 and 6.13 how topics 6 and 10 decompose the “Flash mob gathering” event into its constituent sub-themes. While topic 6 describes flash mob dances in outdoors and near plazas, topic 10 focuses on a flash mob in Hollywood posing in Star Wars costumes and light sabers along with the famous miniature robot R2D2.

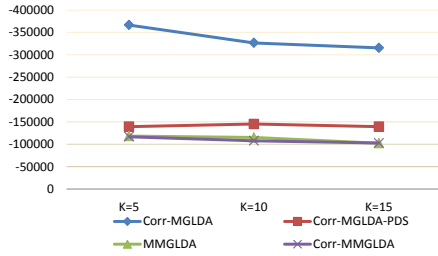


Figure 6.10: Average test ELBOs on all events in the Dev-T set for different topics. **Lower is better**

Topic 10: place fight fake public event gather flash mob music star outdoors war

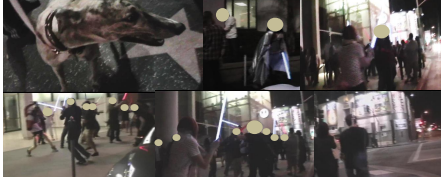


Figure 6.11: Topic 10 for the “Flash mob” event from a ten-topic MMGLDA

Topic 6: mob flash dance people mall outdoors large gather woman public plaza



Figure 6.12: Topic 6 for the “Flash mob” event from a ten-topic Corr-MMGLDA

Topic 10: Hollywood dog star wars robot light saber R2D2 eye lens blvd fight street

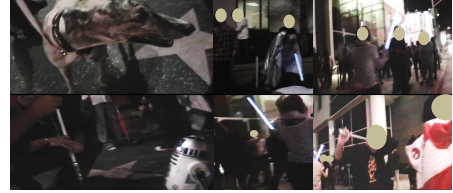


Figure 6.13: Topic 10 for the “Flash mob” event from a ten-topic Corr-MMGLDA

Fig. 6.11 shows the inter-translation of modalities for topic 10 from MMGLDA corresponding to that in Fig. 6.13 from Corr-MMGLDA. Note how the topic loses specificity (e.g. misses “R2D2”, “light,” “saber” within the top few words) and focuses on generality (e.g. flash mob). Topic 6 for MMGLDA is exactly the same as that for Corr-MMGLDA. Table 6.3 shows log of the ratio:

$\frac{\alpha_k}{|\Lambda_k|}$ for the two proposed models and gives us a hint on how “broad” a topic k may be vs. how much variance in the visual summary is it able to capture. A relatively higher value of the ratio means that a topic captures more variance and hence the volume captured by the determinant of the inverse covariance matrix Λ_k , i.e. $|\Lambda_k|$, through its spanning eigenvectors is proportionally less. For MMGLDA, this ratio is always relatively lower in our setting and this means that the model captures more generic patterns first giving rise to a lower $|\Lambda_k|^{-1}$.

Model	$k=6$	$k=10$	$\text{avg}_{k \neq \{6,10\}}$
Corr-MMGLDA	104.628	8.164	40.2398
MMGLDA	104.623	8.102	40.0702

Table 6.3: $\ln \frac{\alpha_k}{|\Lambda_k|}$ values for topics in event 8

The last column in Table 6.3 is the average of the ratios for the other topics from the two models for event eight. The ratios can be large for a more general topic (e.g. topic 6 in Fig. 6.12) owing to a higher α_k too. Although, all values are close due to the use of the broader prior, \mathbf{I} , it is observed that Corr-MMGLDA discovers related words which are qualitatively superior. However, the corresponding most probable frames are almost similar for both models in most cases. Translating related words to frames is best judged manually, but, the ratio we use here can be a viable alternative.

6.4.3 Translating/Summarizing Videos To Text

Table 6.4 reports the ROUGE-1 (henceforth R-1) scores of the predicted 5 to 10 keyword summaries from the different models when compared with the corresponding short human synopses. Sometimes full sentences cannot be generated from the predicted words due to a deficient language model. This and the short nature of human synopses are the primary reasons why we perform only R-1 evaluation. Some examples of the sentences/phrases are shown in Fig. 6.15.

In Table 6.4, OB is the Object Bank baseline (see just ahead of Section 6.3.1)—it confirms the difficulty of detecting objects on this dataset. The quantized OB responses perform poorly since a-priori it is hard to know which object detectors will be needed and existing but irrelevant object detectors can produce an unpredictable pattern of false positives. Creating object models for every genre of data requires expensive annotation efforts. Even if there is a 100% overlap between our training vocabulary and object models, the R-1 scores for OB *may* only increase by 10-folds which is still low.

Purely multinomial topic models showing lower ELBOs can perform quite well in BoW summarization. MMLDA assigns likelihoods based on success and failure of *independent* events and failures contribute highly negative terms to the log likelihoods but this does not indicate the model’s summarization performance where low probability terms are pruned out. Gaussian components can partially remove the *independence* through covariance modeling and fit the data better at the cost of higher time and space complexity. The R-1 scores from MM(G)LDAs are comparable for 5 and 10 keywords with no statistical difference, however, a possible reason for lower R-1 scores for Corr-MMGLDA model is that due to better correspondence to the topic of the GIST energy in the scene, when a topically relevant but non-summary word is chosen upfront, more related but non-summary words are also drawn in. As future research, we also like to do a principled initialization of Gaussian parameter priors as in [Nasios and Bors, 2006]. This translates to better initializations within the multimedia topic modeling platform. We plan to pursue this as part of future research.

However, the high scores with Corr-MGLDA-PDS is entirely co-incidental—the topics are more

	Model	$n=5$	$n=10$	OB
E001	MMLDA	0.182	0.248	0.0*
	Corr-MGLDA-PDS	0.187	0.257	
	MMGLDA	0.179	0.245	
	Corr-MMGLDA	0.139*	0.192*	
E002	MMLDA	0.186	0.249	0.0*
	Corr-MGLDA-PDS	0.182	0.242	
	MMGLDA	0.186	0.237	
	Corr-MMGLDA	0.143*	0.176*	
E003	MMLDA	0.221	0.265	0.012* =1%
	Corr-MGLDA-PDS	0.233	0.263	
	MMGLDA	0.228	0.267	
	Corr-MMGLDA	0.171*	0.230	
E004	MMLDA	0.265	0.302	0.0*
	Corr-MGLDA-PDS	0.263	0.292	
	MMGLDA	0.264	0.321	
	Corr-MMGLDA	0.221	0.247*	
E005	MMLDA	0.167	0.213	0.005* =0.5%
	Corr-MGLDA-PDS	0.180	0.208	
	MMGLDA	0.165	0.205	
	Corr-MMGLDA	0.129*	0.142*	
6-15 Avg.	MMLDA	0.216	0.252	0.001* =0.1%
	Corr-MGLDA-PDS	0.211	0.258	
	MMGLDA	0.210	0.243	
	Corr-MMGLDA	0.179*	0.221	

Table 6.4: Individual and average ROUGE-1 scores on the events—best results from 10/20 latent topics are shown. The value of n represents the top- n most probable keywords. A (*) means significantly **worse** performance at 95% confidence to {MM,MMG}LDAs. These results are only reported for the same hyperparameter settings.

or less uniform and each one covers parts of the sub-events equally. Further, each w_o 's density over those topics is uniform enough to not achieve a reasonable permutation. The same thing happens when PDS is used for our MMGLDA family of models and the summaries completely lose subjective appeal although R-1 scores improve considerably. This is similar to the *qualitatively degenerate approach*—taking the top n frequent words from the event vocabulary and using those as summaries for **every** test video. The scores for MMLDA and MMGLDA are also comparable to this setting. Quantification of the permutation quality has not been done and is left as one of the key aspects to be explored in future research.

Scores in Table 6.4 need to be multiplied by the event classification accuracies to obtain lower bounds for clips having no event labels. The scores become competitive for larger n and much larger K if we topic model on the entire corpus.

6.4.4 Event Classification

Fig. 6.14 shows 5-fold cross-validation on the 15-event training set and also the test accuracies on Dev-T set for event classification. A c -SVM classifier from the libSVM [Chang and Lin, 2011] package is used with default settings for multiclass classification. Although around 50% classification accuracy can be easily achieved using the discrete visual features that we use, higher accuracies can be obtained using better kernels, fusion of classifiers and optimizing Detection-Error-Tradeoff curves while cross validating [Natarajan et al., 2011]. However, these discussions are outside the scope of this thesis (see for example [Perera et al., 2012]).

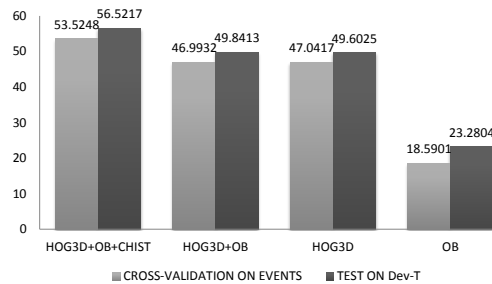


Figure 6.14: Event detection accuracies for cross-validation (light gray bars) and test (dark gray bars) with different features

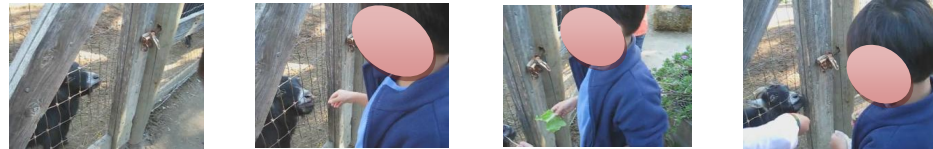
6.4.4.1 NATURAL LANGUAGE GENERATION

To translate a video into multiple sentences from predicted keywords, we use an ordered-sequence template as <subject, verb, object, preposition, scene-noun>. Language models from the data at hand is used to prune impossible sequences. The subjects, objects, verbs and nouns extracted from the training synopses using dependency grammars and POS models appear in each generated sentence only once.

We use the parser in [Klein and Manning, 2003] to score the sequence of words following the template. The sentences are ordered according to bigram and parse tree scores. When complete sentences cannot be generated due to a deficient language model, we output possible bigrams and trigrams (see E004 in Fig. 6.15). Similar corpus based sentence generation techniques can be found in [Yang et al.,



Bag of words: skateboard trick skater indoor jump perform young skate olly man staircase snowboard skateboarder clip ollie
Sentences: Men perform tricks while skateboarding. Men performing tricks on skateboards. Skaters perform jumps on skateboards. Men perform ollies on skateboards.
[Event E001 - board trick] Actual Summary: indoor ollies



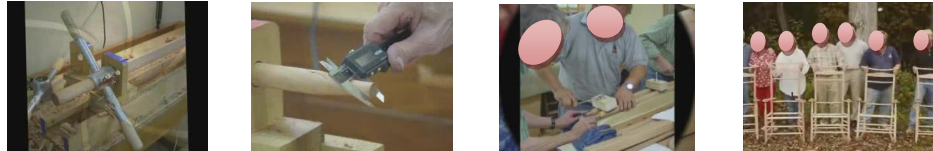
Bag of words: feed bird food outdoors woman eat hand leaf daytime boy giraffe zoo cup girl goat
Sentences: Boys feed birds by hand. Girl feed birds by hand. Girl eats food in zoo. Woman feed birds by hand. Woman feeds goat in zoo.
[Event E002 - Feeding animal] Actual Summary: little boy feeds goat



Bag of words: fish noodle man sea boat bare land big stretch person catch hand stream catfish
Sentences: Men catch fish on boat. Men catch fish by hand. Men catch fish in stream. Men catch fish in boat.
[Event E003 - Landing a fish] Actual Summary: catching big fish off dock



Bag of words: church wed ceremony inside bride groom aisle dress doughnut couple clap hug kiss sign
Sentences: Could not generate sentence – current template and language model is insufficient, but phrases are found like “wedding ceremony”, “couple kissing”, “bride and groom”
[Event E004 - Wedding] Actual Summary: wedding ceremony in church



Bag of words: wood man cut make mill tree piece chainsaw outdoors large wooden guitar automatic lathe
Sentences: Men make cuts to wood.
[Event E005 - Woodworking] Actual Summary: people building wooden chairs

Figure 6.15: Bag of keywords and sentence translations from our proposed MMGLDA ($K=20$) for some clips from the first five events from the Dev-T set

2011, Gupta et al., 2012] but NLG is a research topic in its own right⁴.

6.5 A Thousand Frames in just a Few Sentences – Enhancing Summary Relevancy

In this section, we focus on improving the unigram recall score of the final summaries generated from the keywords predicted using low level video features. Since combining keywords into sentences using a pre-specified grammar template does not increase recall in terms of unigram coverage, we form the hypothesis that there must be some sentences in the training set which also summarize the test videos better than just the predicted keywords.

From a psychoanalytic point of view, our motivation stems from the following: We can imagine a comedian trying to impress an audience. Often times some narrative is quoted as is and then mocked at (sometimes using another contradictory narrative) This implies that the comedian is using a sample from the training set to maximize relevance (measured as the intensity of applause from the audience) for a new scenario.

The problem of generating natural language descriptions of images and videos has been steadily gaining prominence in the computer vision community. The problem is important for three reasons: i) transducing visual data into textual data would permit well understood text-based indexing and retrieval mechanisms essentially *for free*; ii) fine grained object models and region labeling introduce a new level of semantic richness to multimedia retrieval techniques; and iii) grounding representations of visual data in natural language has great potential to overcome the inherent semantic ambiguity prominent in the data-driven high-level vision community (see [Torralba and Efros, 2011] for a discussion of data-set bias and discussion on the different *meanings* common labels can have within and across data sets).

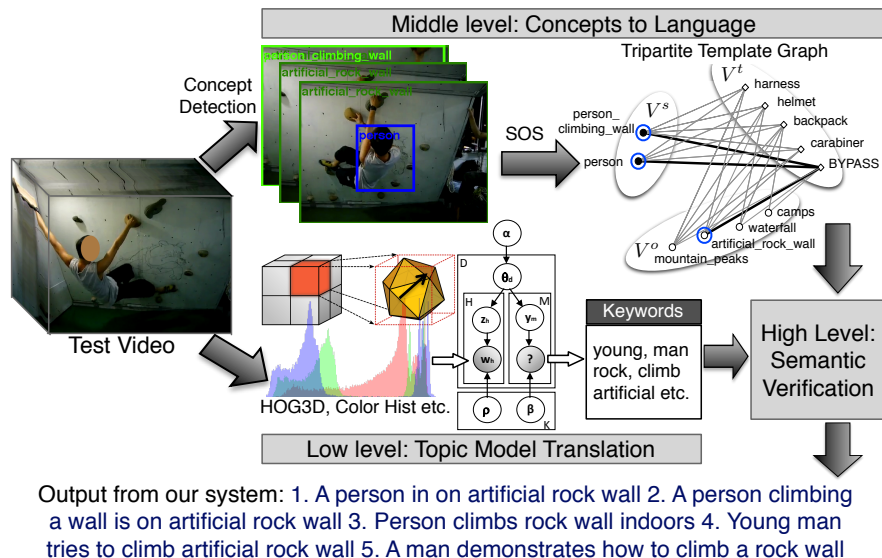


Figure 6.16: A framework of our hybrid system showing a test video being processed and linguistically described through top-down concept detection and bottom-up keyword summarization

⁴http://aclweb.org/aclwiki/index.php?title=Downloadable_NLG_systems

The video to text processing pipeline of our system is shown in Fig. 6.16. A test video with its event category information (here it is “rock climbing”) is processed in three ways—1) in a bottom up fashion where a multimodal latent topic model, using only low level features, efficiently predicts some fixed number of keywords acting as a proxy to the caption of the test video; 2) in a top down fashion where a multitude of frame level concept detections are pruned across the entire length of the video to obtain a smaller relevant set and 3) a high level semantic verification of the predicted caption keywords through the detected concepts that focuses on increasing the information need captured in well formed English sentences which are the final output of our system. To date, the most common approach to such lingual description of images has been to model the joint distribution over low-level image features and language, typically nouns. Early work on multimodal topic models by Blei et al. [Blei and Jordan, 2003] and subsequent extensions [Putthividhya et al., 2010, Feng and Lapata, 2010b, Wang et al., 2009, Cao and Fei-Fei, 2007] jointly model image features (predominantly SIFT and HOG derivatives) and language words as mixed memberships over latent topics with considerable success. Other non-parametric nearest-neighbor and label transfer methods, such as Makadia et al. [Makadia et al., 2008] and TagProp [Guillaumin et al., 2009], rely on large annotated sets to generate descriptions from similar samples. These methods have demonstrated a capability of lingual description on images at varying levels, but they have two main limitations. Being based on low-level features and/or similarity measures, first, it is not clear they can scale up as the richness of the semantic space increases. Second, the generated text has largely been in the form of word-lists without any semantic verification (see Section 6.5.1.3).

Alternatively, a second class of approaches to lingual description of images directly seeks a set of high-level concepts, typically objects but possibly others such as scene categories. Prominent among object detectors are Object Bank (OB) [Li et al., 2010b] and the related deformable parts model (DPM) [Felzenszwalb et al., 2010] which have been successful in the task of “annotating” natural images. Despite being able to guarantee the semantic veracity of the generated lingual description, these methods have found limited use due to the overall complexity of object detection *in-the-wild* and its constituent limitations (i.e., noisy detection), and the challenge of enumerating all relevant world concepts and learning a detector for them (although recent work in propagation on Image Net [Kuettel et al., 2012] shows potential to overcome this hurdle).

Our method does not suffer from the lack of semantic verification that the classical low-level models do, nor does it suffer from the tractability challenges of the high-level concept methods—it can rely on fewer well-trained concept detectors for verification allowing the correlation between different concepts to replace the need for a vast set of concept detectors.

We use multimodal latent topic models to find a proposal distribution over some training vocabulary of textual words (see Fig. 6.16 for an overview of the system), then select the most probable words as potential subjects, objects and verbs through a natural language dependency grammar and part-of-speech tagging on a training set of sentences. We train and run state-of-the-art DPM concept detectors such as “artificial rock wall,” “person climbing wall” etc. We convert detected key-concepts from the middle layer into lingual descriptions through a tripartite template graph, which encodes the relations between concepts. Finally, we enhance the lingual descriptions by selecting similar training sentences through concept verification of the predicted keywords. Currently our semantic verification step is independent of any computer vision framework and works by measuring the number of inversions between two ranked lists of predicted keywords and detected concepts both being conditional on their respective learned topic multinomials.

Images vs. Videos: Recent work in [Berg et al., 2012, Kulkarni et al., 2011, Farhadi et al., 2010, Yang et al., 2011] is mainly focused on generating more verbose and fluent descriptions of a single image—images and not videos. Videos introduce an additional set of challenges, such as temporal variation/articulation and dependencies. Most related work in vision has focused only on the activity classification side: example methods using topic models for activities are the hidden topic markov model [Wanke et al., 2010], semi-latent topic models [W. and M., 2009] and frame-by-frame Markovian topic models [Hospedales et al., 2009], but these methods do not model language and visual topics jointly. A recent activity classification paper of relevance is the Action Bank method [Sadanand and Corso, 2012], which ties high-level actions to constituent low-level action detections, but it does not include any language generation framework.

The two most relevant works to ours are the Khan et al. [Khan et al., 2011] and Barbu et al. [Barbu et al., 2012] systems. Both of these methods extract high-level concepts, such as faces, humans, tables, etc., perform tracking through time and generate language description by template filling. This method relies directly on all high-level concepts being enumerated (the second class of methods introduced above) and hence may be led astray by noisy detection and has a limited vocabulary, unlike our approach which not only uses the high-level concepts but augments them with a large corpus of textual synopses from the bottom-up. Furthermore, their datasets have simpler videos not *in-the-wild*.

We, on the other hand, focus on obtaining the lingual descriptions of general videos (e.g., from YouTube) *directly* through bottom-up visual feature translations to text and top-down concept detections. We leverage both detailed object annotations and loose human lingual synopses. Our proposed hybrid method shows much better relevant content generation over simple keyword annotation of videos alone as observed using quantitative evaluation on two datasets—the TRECVID dataset from NIST and a new in-house dataset consisting of cooking videos collected from YouTube with human lingual descriptions generated through Amazon’s Mechanical Turk (Section 6.5.2).

6.5.1 An Information Extraction and Summarization System Framework

6.5.1.1 LOW LEVEL: TOPIC MODEL

We adapt the GM-LDA model in [Blei and Jordan, 2003] (dubbed MMLDA – short for MultiModalLDA in this paper) to handle a discrete visual feature space. The original model in [Blei and Jordan, 2003] is defined in the continuous space, which presents challenges for the discrete visual features we use (e.g., HOG derivatives [Kläser et al., 2008]), that can become unstable during deterministic approximate optimization due to extreme values in high-dimensions and its inherent non-convexity [Wainwright and Jordan, 2008].

We briefly revisit the MMLDA model (see Chapters 1 and 4) and demonstrate how it is instantiated and differs from the original version in [Blei and Jordan, 2003]. First, we use an asymmetric Dirichlet prior, α for the document level topic proportions θ_d following [Wallach et al., 2009] unlike the symmetric one in [Blei and Jordan, 2003]. In Fig. 6.4, D is the number of documents, each consisting of video and textual synopsis (the text is only available during training). The number of discrete visual words and discrete textual words per video document d are N and M . The parameters for corpus level topic multinomials over visual words are $\rho_{1:K}$. The parameters for corpus level topic multinomials over

textual words are $\beta_{1:K}$ —only the training instance of this parameter is used while keyword prediction. The indicator variables for choosing a topic are $\{z_{d,n}\}$ and $\{y_{d,m}\}$; $w_{d,m}$ is the for text word at position m in video “document” d with vocabulary size V . Each $w_{d,n}$ is a visual feature from a bag-of-discrete-visual-words at position n with vocabulary size $corrV$ and each $w_{d,n}$ represents a HOG3D [Kläser et al., 2008], OB, color histogram, transformed color histogram [van de Sande et al., 2010] etc.

We use the mean field method of optimizing a lowerbound to the true likelihood of the data. A fully factorized q distribution with “free” variational parameters γ , ϕ and λ is imposed by: $q(\theta, \mathbf{z}, \mathbf{y} | \gamma, \phi, \lambda) =$

$$\prod_{d=1}^D q(\theta_d | \gamma_d) \left[\prod_{n=1}^{N_d} q(z_{d,n} | \phi_{d,n}) \prod_{m=1}^{M_d} q(y_{d,m} | \lambda_{d,m}) \right] \quad (6.38)$$

The optimal values of free variables and parameters are found by optimizing the lower bound on $\ln p(\mathbf{w}_M, \mathbf{w}_N | \alpha, \beta, \rho)$. The free multinomial parameters of the variational topic distributions ascribed to the corresponding data are ϕ_d s. The free parameters of the variational word-topic distribution are λ_d s. The surrogate for the K -dimensional α is γ_d which represents the expected number of observations per document in each topic k . The optimal value expressions of the hidden variables in video document d for the MMLDA model are as follows:

$$\phi_{d,n,k} \propto \exp \{ \psi(\gamma_{d,k}) + \ln \rho_{k,w_{d,n}} \} \quad (6.39)$$

$$\lambda_{d,m,k} \propto \exp \{ \psi(\gamma_{d,k}) + \ln \beta_{k,w_{d,m}} \} \quad (6.40)$$

$$\gamma_{d,k} = \alpha_k + \sum_{n=1}^{N_d} \phi_{d,n,k} + \sum_{m=1}^{M_d} \lambda_{d,m,k} \quad (6.41)$$

where Ψ is the digamma function. The expressions for the maximum likelihood of the topic parameters are:

$$\rho_{k,j} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{j=1}^{corrV} \phi_{d,n,i} \delta(w_{d,n}, j) \quad (6.42)$$

$$\beta_{k,j} = \sum_{d=1}^D \sum_{m=1}^{M_d} \sum_{j=1}^V \lambda_{d,m,i} \delta(w_{d,m}, j) \quad (6.43)$$

The asymmetric α is optimized using the formulations given in [Blei et al., 2003] using gradient ascent using Newton steps as search directions.

A strongly constrained model, Corr-LDA, is also introduced in [Blei and Jordan, 2003] that uses real valued visual features and shows promising image annotation performance. We experiment with the model to use our discrete visual feature space (and name it Corr-MMLDA) but finally opted to not use it in our final experiments due to the following reasons.

The correspondence between $w_{d,m}$ and $z_{d,n}$ necessitates checking for correspondence strengths over all possible dependencies between $w_{d,m}$ and $w_{d,n}$ multiplied by that between $w_{d,n}$ and

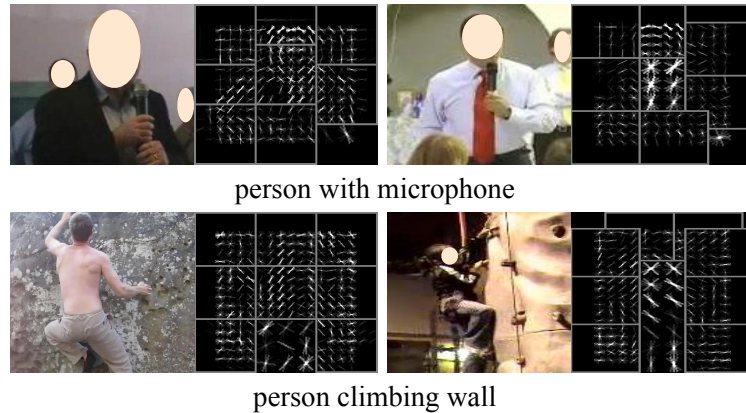


Figure 6.18: Examples of DPM based concept detector.

$z_{d,n}$. This assumption is relaxed in the MMLDA model and removes the bottleneck in run-time efficiency for high dimensional video features without showing significant performance drain.

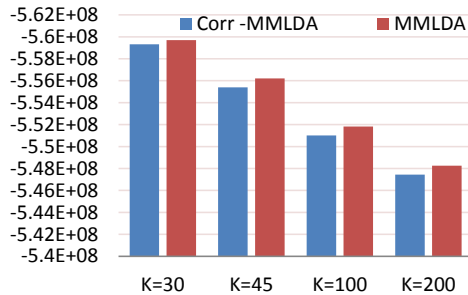


Figure 6.17: Prediction ELBOs from the two topic models for the videos in TRECVID dataset. Lower is better

have approximately the same fit and word prediction power. We choose the MMLDA model since it is computationally less expensive.

$K=200$	$n=1$	$n=5$	$n=10$	$n=15$
MMLDA	0.03518	0.11204	0.18700	0.24117
Corr-MMLDA	0.03641	0.11063	0.18406	0.24840

Table 6.5: Average word prediction 1-gram recall for different topic models with 200 topics when the full corpus is used. The numbers are slightly lower for lower number of topics but not statistically insignificant.

6.5.1.2 MIDDLE LEVEL: VISUAL CONCEPT EXTRACTION TO LANGUAGE

The middle level is a top-down approach that detects concepts sparsely throughout the video, matches them over time, which we call stitching, and relates them to a tripartite template graph for generating language output.

Concept Detectors:

Instead of using publicly available object detectors from datasets like the PASCAL VOC [Everingham et al., 2010], or training independent object detectors for objects such as *microphone*, we build the

concept object detectors like *microphone with upper body*, *group of people* etc., where objects together form a single concept. A concept detector captures more richness semantic information (from object, action and scene level) than the object detectors alone, and usually reduces the visual complexity compared to individual objects, which requires less training examples for an accurate detector. Concept detectors can be looked upon as a generalization of *visual phrases* in [Sadeghi and Farhadi, 2011], to handle more general cases in a video.

We use state-of-the-art Deformable Parts based Model (DPM) [Felzenszwalb et al., 2010] for training concept detectors, some examples of which are visualized in Fig. 6.18. The specific concepts we choose are based on the most frequently occurring objects in the human synopses from the training videos. We use the VATIC tool [Vondrick et al., 2010b] to annotate the trajectories of concept detectors in training videos, which are used in Section 6.5.1.2 for extracting concept relations.

Sparse Object Stitching (SOS):

Concept detectors act as a proxy to the trajectories being tracked in a video. However, tracking over detection is a challenging and open problem for videos *in-the-wild*. First, camera motion and the frame rate are unpredictable, rendering the existed tracking methods useless. Second, the scale of our dataset is huge (thousands of video hours), we hence need a fast alternative. Our approach is called *sparse object stitching*; we sparsely obtain the concept detections in a video and then sequentially group frames based on commonly detected concepts.

For a given video \mathcal{M} , we run the set of concept detectors \mathcal{L} on N sparse distributed frames (e.g. 1 frame/sec) and denote the set of positive detections on each frame as D_i . The algorithm tries to segment the video \mathcal{M} into a set of concept shots $\mathcal{S} = \{S_1, S_2, \dots, S_M\}$, where $\mathcal{S} = \cup D_i$, and $M \ll N$, so that each S_j can be independently described by some sparse detections similar in spirit to [Khan et al., 2011]. We start by uniformly splitting \mathcal{S} into K proposal shots $\{S'_1, S'_2, \dots, S'_K\}$, each captures certain detections in a video. Then we traverse the proposed shots one by one considering two neighboring shots S'_k and S'_{k+1} at a time. If the Jaccard distance $J(S'_k, S'_{k+1}) = 1 - \frac{|S'_k \cap S'_{k+1}|}{|S'_k \cup S'_{k+1}|}$ is lower than a threshold σ (set as 0.5 using cross-validation), then we merge these two proposed shots into one shot and compare it with the next shot, otherwise shot S'_k is an independent shot. For each such concept shot, we match it to a tripartite template graph, as we describe next.

Tripartite Template Graph:

We use a tripartite graph $\mathcal{G} = (V^s, V^t, V^o, E)$ — V^s for human subjects, V^t for tools, and V^o for objects—that takes the concept detections from S_j to generate template based language description. The vertex set $\mathcal{V} = V^s \cup V^t \cup V^o$ is identical to the set of concept detectors \mathcal{L} in the domain at hand. Each concept detector is assigned to one of the three vertex sets. The set of paths $\mathcal{P} = \{(E_{i,j}, E_{j,k}) | i \in V^s, j \in V^t, k \in V^o\}$ is defined as all valid paths from V^s to V^o through V^t , and each forms a possible language output.

We use the annotated object trajectories in training videos to build the edge set E . Our observation is that if there is an edge between two nodes from different vertex sets, then the trajectories of the corresponding concept detector annotations have certain overlaps. Based on this observation, we build the graph by counting the temporal coherence among the concept detector annotations.

Language output: Given the top confident concept detections $\mathcal{L}_c \subset \mathcal{L}$ in one concept shot S_j , we activate the set of paths $\mathcal{P}_c \subset \mathcal{P}$. A natural language sentence is output for paths containing a common

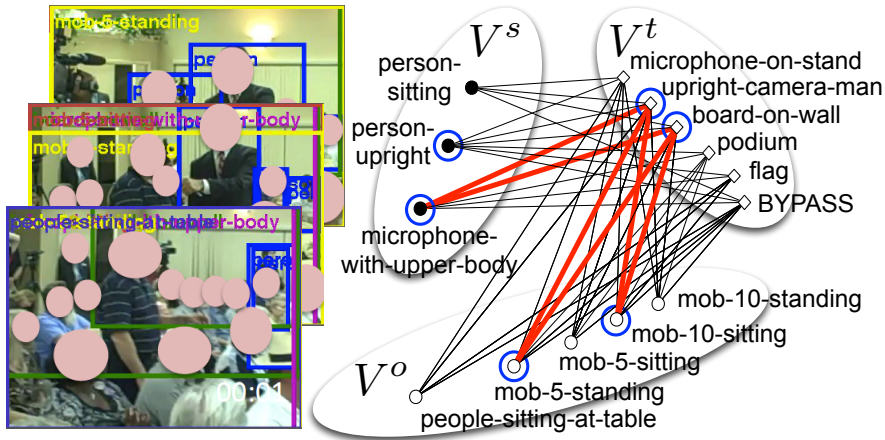


Figure 6.19: Linguistic descriptions from tripartite template graphs consisting of concepts as vertices

subject using the template $\langle V^s, V^t, V^o \rangle$. For situations where $\mathcal{L}_c \cap V^t = \emptyset$, the consistency of the tripartite graph is maintained through a default “BYPASS” node in V^t (see Figs. 6.16 and 6.19). This node acts as a “backspace” production rule in the final lingual output thereby connecting the subject to an object effectively through a single edge. There is, similarly, a BYPASS node in V^o as well. We generally do not consider the situation that $\mathcal{L}_c \cap V^s = \emptyset$, in which no human subject is present. Histogram counts are used for ranking the concept nodes for the lingual output.

Fig. 6.19 depicts a visual example of this process. The edges represent the action phrases or function words that stitch the concepts together cohesively. For example, consider the following structure: “([a person with microphone]) *is speaking to* ([a large group of standing people] and [a small group of sitting people]) *with* ([a camera man] and [board in the back]).” Here the parentheses encapsulate a simple conjunctive production rule and the phrases inside the square brackets denote human subjects, tools or objects. The edge labels in this case are “is speaking to” and “with” which are part of the template $\langle V^s, V^t, V^o \rangle$. In the figure, \mathcal{L}_c is colored blue and edges in \mathcal{P}_c with the common vertex “microphone-with-upper-body” are colored red. We delete repeated sentences as well in the final summary.

6.5.1.3 HIGH LEVEL: FINAL LINGUAL DESCRIPTIONS

The final level of our system joins the two earlier lingual descriptions to enhance the set of sentences given from the middle level and at the same time to filter the sentences from the low level. Our method takes the predicted words from the low level and tags their part of speech with standard NLP tools; these are then used to retrieve weighted nearest neighbors from the training synopses, which are then ranked according to predictive importance, similar in spirit to how Farhadi et al. [Farhadi et al., 2010] select sentences, but we rank over semantically verified low level sentences, giving higher weight to shorter sentences while always ranking the template generated sentences from the middle level, as the best ones.

We use the dependency grammar and part-of-speech (POS) models in the Stanford NLP Suite⁵ to create annotated dictionaries based on word morphologies; the human synopses provide the input. The predicted keywords from the low level topic models are labeled through these dictionaries. For more than two POSs for the same morphology, we prefer verbs, but other variants can be retained as well without loss of generality. For the video in Fig. 6.19, we obtain the following labeled top 15 key-

⁵nlp.stanford.edu/software/corenlp.shtml

words: “hall/OBJ town/NOUN meeting/VERB man/SUBJ-HUMAN speaks/VERB microphone/OBJ talking/VERB representative/SUBJ-HUMAN health/NOUN care/NOUN politician/SUBJ-HUMAN chairs/NOUN flags/OBJ people/OBJ crowd/OBJ.” The word annotation classes used are Subjects, Verbs, Objects, Nouns and “Other.” Subjects which can be humans (SUBJ-HUMAN) are determined using WordNet synsets. Note that this form of keyword generation “summarizes” the video as a bag-of-words which falls short of making the video humanly understandable w.r.t concept importance [Berg et al., 2012].

To obtain the final lingual description of a test video, the output from tripartite template graphs are used first. If there happen to be no detections, we just output the sentences selected using keywords. For semantic verification, we train MMLDA on a vocabulary of training synopses and training concept annotations available using VATIC. Then we compute the number of topic rank inversions for two ranked lists of the top P predictions and top C detections from a test video as:

$$\begin{aligned}
 L_{keywords} &= \left\langle \left\{ k : \sum_{j=1}^V \sum_{m=1}^P p(w_m | \beta_k) \delta(w_m^j) \right\}^\uparrow \right\rangle \\
 L_{concepts} &= \left\langle \left\{ k : \sum_{j=1}^{corrV} \sum_{n=1}^C p(w_n | \rho_k) \delta(w_n^j) \right\}^\uparrow \right\rangle
 \end{aligned} \tag{6.44}$$

If the number of inversions is less than a threshold ($\leq \sqrt{P + C}$) then the keywords are semantically verified by the detected concept list. Finally, we retrieve nearest neighbor sentences from the training synopses by a ranking function. Each sentence s is ranked as: $r_s = bh(w_1 x_{s_1} + w_2 x_{s_2})$ where b is a boolean variable indicating that a sentence must have at least two of the labeled predictions which are verified by the class of words to which the concept models belong. The boolean variable h indicates the presence of at least one human subject in the sentence. The variable indicating the total number of matches divided by the number of words in the sentence is x_{s_1} —this penalizes longer and irrelevant sentences. The sum of the weights of the predicted words from the topic model in the sentence is x_{s_2} —the latent topical strength is reflected here. Each of x_{s_1} and x_{s_2} is normalized over all matching sentences. The weights for sentence length penalty and topic strength respectively are w_1 and w_2 (we set these to be equal in our implementation). The transfer of training sentences *in toto* to describe a test video may be suboptimal in differentiating one video from another in the same event, however, as a summary (sentence ordering aside), it covers more relevant information than just pulling in more keywords upto a certain summary length (see Table 6.7).

6.5.2 Further experiments and Results

6.5.2.1 DATASETS AND FEATURES

TRECVID MED12 dataset: The first dataset that we use for generating lingual descriptions of real life videos is released as part of NIST’s 2012 TRECVID Multimedia Event Detection (MED12). The training set is organized into 25 event categories each containing about 200 videos of positive and related instances of the event descriptions. For choosing one topic model over another (Section ??) we use the positive videos and synopses in the 25 training events and predict the words for the positive videos for

the first five events in the Dev-T collection⁶. The synopses in the training set consist of short and very high level descriptions of the corresponding videos and ranges from 2 to 42 words but on average **10** words with stopwords.

A separate dataset released as part of the Multimedia Event Recounting (MER) task contains six test videos per event where the five events are selected from among the 25 events for MED12. These five events are: 1) *Cleaning an appliance*; 2) *Renovating a home*; 3) *Rock Climbing*; 4) *Town hall meeting*; 5) *Working on a metal crafts project*. Since this MER12 test set cannot be released to public for obtaining synopses, we use in-house annotators to write one short sentence for each of the videos.

In-house “YouCook” dataset on cooking videos: The cooking dataset consists of 88 videos downloaded from YouTube. The training set consists of 49 videos with concept annotations. We use these annotations to train concept DPM models. The test set consists of 39 videos which consists of extensive camera motion and zooming in and out so as to make it a hard test set for annotation. The concept object models for this set roughly fall in the categories of utensils (31%), bowls (38%), dairy (4%), vegetables and fruits (4%), meats (8%), condiments (4%) and miscellaneous ingredients (11%).

We use Amazon’s Mechanical Turk to obtain synopses corresponding to the collected videos. The users are shown an example video with a sample description focusing on the actions and objects therein. Participants in MTurk are instructed to watch a cooking video as many times as required to linguistically describe the video in *at least* three sentences totaling a *minimum* of 15 words. We set our minimum due to the complex nature of the micro-actions in this dataset. After analyzing the synopses, the average number of words per summary is 67, the average number of words per sentence in the summary is 10 with stopwords and the average number of summaries per video is eight. There is a recent data set also about cooking [Rohrbach et al., 2012] but it is a fixed scene dataset and no object annotations are included.

Low Level Features for Topic Model: We use different types of low level video features for different events in the MED12 dataset. HOG3D [Kläser et al., 2008] is the common feature that we use for all events. We resize the video frames such that the largest dimension (height or width) was 160 pixels, and extract HOG3D features from a dense sampling of frames. We then use K-means clustering to create a 4000-word codebook following from a random sampling of the MED12 training data. Instead of color histograms, we use transformed color histogram (TCH) features [van de Sande et al., 2010] for the videos in MED12 dataset due to poor resolution in most of them. We use a 4096 dimensional codebook for TCH. The different visual features that we use for five events in the MED12 and MER12 test datasets are as follows: for events 1, 2, 3 and 4 we use HOG3D and TCH; for event 5 we only use HOG3D only. These decisions are made through 5-fold cross validation performance using the quantitative evaluation mentioned in Section 6.5.2.2 on the MED12 training data for the five events in MER12 test set. Held-out data log likelihoods only help in setting a minimum number of topics which is found to be 10. A higher number of topics do not improve end task performance significantly.

For our YouCook dataset, the low level features we use are HOG3D and color histograms. We use a 1000-word codebook for the HOG3D features due to sparsity of the training set and also following [Bilinski and Bremond, 2011]. Color histograms are computed using 512 RGB color bins. Further, they are computed over each frame and merged across the video. Due to large deviations in the extreme values, we use the histogram between the 15th and 85th percentiles averaged over the entire video.

⁶<http://www.nist.gov/itl/iad/mig/med12.cfm>

Events	P-T-KW	P-D-S	P-T-S	P-DT-S	R-T-KW	R-D-S	R-TM-S	R-DT-S
Cleaning appliance	20.03	11.69 ^(*)	17.52	10.68 ^(*)	19.16 ⁽⁻⁾	35.76	32.60	48.15
Renovating home	6.66	12.55	15.29	9.99	7.31 ⁽⁻⁾	30.67	43.41	49.52
Rock climbing	24.45	24.52	16.21 ^(*)	12.61 ^(*)	44.09	46.23	59.22	65.84
Town hall meeting	17.35	27.56	14.41	13.36	13.80 ⁽⁻⁾	45.55	28.66	56.44
Metal crafts project	16.73	31.68	18.12	15.63	19.01 ⁽⁻⁾	25.87	41.87	54.84

Table 6.6: ROUGE-1 precision and recall scores for MER12 test set. A ⁽⁻⁾ for the R-T-KW column means significantly lower performance than the next 2 columns. The **bold** numbers in the last column is significantly better than the previous 3 columns in terms of recall. The **bold** numbers in P-D-S column are significantly better than those in P-T-KW column. A ^(*) in columns 3, 4 or 5 means significantly lower than P-T-KW. A 95% confidence interval is used for significance testing.

Since the events for the test videos are known, separate topic models are built for each event that vary based on the text vocabulary—the visual feature codebooks are not event specific. This also improves performance and reduces compute time.

6.5.2.2 QUANTITATIVE EVALUATION

We use the ROUGE [Lin and Hovy, 2003] tool to evaluate whether the flow of evidence bottom up results in more relevant content being generated as a *video summary*. As used in [Yang et al., 2011], ROUGE is a standard for comparing text summarization systems which focuses on recall of relevant information coverage. The BLEU [Papineni et al., 2002] scorer is more precision oriented and is useful for comparing accuracy and fluency (usually using 4-grams) of the outputs of text translation systems as used in [Belz and Reiter, 2006, Kulkarni et al., 2011] which is not our end task.

The problem presented here is extremely difficult to solve—in the UIUC PASCAL sentence dataset [Rashtchian et al., 2010], five sentences are used *per image*. On the other hand we only allow at most 5 sentences *per video* per level – low or middle up to a maximum of ten. A human, on the other hand, can describe a video, on average, in just one sentence!

Table 6.6 shows the ROUGE-1 recall and precision scores obtained from the different outputs from our system for the MER12 test set. In tables 6.6 and 6.7, T means the low level topic model and D is the DPM model in [Felzenszwalb et al., 2010]. R means “Recall” and P means “Precision.” Note that we use top 15 keywords with redundancy particularly retaining subjects like “man,” “woman” etc. and verb morphologies (which otherwise stem to the same prefix) as proxies for ten-word training synopses. KW means keywords and S means sentences. The baseline system is a single set of keywords—the output from our lowest level.

From Table 6.6, it is clear that lingual descriptions from both the lower and middle levels of our system cover more relevant information, albeit, at the cost of introducing additional words. Increasing the number of keywords improves recall but precision drops dramatically. The drop in precision for our final output is also due to increased length of the descriptions. However, the scores remain within the 95% confidence interval of that from the keywords for “Renovating home,” “Town hall meeting” and “Metal crafts project” events. The “Rock climbing” event has very short synopses as reference summaries and the “Cleaning an appliance” event is a very hard event both for DPM as well as MMLDA since multiple related concepts indicative of appliances in context appear in prediction and detection. From Table 6.6 we see the efficacy of the short lingual descriptions from sparse object stitching in terms of precision while the final output of our system significantly outperforms relevant content coverage of

the lingual descriptions from the other individual levels with regards to recall.

P2-T - KW	P1-T - KW	R2-T - KW	R1-T - KW	P2- DT-S	P1- DT-S	R2- DT-S	R1- DT-S
6E-4	15.47	6E-4	19.02	5.04	24.82	6.81	34.2

Table 6.7: ROUGE scores for our “YouCook” dataset





Table 6.7 shows ROUGE scores for both 1-gram and 2-gram comparisons. R1 means ROUGE-1-Recall and P1 means ROUGE-1-Precision. Similarly for R2 and P2. The length of the all system summaries are truncated at 67 words based on the average human synopses lengths. The sentences from the low level are chosen based on the top 15 predictions only. For fair comparison on recall, the number of keywords (KW columns in Table 6.7) is chosen to be 67. The numbers in bold are significant at 95% confidence over corresponding columns in the left. R-2 is non-zero for keywords since some paired keywords are indeed phrases. Our method thus performs significantly well even when compared against longer synopses. Our lingual descriptions built on top of concept labels and just a few keywords significantly outperform labeling with even *four times* as large a set of keywords! This can also tune language models to context since creating a sentence out of the predicted nouns and verbs do not increase recall based on unigrams.

6.5.2.3 QUALITATIVE EXAMPLES

The first four rows in Fig. 6.20 show examples from the MER12 test set. The first one or two italicized sentences in each row are the result of tripartite template graph output. The “health care reform” in the second row is a noise phrase that actually cannot be verified though our middle level but remains in the sentence due to our conservative ranking formula. Next we show a good and a bad example from our YouCook dataset. The two human synopses in last 2 rows are shown for the purpose of illustrating their variance and yet their relevancy. The last cooking video has a low R1-R score of 21% due to imprecise predictions and detections.

6.6 Summary

Documents containing video and text are becoming more and more widespread and yet content analysis of those documents depends primarily on the text. Although automated discovery of semantically related words from text improves free text query understanding, translating videos into text summaries can also yield better information needs that capture the semantic content of the videos as well. This facilitates better video search particularly in the absence of accompanying text. In this paper, we propose a multimedia topic modeling framework suitable for providing a basis for automatically discovering and translating semantically related words obtained from textual metadata of multimedia documents to semantically related videos or frames from videos. The framework jointly models video and text and is flexible enough to handle different types of document features in their constituent domains such as discrete and real valued features from videos representing actions, objects, colors and scenes as well as discrete features from text. Our proposed models show much better fit to the multimedia data in terms of held-out data log likelihoods. For a given query video, our models translate low level vision features into bag of keyword summaries, which can be further translated using simple natural language generation techniques into human readable paragraphs ideal for information extraction needs. We quantitatively

	<p>Keywords: refrigerator/OBJ cleans/VERB man/SUBJ-HUMAN clean/VERB blender/OBJ cleaning/VERB woman/SUBJ-HUMAN person/SUBJ-HUMAN stove/OBJ microwave/OBJ sponge/NOUN food/OBJ home/OBJ hose/OBJ oven/OBJ</p> <p>Sentences from Our System 1. <i>A person is using dish towel and hand held brush or vacuum to clean panel with knobs and washing basin or sink</i> 2: Man cleaning a refrigerator. 3: Man cleans his blender. 4: Woman cleans old food out of refrigerator. 5: Man cleans top of microwave with sponge.</p> <p>Human Synopsis: Two standing persons clean a stove top with a vacuum clean with a hose.</p>
	<p>Keywords: meeting/VERB town/NOUN hall/OBJ microphone/OBJ talking/VERB people/OBJ podium/OBJ speech/OBJ woman/SUBJ-HUMAN man/SUBJ-HUMAN chairs/NOUN clapping/VERB speaks/VERB questions/VERB giving/VERB</p> <p>Sentences from Our System 1: <i>A person is speaking to a small group of sitting people and a small group of standing people with board in the back.</i> 2: A person is speaking to a small group of standing people with board in the back 3: Man opens town hall meeting. 4: Woman speaks at town meeting. 5: Man gives speech on health care reform at a town hall meeting.</p> <p>Human Synopsis: A man talks to a mob of sitting persons who clap at the end of his short speech.</p>
	<p>Keywords: people/SUBJ-HUMAN, home/OBJ, group/OBJ, renovating/VERB, working/VERB, montage/OBJ, stop/VERB, motion/OBJ, appears/VERB, building/VERB, floor/OBJ, tiles/OBJ, floorboards/OTHER, man/SUBJ-HUMAN, laying/VERB</p> <p>Sentences from Our System: 1. <i>A person is using power drill to renovate a house.</i> 2. <i>A crouching person is using power drill to renovate a house.</i> 3. A person is using trowel to renovate a house. 4: man lays out underlay for installing flooring. 5: A man lays a plywood floor in time lapsed video.</p> <p>Human Synopsis: Time lapse video of people making a concrete porch with sanders, brooms, vacuums and other tools.</p>
	<p>Keywords: metal/OBJ man/SUBJ-HUMAN bending/VERB hammer/VERB piece/OBJ tools/OBJ rods/OBJ hammering/VERB craft/VERB iron/OBJ workshop/OBJ holding/VERB works/VERB steel/OBJ bicycle/OBJ</p> <p>Sentences from Our System 1. <i>A person is working with pliers.</i> 2 Man hammering metal. 3. Man bending metal in workshop. 4. Man works various pieces of metal. 5. A man works on a metal craft at a workshop.</p> <p>Human Synopsis: A man is shaping a star with a hammer.</p>

In the images below, no detections are shown for clarity. *The sentences in italics* are output through tripartite template graphs (Section 6.5.1.2)



	<p>Keywords: bowl/OBJ pan/OBJ video/OBJ adds/VERB lady/OBJ pieces/OBJ ingredients/OBJ oil/OBJ glass/OBJ liquid/OBJ butter/SUBJ-HUMAN woman/SUBJ-HUMAN add/VERB stove/OBJ salt/OBJ</p> <p>Sentences from Our System: 1. <i>A person is cooking butter with bowl and stovetop.</i> 2. In a pan add little butter. 3. She adds some oil and a piece of butter in the pan. 4. A woman holds up Bisquick flour and then adds several ingredients to a bowl. 5. A woman adds ingredients to a blender.</p> <p>Human Synopsis1: A lady wearing red colored dress, blending (think butter) in a big sized bowl. Besides there is 2 small bowls containing white color powders. It may be maida flour and sugar. After she is mixing the both powders in that big bowl and blending together. Human Synopsis2: In this video, a woman first adds the ingredients from a plate to a large porcelain bowl. She then adds various other ingredients from various different bowls. She then mixes all the ingredients with a wooden spoon.</p>
	<p>Keywords: bowl/OBJ pan/OBJ video/OBJ adds/VERB ingredients/OBJ lady/OBJ woman/SUBJ-HUMAN add/VERB pieces/OBJ stove/OBJ oil/OBJ put/VERB added/VERB mixes/VERB glass/OBJ</p> <p>Sentences from Our System: 1. <i>A person is cooking with pan and bowl.</i> 2. <i>A person is cooking with pan.</i> 2. A woman adds ingredients to a blender. 2. In this video, a woman adds a few ingredients in a glass bowl and mixes them well. 3. In this video, a woman first adds the ingredients from a plate to a large porcelain bowl 4. The woman is mixing some ingredients in a bowl. 5. the woman in the video has a large glass bowl.</p> <p>Human Synopsis1: The woman is giving directions on how to cook bacon omelette. She shows the ingredients for cooking and was frying the bacon, scrambling the egg, melting the butter and garnishing it with onions and placed some cheese on top. The woman then placed the scrambled egg and bacon to cook and then placed it on a dish. Human Synopsis2: in this video the woman takes bacon, eggs, cheese, onion in different containers. On a pan she cooks the bacon on low flame. Side by side she beats the eggs in a bowl, she removes the cooked bacon on a plate. In the pan she fries onions and then adds the beaten eggs. She sprinkles grated cheese on the pan and cooks well. She then adds the fried bacon on the eggs in the pan and cook well. She transfers the cooked egg with bacon to as serving plate.</p>

Figure 6.20: Qualitative results from MER12 test and our “YouCook” dataset. Only top 5 sentences from our system are shown.

compare the results of video to text translation in the bag of words form against a state-of-the-art baseline object recognition model from computer vision. We show that text translations from multimodal topic models vastly outperform the baseline on a multimedia dataset downloaded from the Internet.

In general Corr-MMGLDA improves on text to video translation while the non-correspondence versions perform better in video to text summarization. Video summarization through topic models significantly out-perform that through state-of-the-art object detectors and thus can be used as new baselines. Our NLG component has suffered from severe data sparsity and impoverished language models and we wish to overcome these using external knowledge bases.

We further combine the best aspects of top-down and bottom-up methods of producing lingual descriptions of videos *in-the-wild* that exploit the rich semantic space of both text and visual features. Our contribution is unique in that the class of concept detectors semantically verify low level predictions from bottom up and leverage both sentence generation and selection that together outperforms the coverage of information need output from independent modules.

6.7 Acknowledgements

We thank the Mori group at Simon Fraser University, Kitware Inc. and Scott McCloskey at Honeywell ACS Labs for helpful discussions and feature extraction, and Philip Rosebrough, Cody Boppert, Yao Li, and David Molik for their work on data curation and annotation. This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF-11-1-0090), the DARPA Minds Eye program (W911NF-10-2-0062), and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, DARPA, ARO, NSF or the U.S. Government.

6.8 Appendix

6.8.1 Some Important Derivations

In order to derive the optimal solution for $q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$ we start with its factorization and selecting only those terms which depend on $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$.

$$\ln q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{i=1}^K \ln p(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) + E_{q(\mathbf{Z})}[\ln p(\mathbf{Z}|\boldsymbol{\theta})] + \sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{Z})}[z_{d,o,i}] \ln \mathcal{N}(\mathbf{w}_{d,o}|\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) + const \quad (6.45)$$

where $\ln \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) = 1/2 [-\ln(2\pi) + \ln |\boldsymbol{\Lambda}_i| - (\mathbf{w}^T \boldsymbol{\Lambda}_i \mathbf{w} - \boldsymbol{\mu}_i^T \boldsymbol{\Lambda}_i \mathbf{w} - \mathbf{w}^T \boldsymbol{\Lambda}_i \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T \boldsymbol{\Lambda}_i \boldsymbol{\mu}_i)]$. Denoting P to be the dimensionality of \mathbf{w} and expanding, we have:

$$\begin{aligned} \ln q^*(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) &= (-\kappa_0/2) \left((\boldsymbol{\mu}_i - \mathbf{m}_0)^T \boldsymbol{\Lambda}_i (\boldsymbol{\mu}_i - \mathbf{m}_0) \right) + (1/2) \ln |\boldsymbol{\Lambda}_i| - (1/2) \text{Tr}(\boldsymbol{\Lambda}_i \mathbf{W}_0^{-1}) \\ &+ \frac{\nu_0 - P - 1}{2} \ln |\boldsymbol{\Lambda}_i| - (1/2) \sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{Z})}[z_{d,o,i}] \left((\mathbf{w}_{d,o} - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i (\mathbf{w}_{d,o} - \boldsymbol{\mu}_i) \right) \\ &+ (1/2) \left(\sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{Z})}[z_{d,o,i}] \right) \ln |\boldsymbol{\Lambda}_i| + const \end{aligned} \quad (6.46)$$

Using the product rule of probability, we can express $\ln q^*(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$ as $\ln q^*(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i) + \ln q^*(\boldsymbol{\Lambda}_i)$. To identify the distribution for $\boldsymbol{\mu}_i$, we select the terms on the right hand side of Equ. 6.46 which depend on $\boldsymbol{\mu}_i$, yielding:

$$\begin{aligned} \ln q^*(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i) &= -(1/2) \boldsymbol{\mu}_i^T \left[\kappa_0 + \sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{Z})}[z_{d,o,i}] \right] \boldsymbol{\Lambda}_i \boldsymbol{\mu}_i \\ &+ \boldsymbol{\mu}_i^T \boldsymbol{\Lambda}_i \left[\kappa_0 + \sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{Z})}[z_{d,o,i}] \mathbf{w}_o \right] + const \\ &= -(1/2) \boldsymbol{\mu}_i^T [\kappa_0 + N_i] \boldsymbol{\Lambda}_i \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T \boldsymbol{\Lambda}_i [\kappa_0 \mathbf{m}_0 + N_i \bar{\mathbf{x}}_i] + const \end{aligned} \quad (6.47)$$

where we have made use of the expressions in 6.27. Thus we see that $\ln q^*(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i)$ is a Gaussian

distribution i.e. a quadratic in $\boldsymbol{\mu}_i$. Completing the squares of the quadratic expression in Equ. 6.47 and that for the Gaussian distribution and identifying the coefficients allows us to determine the mean and precision of this Gaussian, yielding:

$$q^*(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i) = \mathcal{N}(\boldsymbol{\mu}_i|\mathbf{m}_i, (\kappa_i\boldsymbol{\Lambda}_i)) \quad (6.48)$$

where

$$\kappa_i = \kappa_0 + N_i; \quad \mathbf{m}_i = \frac{1}{\kappa_i}(\kappa_0\mathbf{m}_0 + N_i\bar{\mathbf{x}}_i) \quad (6.49)$$

The expressions in Equ. 6.49 are intuitive and follows from the conjugate prior properties of the parameters. The posterior for κ_i is effectively reflecting the amount of the expected *number* of observations in component i added to the initial number of pseudo observations. The posterior for \mathbf{m}_i captures the updated *value* of \mathbf{m} through the first order sufficient statistics of component i .

Next we determine the distributional form of $\ln q^*(\boldsymbol{\Lambda}_i)$ by using the fact that

$$\ln q^*(\boldsymbol{\Lambda}_i) = \ln q^*(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) - \ln q^*(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i) \quad (6.50)$$

Substituting Eqs. 6.48 and 6.49 in the right hand side of Equ. 6.50 and keeping only those terms that depend on $\boldsymbol{\Lambda}_i$ we obtain the following:

$$\begin{aligned} \ln q^*(\boldsymbol{\Lambda}_i) &= (-\kappa_0/2) \left((\boldsymbol{\mu}_i - \mathbf{m}_0)' \boldsymbol{\Lambda}_i (\boldsymbol{\mu}_i - \mathbf{m}_0) \right) + (1/2) \ln |\boldsymbol{\Lambda}_i| - (1/2) \text{Tr}(\boldsymbol{\Lambda}_i \mathbf{W}_0^{-1}) \\ &+ \frac{\nu_0 - P - 1}{2} \ln |\boldsymbol{\Lambda}_i| - (1/2) \sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{z})}[z_{d,o,i}] \left((\mathbf{w}_{d,o} - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i (\mathbf{w}_{d,o} - \boldsymbol{\mu}_i) \right) \\ &+ \frac{1}{2} \left(\left(\sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{z})}[z_{d,o,i}] \right) \ln |\boldsymbol{\Lambda}_i| + \kappa_0 \left((\boldsymbol{\mu}_i - \mathbf{m}_0)^T \boldsymbol{\Lambda}_i (\boldsymbol{\mu}_i - \mathbf{m}_0) \right) - \ln |\boldsymbol{\Lambda}_i| \right) + \text{const} \\ &= \frac{\nu_i - P - 1}{2} \ln |\boldsymbol{\Lambda}_i| - (1/2) \text{Tr}(\boldsymbol{\Lambda}_i \mathbf{W}_i^{-1}) + \text{const} \end{aligned} \quad (6.51)$$

The expression for $q^*(\boldsymbol{\Lambda}_i)$ in Equ. 6.51 has the functional form of the probability density function of the Wishart distribution defined in general as $\ln \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) = \ln Z(\mathbf{W}, \nu) + (\nu - P - 1)/2 \ln |\boldsymbol{\Lambda}| - (1/2) \text{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})$ where $Z(\mathbf{W}, \nu)$ is the normalization constant given by $Z(\mathbf{W}, \nu) = |\mathbf{W}|^{-\nu/2} \left(2^\nu P/2 \pi^{P(P-1)/4} \prod_{p=1}^P \Gamma\left(\frac{\nu+1-p}{2}\right) \right)^{-1}$. By expanding and matching coefficients we obtain:

$$\begin{aligned} \mathbf{W}_i^{-1} &= \mathbf{W}_0^{-1} + \kappa_0 (\boldsymbol{\mu}_i - \mathbf{m}_0)(\boldsymbol{\mu}_i - \mathbf{m}_0)^T + \sum_{d=1}^D \sum_{o=1}^O E_{q(\mathbf{z})}[z_{d,o,i}] (\mathbf{w}_{d,o} - \boldsymbol{\mu}_i)(\mathbf{w}_{d,o} - \boldsymbol{\mu}_i)^T \\ &- \kappa_0 (\boldsymbol{\mu}_i - \mathbf{m}_k)(\boldsymbol{\mu}_i - \mathbf{m}_k)^T \\ &= \mathbf{W}_0^{-1} + N_i \mathbf{S}_i + \frac{\kappa_0 N_i}{\kappa_0 + N_i} (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_0)^T \end{aligned} \quad (6.52)$$

and

$$\nu_i = \nu_0 + \sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{z})}[z_{d,o,i}] = \nu_0 + N_i \quad (6.53)$$

where we have used the result:

$$\sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{Z})}[z_{d,o,i}] \mathbf{w}_{d,o} \mathbf{w}_{d,o}^T = \sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{Z})}[z_{d,o,i}] (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i) (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i)^T + N_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \quad (6.54)$$

The terms involving $\boldsymbol{\mu}_i$ have canceled out in Equ. 6.52 and this is intuitive since $q^*(\boldsymbol{\Lambda}_i)$ is independent of $\boldsymbol{\mu}_i$. Thus $q^*(\boldsymbol{\Lambda}_i)$ is a Wishart distribution of the form $q^*(\boldsymbol{\Lambda}_i) = \mathcal{W}(\boldsymbol{\Lambda}_i | \mathbf{W}_i, \nu_i)$. The interpretation for the posteriors over ν_i and \mathbf{W}_i are exactly the same as that for κ_i and \mathbf{m}_i , the only difference being the use of second order sufficient statistics.

It is interesting to note that if we do not use any priors over the Gaussian parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i^{-1} = \boldsymbol{\Sigma}_i$, then we fall back to the MLE for $\boldsymbol{\Sigma}_i$ as in [Blei and Jordan, 2003] as follows:

$$\mathcal{L}_{[\boldsymbol{\Sigma}_i]} = E_{q(\mathbf{Z}^{(O)})} \log p(\mathbf{W}, \mathbf{W}^{(O)}, \mathbf{W}^{(H)}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{Z}^{(O)}, \mathbf{Z}^{(H)}) = \frac{1}{2} \sum_{d=1}^D \sum_{o=1}^O \sum_{i=1}^K \phi_{d,o,i}^{(O)} [\log |\boldsymbol{\Sigma}_i|^{-1} - \text{Tr}(\boldsymbol{\Sigma}_i^{-1} \mathbf{S}_i)] \quad (6.55)$$

where as in Equ. 6.27,

$$N_i = \sum_{d=1}^D \sum_{o=1}^{O_d} \phi_{d,o,i}^{(O)}; \quad \bar{\mathbf{x}}_i = \frac{1}{N_i} \sum_{d=1}^D \sum_{o=1}^{O_d} \phi_{d,o,i}^{(O)} \mathbf{w}_{d,o}$$

$$\mathbf{S}_i = \frac{\sum_{d=1}^D \sum_{o=1}^{O_d} \phi_{d,o,i}^{(O)} (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i) (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i)'}{N_i}$$

Taking derivatives w.r.t $\boldsymbol{\Sigma}_i$ we have,

$$\therefore 2 \frac{\partial \mathcal{L}_{[\boldsymbol{\Sigma}_i]}}{\partial \boldsymbol{\Sigma}_i^{-1}} = \frac{\partial}{\partial \boldsymbol{\Sigma}_i^{-1}} \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \log |\boldsymbol{\Sigma}_i^{-1}| \right] - \frac{\partial}{\partial \boldsymbol{\Sigma}_i^{-1}} \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \text{Tr}(\boldsymbol{\Sigma}_i^{-1} (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i) (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i)') \right] \quad (6.56)$$

Now if we denote $\boldsymbol{\Sigma}_i^{-1} = \mathbf{X}$, then we have the following [Harville, 2008]:

$$\frac{\partial |\mathbf{X}|}{\partial x_{i,j}} = \text{Tr} \left[\text{Adj}(\mathbf{X}) \frac{\partial \mathbf{X}}{\partial x_{i,j}} \right] = |\mathbf{X}| \text{Tr} \left[\mathbf{X}^{-1} \frac{\partial \mathbf{X}}{\partial x_{i,j}} \right]$$

$$\therefore \frac{\partial \log |\mathbf{X}|}{\partial x_{i,j}} = \text{Tr} \left[\mathbf{X}^{-1} \frac{\partial \mathbf{X}}{\partial x_{i,j}} \right] = \text{Tr} [\mathbf{X}^{-1} \mathbf{e}_i \mathbf{e}_j'] = \text{Tr} [\mathbf{e}_j' \mathbf{X}^{-1} \mathbf{e}_i] = y_{j,i} \quad (6.57)$$

where $y_{j,i}$ is the j, i^{th} element of \mathbf{X}^{-1} i.e. the i, j^{th} element of $[\mathbf{X}^{-1}]'$. Also, $\text{Adj}(\mathbf{X})$ is the adjoint of the square matrix \mathbf{X} and \mathbf{e}_i is the unit vector whose i^{th} component is 1 and 0 everywhere else.

$$\therefore \frac{\partial \log |\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})' = \mathbf{X}^{-1} \quad \text{iff } x_{i,j} = x_{j,i} \quad (6.58)$$

Thus Equ. 6.56 becomes:

$$\therefore \frac{\partial \mathcal{L}_{[\boldsymbol{\Sigma}_i]}}{\partial \boldsymbol{\Sigma}_i^{-1}} = \frac{1}{2} \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \boldsymbol{\Sigma}_i - \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \text{Tr} \left(\frac{\partial}{\partial (\boldsymbol{\Sigma}_i^{-1})_{m,n}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i) (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i)') \right) \right] \right]$$

$$\begin{aligned}
&= \frac{1}{2} \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \boldsymbol{\Sigma}_i - \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \text{Tr} \left(\mathbf{e}_m \mathbf{e}_n' (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i) (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i)' \right) \right] \right] \\
&= \frac{1}{2} \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \boldsymbol{\Sigma}_i - \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \text{Tr} \left(\mathbf{e}_n' (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i) (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i)' \mathbf{e}_m \right) \right] \right] \\
&= \frac{1}{2} \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \boldsymbol{\Sigma}_i - \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \left(\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i \right) \left(\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i \right)' \right]_{n,m} \right] \\
&= \frac{1}{2} \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \boldsymbol{\Sigma}_i - \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \left(\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i \right) \left(\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i \right)' \right] \right] \\
&= \frac{1}{2} \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \boldsymbol{\Sigma}_i - \left[\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} \left(\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i \right) \left(\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i \right)' \right] \right] \text{ by symmetry of covariance matrix} \\
&= \frac{1}{2} [N_i \boldsymbol{\Sigma}_i - [N_i \mathbf{S}_i]] \tag{6.59}
\end{aligned}$$

Setting Equ. 6.59 to 0 for maximum likelihood estimation of $\boldsymbol{\Sigma}_i$, we obtain:

$$\boldsymbol{\Sigma}_i = \mathbf{S}_i = \frac{\sum_{d=1}^D \sum_{o=1}^O \phi_{d,o,i}^{(O)} (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i) (\mathbf{w}_{d,o} - \bar{\mathbf{x}}_i)'}{N_i} \tag{6.60}$$

To obtain $E_{q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)}[(\mathbf{w}_o - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i (\mathbf{w}_o - \boldsymbol{\mu}_i)]$, we first write down the double integral:

$$E_{q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)} [(\mathbf{w}_o - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i (\mathbf{w}_o - \boldsymbol{\mu}_i)] = \int \int \text{Tr} \left\{ \boldsymbol{\Lambda}_i (\mathbf{w}_o - \boldsymbol{\mu}_i) (\mathbf{w}_o - \boldsymbol{\mu}_i)^T \right\} q^*(\boldsymbol{\mu}_i | \boldsymbol{\Lambda}_i) q^*(\boldsymbol{\Lambda}_i) d\boldsymbol{\mu}_i d\boldsymbol{\Lambda}_i \tag{6.61}$$

Next we use the result $q^*(\boldsymbol{\mu}_i | \boldsymbol{\Lambda}_i) = \mathcal{N}(\boldsymbol{\mu}_i | \mathbf{m}_i, (\kappa_i \boldsymbol{\Lambda}_i))$ to perform integration over $\boldsymbol{\mu}_i$. Using the standard expressions for expectations under a Gaussian distribution we have:

$$E_{q(\boldsymbol{\mu}_i)}[\boldsymbol{\mu}_i] = \mathbf{m}_i \tag{6.62}$$

$$\begin{aligned}
E_{q(\boldsymbol{\mu}_i)}[\boldsymbol{\mu}_i \boldsymbol{\mu}_i^T] &= \kappa_i^{-1} \boldsymbol{\Lambda}_i^{-1} - E_{q(\boldsymbol{\mu}_i)} \left[-\boldsymbol{\mu}_i \mathbf{m}_i^T - \mathbf{m}_i \boldsymbol{\mu}_i^T + \mathbf{m}_i \mathbf{m}_i^T \right] \\
&= \kappa_i^{-1} \boldsymbol{\Lambda}_i^{-1} - \left[-2\mathbf{m}_i \mathbf{m}_i^T + \mathbf{m}_i \mathbf{m}_i^T \right] \\
&= \mathbf{m}_i \mathbf{m}_i^T + \kappa_i^{-1} \boldsymbol{\Lambda}_i^{-1} \tag{6.63}
\end{aligned}$$

from which we obtain the expression with respect to $\boldsymbol{\mu}_i$ in the form

$$E_{q(\boldsymbol{\mu}_i)} [(\mathbf{w}_o - \boldsymbol{\mu}_i) (\mathbf{w}_o - \boldsymbol{\mu}_i)^T] = \mathbf{w}_o \mathbf{w}_o^T - \mathbf{w}_o \mathbf{m}_i^T - \mathbf{m}_i \mathbf{w}_o^T + \mathbf{m}_i \mathbf{m}_i^T + \kappa_i^{-1} \boldsymbol{\Lambda}_i^{-1} \tag{6.64}$$

$$= (\mathbf{w}_o - \mathbf{m}_i) (\mathbf{w}_o - \mathbf{m}_i)^T + \kappa_i^{-1} \boldsymbol{\Lambda}_i^{-1} \tag{6.65}$$

Finally taking expectation w.r.t. $\boldsymbol{\Lambda}_i$ we have:

$$\begin{aligned}
&E_{q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)} [(\mathbf{w}_{d,o} - \boldsymbol{\mu}_i)^T \boldsymbol{\Lambda}_i (\mathbf{w}_{d,o} - \boldsymbol{\mu}_i)] \\
&= \int \text{Tr} \left\{ \boldsymbol{\Lambda}_i [(\mathbf{w}_{d,o} - \mathbf{m}_i) (\mathbf{w}_{d,o} - \mathbf{m}_i)^T] + \kappa_i^{-1} \boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^{-1} \right\} q^*(\boldsymbol{\Lambda}_i) d\boldsymbol{\Lambda}_i \\
&= \int \left\{ (\mathbf{w}_{d,o} - \mathbf{m}_i)^T \boldsymbol{\Lambda}_i (\mathbf{w}_{d,o} - \mathbf{m}_i) + P \kappa_i^{-1} \right\} q^*(\boldsymbol{\Lambda}_i) d\boldsymbol{\Lambda}_i \\
&= P \kappa_i^{-1} + \nu_i (\mathbf{w}_{d,o} - \mathbf{m}_i)^T \mathbf{W}_i (\mathbf{w}_{d,o} - \mathbf{m}_i) \tag{6.66}
\end{aligned}$$

Here we have used $q^*(\Lambda_i) = \mathcal{W}(\Lambda_i | \mathbf{W}_i, \nu_i)$ together with the standard result for the expectation under a Wishart distribution to obtain $E_{q(\Lambda_i)}[\Lambda_i] = \nu_i \mathbf{W}_i$. For the Wishart distribution expressions, we have,

$$\mathcal{W}(\Lambda | \mathbf{W}, \nu) = Z(\mathbf{W}, \nu) |\Lambda|^{(\nu-P-1)/2} \exp\left\{-\frac{1}{2} \text{Tr}(\mathbf{W}^{-1} \Lambda)\right\} \quad (6.67)$$

where,

$$\begin{aligned} Z(\mathbf{W}, \nu) &= |\mathbf{W}|^{-\nu/2} \left(2^{\nu P/2} \pi^{P(P-1)/4} \prod_{p=1}^P \Gamma\left(\frac{\nu+1-p}{2}\right) \right)^{-1} \\ E_{q(\Lambda_i)}[\Lambda] &= \nu \mathbf{W} \\ E_{q(\Lambda_i)}[\ln |\Lambda|] &= \sum_{p=1}^P \Psi\left(\frac{\nu+1-p}{2}\right) + P \ln 2 + \ln |\mathbf{W}| \\ H[\Lambda] &= -\ln Z(\mathbf{W}, \nu) - \frac{\nu-P-1}{2} E[\ln |\Lambda|] + \frac{\nu P}{2} \end{aligned} \quad (6.68)$$

Here \mathbf{W} is a $P \times P$ symmetric positive definite matrix and $\Psi(\cdot)$ is the digamma function. The parameter ν is the degrees of freedom of the distribution and is restricted to be $\geq P$.

To derive the expression: $E[\ln p(\mathbf{w}_O | \mathbf{z}_O, \boldsymbol{\mu}, \Lambda)]$ for every video document d , we use Equ. 6.61 and the expression for $E[\ln |\Lambda|] \equiv \ln \hat{\Lambda}_i$

$$\begin{aligned} &E_{q(\mathbf{z}, \boldsymbol{\mu}_i, \Lambda_i)}[\ln p(\mathbf{w}_O | \mathbf{z}_O, \boldsymbol{\mu}, \Lambda)] \\ &= \frac{1}{2} \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{z})}[z_{d,o,i}] \left\{ E_{q(\Lambda_i)}[\ln \Lambda_i] - E_{q(\boldsymbol{\mu}_i, \Lambda_i)}[(\mathbf{w}_{d,o} - \boldsymbol{\mu}_i) \Lambda_i (\mathbf{w}_{d,o} - \boldsymbol{\mu}_i)] - P \ln(2\pi) \right\} \\ &= \frac{1}{2} \sum_{o=1}^O \sum_{i=1}^K E_{q(\mathbf{z})}[z_{d,o,i}] \left\{ \ln \hat{\Lambda}_i - P \kappa_i^{-1} - \nu_i (\mathbf{w}_{d,o} - \mathbf{m}_i)^T \mathbf{W}_i (\mathbf{w}_{d,o} - \mathbf{m}_i) - P \ln(2\pi) \right\} \end{aligned} \quad (6.69)$$

Typically, for a mixture model like mixture of Gaussians, where the document hierarchy is absent, we can derive an equivalent expression as follows, using Equ. 6.54:

$$\begin{aligned} E_{q(\mathbf{z}, \boldsymbol{\mu}_i, \Lambda_i)}[\ln p(\mathbf{w}_O | \mathbf{z}_O, \boldsymbol{\mu}, \Lambda)] &= \frac{1}{2} \sum_{i=1}^K E_{q(\mathbf{z})}[z_{d,o,i}] \left\{ \ln \hat{\Lambda}_i - P \kappa_i^{-1} - \nu_i \text{Tr}(\mathbf{S}_i \mathbf{W}_i) \right. \\ &\quad \left. - \nu_i (\bar{\mathbf{x}}_i - \mathbf{m}_i)^T \mathbf{W}_i (\bar{\mathbf{x}}_i - \mathbf{m}_i) - P \ln(2\pi) \right\} \end{aligned} \quad (6.70)$$

In our scenario, the LDA family of models is a mixed membership model where each document itself is a distribution over the mixture component proportions. Thus in the VB-EM framework, we need to locally optimize over the tractable family of distributions to find the best one which provides a better improvement to the lower bound of the log likelihood w.r.t. the original model. Had this not been the case, plugging in the expression in Equ. 6.70 for a mixture model is the best option both computationally and numerically.

To derive the expressions for, $\sum_{i=1}^K E_{q[\boldsymbol{\mu}_i, \Lambda_i]}[\ln p(\boldsymbol{\mu}_i, \Lambda_i)]$, we have:

$$\sum_{i=1}^K E_{q(\boldsymbol{\mu}_i, \Lambda_i)}[\ln p(\boldsymbol{\mu}_i, \Lambda_i)] = \frac{1}{2} \sum_{i=1}^K \left\{ P \ln\left(\frac{\kappa_0}{2\pi}\right) + \ln \hat{\Lambda}_i - \frac{\kappa_0 P}{\kappa_i} - \kappa_0 \nu_i (\mathbf{m}_i - \mathbf{m}_0)' \mathbf{W}_i (\mathbf{m}_i - \mathbf{m}_0) \right\}$$

$$+ K \ln Z(\mathbf{W}_0, \nu_0) + \frac{\nu_0 - P - 1}{2} \sum_{i=1}^K \ln \hat{\Lambda}_i - \frac{1}{2} \sum_{i=1}^K \nu_i \text{Tr}(\mathbf{W}_0^{-1} \mathbf{W}_i) \quad (6.71)$$

We start with the definition of the prior $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ to obtain an expression for $E_{q_{[\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i]}}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})]$. We have:

$$\begin{aligned} E_{q_{(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)}}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] &= \frac{1}{2} \sum_{i=1}^K \left\{ P \ln \kappa_0 - P \ln(2\pi) + E[\ln |\boldsymbol{\Lambda}_i|] - \kappa_0 E[(\boldsymbol{\mu}_i - \mathbf{m}_0)^T \boldsymbol{\Lambda}_i (\boldsymbol{\mu}_i - \mathbf{m}_0)] \right\} \\ &+ K \ln Z(\mathbf{W}_0, \nu_0) + \sum_{i=1}^K \left\{ \frac{\nu_0 - P - 1}{2} E[\ln |\boldsymbol{\Lambda}_i|] - \frac{1}{2} \text{Tr}(\mathbf{W}_0^{-1} E[\boldsymbol{\Lambda}_i]) \right\} \end{aligned} \quad (6.72)$$

To evaluate the term $E[(\boldsymbol{\mu}_i - \mathbf{m}_0)^T \boldsymbol{\Lambda}_i (\boldsymbol{\mu}_i - \mathbf{m}_0)]$, we first perform expectation w.r.t. $q^*(\boldsymbol{\mu}_i | \boldsymbol{\Lambda}_i)$ then w.r.t. $q^*(\boldsymbol{\Lambda}_i)$. Using the standard results of moments under a Gaussian, we have $E[\boldsymbol{\mu}_i] = \mathbf{m}_i$ and $E[\boldsymbol{\mu}_i \boldsymbol{\mu}_i^T] = \mathbf{m}_i \mathbf{m}_i^T + \kappa_i \boldsymbol{\Lambda}_i^{-1}$. Using these we can obtain:

$$\begin{aligned} E_{q_{(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)}}[(\boldsymbol{\mu}_i - \mathbf{m}_0)^T \boldsymbol{\Lambda}_i (\boldsymbol{\mu}_i - \mathbf{m}_0)] &= \text{Tr} \left(E_{\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i} \left[\boldsymbol{\Lambda}_i (\boldsymbol{\mu}_i - \mathbf{m}_0) (\boldsymbol{\mu}_i - \mathbf{m}_0)^T \right] \right) \\ &= \text{Tr} \left(E_{\boldsymbol{\Lambda}_i} \left[\boldsymbol{\Lambda}_i (\kappa_i^{-1} \boldsymbol{\Lambda}_i^{-1} + \mathbf{m}_i \mathbf{m}_i^T - \mathbf{m}_0 \mathbf{m}_0^T - \mathbf{m}_i \mathbf{m}_0^T + \mathbf{m}_0 \mathbf{m}_0^T) \right] \right) \\ &= K \kappa_i^{-1} + (\mathbf{m}_i - \mathbf{m}_0)^T E[\boldsymbol{\Lambda}_i] (\mathbf{m}_i - \mathbf{m}_0) \end{aligned} \quad (6.73)$$

We now just substitute the expressions $E_{q(\boldsymbol{\Lambda}_i)}[\boldsymbol{\Lambda}_i] = \nu_i \mathbf{W}_i$ and $E_{q(\boldsymbol{\Lambda}_i)}[\ln |\boldsymbol{\Lambda}_i|] = \ln \hat{\Lambda}_i$ to obtain Equ. 6.71.

Finally, we derive the expressions for $E_{q(\boldsymbol{\mu}, \boldsymbol{\Lambda})}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})]$ and $H[q(\boldsymbol{\Lambda}_i)]$ and show them to be:

$$E_{q(\boldsymbol{\mu}, \boldsymbol{\Lambda})}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \sum_{i=1}^K \left\{ \frac{1}{2} \ln \hat{\Lambda}_i + \frac{P}{2} \ln \frac{\kappa_i}{2\pi} - \frac{P}{2} - H[q(\boldsymbol{\Lambda}_i)] \right\} \quad (6.74)$$

$$H[q(\boldsymbol{\Lambda}_i)] = -\ln Z(\mathbf{W}_i, \nu_i) - \frac{(\nu_i - P - 1)}{2} \ln \hat{\Lambda}_i + \frac{\nu_i P}{2} \quad (6.75)$$

$$(6.76)$$

where the following expressions are obtained using standard results

$$\diamond Z(\mathbf{W}_i, \nu_i) = |\mathbf{W}_i|^{-\nu_i/2} \left(2^{\nu_i P/2} \pi^{P(P-1)/4} \prod_{p=1}^P \Gamma \left(\frac{\nu_i + 1 - p}{2} \right) \right)^{-1} \quad (6.77)$$

$$\diamond \ln \hat{\Lambda}_i = E_{q(\boldsymbol{\Lambda}_i)}[\ln |\boldsymbol{\Lambda}_i|] = \sum_{p=1}^P \Psi \left(\frac{\nu_i + 1 - p}{2} \right) + P \ln 2 + \ln |\mathbf{W}_i| \quad (6.78)$$

Note that Ψ is the digamma function and Ψ' is the trigamma function. To compute the entropy of Gaussian-Wishart distribution we first note that $\ln q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) = \ln q(\boldsymbol{\mu}_i | \boldsymbol{\Lambda}_i) + \ln q(\boldsymbol{\Lambda}_i)$. Now, $q(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) = q(\boldsymbol{\mu}_i | \boldsymbol{\Lambda}_i) q(\boldsymbol{\Lambda}_i)$ is distributed as $\mathcal{N}(\boldsymbol{\mu}_i | \mathbf{m}_i, (\kappa_i^{-1} \boldsymbol{\Lambda}_i^{-1})) \mathcal{W}(\boldsymbol{\Lambda}_i | \mathbf{W}_i, \nu_i)$. So when we are taking expectation w.r.t $\boldsymbol{\mu}_i$ of the expression $E[\ln q(\boldsymbol{\mu}_i | \boldsymbol{\Lambda}_i)]$, we can make use of the standard result of

the entropy of the Gaussian, $H[\mathbf{x}] = \frac{1}{2} \ln |\mathbf{\Lambda}| + \frac{P}{2}(1 + \ln(2\pi))$ if $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda})$ to yield:

$$\begin{aligned} E_{q(\boldsymbol{\mu}_i, \mathbf{\Lambda}_i)}[\ln q(\boldsymbol{\mu}_i|\mathbf{\Lambda}_i)] &= E_{q(\mathbf{\Lambda}_i)} \left[\frac{1}{2} \ln |\mathbf{\Lambda}_i| + \frac{P}{2}(\ln \kappa_i - 1 - \ln(2\pi)) \right] \\ &= \frac{1}{2} \ln \hat{\mathbf{\Lambda}}_i + \frac{P}{2}(\ln \kappa_i - 1 - \ln(2\pi)) \end{aligned} \quad (6.79)$$

The term $E_{q(\mathbf{\Lambda}_i)}[\ln q(\mathbf{\Lambda}_i)]$ is simply the negative entropy of the Wishart distribution which we write as $-H[q(\mathbf{\Lambda}_i)]$.

6.8.2 Object Bank Vocabulary

In this section, we list the labels from Object Bank[Li et al., 2010b] that intersect with the human summary vocabularies obtained from the training and the Dev-T sets. Note that tokens in a phrase like “open air” or “dish washer” etc. have been treated separately since unigrams help in the ROUGE-1 scoring of summaries. The token sets are as follows:

For the Training set:

{air, animal, art, auto, automobile, bag, ball, balloon, band, baseball, basket, basketball, bath, bathtub, be, beach, bear, bed, bench, big, bike, bird, blanket, board, boat, book, boot, bottle, bouquet, bowl, box, boy, bride, bridegroom, bridge, bubble, build, bus, button, cabinet, camera, candle, car, carpet, cat, chair, child, clock, clothe, computer, concert, couch, counter, court, cover, cow, cross, crowd, curtain, desk, dinner, dirt, dish, dishwasher, display, document, dog, door, drawer, dress, drum, duck, electric, electrical, elephant, face, facility, fan, female, fence, filter, fish, flag, floor, flower, fork, french, fridge, fruit, garage, girl, glass, glove, goggles, gown, grass, groom, guitar, hat, helmet, hook, hoop, horn, horse, human, image, individual, kid, king, kitchen, knife, lamp, lid, life, light, little, lounge, machine, male, mallet, man, mask, microphone, microwave, military, monkey, motorcycle, mount, mountain, mouse, mug, museum, musician, napkin, necklace, newspaper, ocean, of, open, outdoors, oven, paper, participant, passenger, pavement, pen, people, person, phone, picture, pillow, pin, place, plain, plane, plant, plate, playground, plunger, podium, pool, process, public, rabbit, rack, rail, refrigerator, rock, roller, room, rope, rubber, runway, sail, sand, screen, seashore, seat, set, sheet, shelf, ship, shirt, shoe, shower, sidewalk, sign, singer, sink, ski, slide, soap, soccer, sofa, someone, son, speaker, stage, stair, station, stove, step, stick, stone, stop, street, student, stump, suit, swing, switch, system, t-shirt, table, tank, teacher, tent, tire, tissue, toilet, tool, towel, tower, traffic, train, tree, truck, tub, turtle, tv, umbrella, uniform, vehicle, veil, video, wall, wash, washer, water, wave, wax, wed, wheel, window, wing, write}. Match = 9.15%

For the Dev-T set:

{air, animal, art, bag, ball, balloon, band, basket, bath, bathtub, be, beach, bear, bench, big, bike, bird, blind, board, boat, book, bookshelf, boot, bottle, bouquet, bowl, box, boy, bridal, bride, bridge, bubble, buff, build, cabinet, camera, candle, car, carpet, cat, chair, child, corsage, couch, cover, cow, cross, dish, dishwasher, display, diver, dog, door, dress, drum, duck, electric, face, fan, faucet, fence, field, fin, fish, floor, flower, french, fridge, garage, girl, glove, goggles, gown, grass, groom, guitar, hat, hook, horse, kid, king, kitchen, knife, lamp, life, light, little, machine, mallet, man, microphone, military, monkey, motorcycle, museum, necklace, ocean, of, open, outdoors, paper, pavement, pen, people, person, phone, picture, pin, place, plane, plate, playground, pool, public, rabbit, rack, radio, rail, rock, room, rope, rubber, sailboat, sand, scuba, set, shelf, shoe, sidewalk, ski, slide, snail, soccer, somebody, someone, son, stage, stair, step, stick, stone, stop, street, stump, suit, table, tank, tire, tool, towel, train, tree, truck, turtle, umbrella, vase, vehicle, video, wall, wash, washer, water, wave, wax, wed, wheel}. Match = 13.6%

In all our experiments we removed standard English stopwords and ran ROUGE [Lin and Hovy, 2003] using the following parameters: -n 2 -x -m -2.4 -u -c 95 -f A -p 0.5 -t 0 -a -d.

6.8.3 Implementation

In this section, we outline the algorithmic procedures to implement the proposed models in the paper. The baseline topic models can be implemented similarly.

Recall that for the Corr-MMGLDA model,

$E_q[\ln p(\mathbf{w}_M | \mathbf{z}_{\mathbf{y}_M}, \beta)]$ expands out to be:

$$\sum_{m=1}^{M_d} \sum_{k=1}^K \left(\sum_{o=1}^{O_d} \phi_{d,m,o} \phi_{d,o,k}^{(O)} \right) \ln \beta_{k,w_{d,m}} \quad (6.80)$$

We mentioned in the paper that equation (6.80) is both the bottleneck and the strong point of the correspondence class of the LDA models. It is a computational bottleneck because finding the confidence of the word w_m over topic i necessitates the elimination of uncertainties of w_m 's dependence on w_o and w_o 's dependence on topic i . This is also a strong point since the summation suggests a stronger influence of a topic on a summary word if that influence is justified by most w_o s.

The computational burden for the correspondence class of models increase precisely for the inner sum appearing in the variational distribution update expressions in the E-step. For a sufficiently large number of topics, K , and a sufficiently large number of corresponding observations O , Corr-MMGLDA is computationally more expensive. If these conditions are not true, as is the case in our experiments, we observed that both models took approximately the same training time.

The rest of this supplementary material outlines the key steps for implementing the multimedia topic models. We have used the well tested GNU Scientific Library⁷ (GSL) to implement the linear algebraic operations but the rest of the code is developed in-house and written in C++. Note that the tractable mean field distribution vectors for $\phi^{(H)}$ and $\phi_{d,m,i}$ or $\phi_{d,m,o}$ can be thrown out as each document is processed but the $\phi^{(O)}$ s need to be stored and updated for each iteration since they are needed in the Maximization step for updating the priors for the Gaussians.

6.8.4 Algorithm and Pseudocodes

ALGORITHM 2: vb_em

- 1: **if** *algorithm_mode* == "training" **then**
- 2: initialize_statistics(); {use random or seeded initialization for ρ and β and random initialization for \mathbf{m} and \mathbf{W} from randomly initialized real valued data vectors}
- 3: vb_m_step();
- 4: **end if**
- 5: **if** *algorithm_mode* == "predict" **then**
- 6: load up the parameters; { ρ , \mathbf{m} , \mathbf{W} , κ and ν s are modified through discrete and real valued distributions in test set and the *trained* β is accessed only after convergence to obtain word predictions. A simple to enforce the latter is to set up a dummy textual word index and set its count to 0. All $\beta_{i,j}$ s will be 0 in this way which is equivalent to not looking at β at all.}
- 7: **end if**

⁷<http://www.gnu.org/software/gsl/>

```

8:  $elbo\_prev \leftarrow 0$ 
9:  $elbo\_current \leftarrow 0$ ;
10:  $iters \leftarrow 0$ 
11: while NOT_CONVERGED do
12:    $elbo\_current \leftarrow vb\_e\_step()$  {update hidden variables}
13:    $vb\_m\_step()$  {update model parameters}
14:    $converged \leftarrow (elbo\_prev - elbo\_current)/(elbo\_prev)$ 
15:    $elbo\_prev \leftarrow elbo\_current$ 
16:    $iters \leftarrow iters + 1$ 
17: end while

```

ALGORITHM 3: *vb_e_step*

```

1: zero_initialize_statistics();
2:  $elbo\_current \leftarrow 0$ 
3: for  $d = 1$  to  $D$  do
4:    $doc \leftarrow corpus.video\_document\_vec[d]$ 
5:    $elbo\_current += doc\_e\_step(d, doc)$ 
   {also accumulate sufficient statistics for  $\rho_i, \beta_i, N_i, \bar{x}_i$  and  $S_i$  for each topic  $i$ .
   The doc_e_step() routine updates  $\gamma_{d,i}, \phi_{d,h,i}^{(H)}, \phi_{d,o,i}^{(O)}$  and  $\phi_{d,m,i}$  or  $\phi_{d,m,o}$  based on whether the model is
   MMGLDA or Corr-MMGLDA.}
   {Update sufficient statistics for alpha as follows;}
6:   if symmetric_dirichlet then
7:      $gamma\_sum \leftarrow 0$ 
8:     for  $k = 1 \rightarrow K$  do
9:        $gamma\_sum += \gamma[d][k]$ 
10:       $alpha\_ss += \Psi(\gamma[d][k])$  {alpha_ss holds sufficient statistics for symmetric alpha}
11:     end for
12:      $alpha\_ss - = K \times \Psi(gamma\_sum)$  { $\Psi(\cdot)$  is the digamma function}
13:   end if
14:   if asymmetric_dirichlet then
15:      $gamma\_sum \leftarrow 0$ 
16:     for  $k = 1 \rightarrow K$  do
17:        $gamma\_sum += \gamma[d][k]$ 
18:        $alpha\_ss[k] += \Psi(\gamma[d][k])$ 
19:        $alpha\_ss\_exp\_aux[k] = \Psi(\gamma[d][k])$  {alpha_ss_exp_aux is an auxilliary array}
20:     end for
21:     for  $k = 1 \rightarrow K$  do
22:        $alpha\_ss[k] - = \Psi(gamma\_sum)$ 
23:        $alpha\_ss\_exp\_aux[k] - = \Psi(\gamma[d][k])$ 
24:     end for
25:     for  $k = 1 \rightarrow K$  do
26:        $expo \leftarrow \exp(alpha\_ss\_exp\_aux[k])$ 
27:        $alpha\_ss\_exp[k] += expo$ 
28:        $alpha\_ss\_exp\_square[k] += expo \times expo$  {alpha_ss_exp and alpha_ss_exp_square hold sufficient
       statistics for asymmetric alpha}
29:     end for
30:   end if

```

```

31: end for
32: for  $k = 1 \rightarrow K$  do
33:    $\text{elbo\_current} += E_{q[\mu_i, \Lambda_i]} [\ln p(\mu_i, \Lambda_i)]$ 
34:    $\text{elbo\_current} -= E_{q[\mu_i, \Lambda_i]} [\ln q(\mu_i, \Lambda_i)]$ 
35: end for
36: return  $\text{elbo\_current}$ ;

```

ALGORITHM 4: doc_e_step

```

1:  $\gamma_{d,i} = \alpha + \frac{(\text{doc.total\_num\_words} + \text{doc.total\_num\_corr\_words} + \text{doc.num\_real\_valued\_observations})}{K}$ 
2:  $\phi_{d,h,i}^{(H)} = \frac{1.0}{K}$ 
3:  $\phi_{d,o,i}^{(O)} = \frac{1.0}{K}$ 
4:  $\phi_{m,i} = \frac{1.0}{K}$  {If model is MMGLDA OR}
    $\phi_{m,o} = \frac{1.0}{\text{doc.num.real.valued.observations}}$  {If model is Corr-MMGLDA}
5:  $\text{elbo\_current} \leftarrow 0$ ;
6: while not converged do
7:   update  $\phi_{d,h,i}^{(H)}$ 
8:   update  $\phi_{d,o,i}^{(O)}$ 
9:   update  $\phi_{d,m,i}$  {If model is MMGLDA} {OR} update  $\phi_{d,m,o}$  {If model is Corr-MMGLDA}
10:  update  $\gamma_{d,i}$ 
11:   $\text{elbo\_current} \leftarrow \text{compute\_likelihood}()$  {To compute likelihoods use the expressions in  $\mathcal{L}_{(MMGLDA)}$  for MMGLDA or  $\mathcal{L}_{(Corr-MMGLDA)}$  for Corr-MMGLDA}
12: end while
13: return  $\text{elbo\_current}$ ;

```

ALGORITHM 5: vb_m_step

```

1: for all  $i \in 1, \dots, K, v \in 1, \dots, V, \text{corr}_v \in 1, \dots, \text{corr}V$  do
2:   update  $\rho_{i,\text{corr}_v}$  and  $\beta_{i,v}$  from sufficient statistics
3:   update  $\mathbf{m}_i, \mathbf{W}_i, \kappa_i$  and  $\nu_i$  from sufficient statistics
4:   update  $\alpha$  {For symmetric  $\alpha$ , several publicly available VB implementations corresponding to [Blei et al., 2003] can be used as a blackbox; for asymmetric  $\alpha$ , algorithms 4 through 7 implement the Newton Raphson method of optimizing the  $\alpha$  parameter using the derivations in [Blei et al., 2003] and [Minka, 2009] }
5: end for

```

ALGORITHM 6: initialize_guesses

```

1: input:  $\text{alpha\_ss\_exp}, \text{alpha\_ss\_exp\_square}, D$  and  $K$ 
2: output:  $\text{init}_a$  - vector containing initial starting points for optimizing  $\alpha[k]$  obtained from data
3: for  $k = 1 \rightarrow K$  do
4:    $\text{init}_a[k] \leftarrow \text{alpha\_ss\_exp}[k]/D$ 
5:    $m[k] \leftarrow \text{alpha\_ss\_exp\_square}[k]/D$ 
6:    $s[k] \leftarrow (\text{init}_a[k] - m[k]) / (m[k] - \text{init}_a[k] \times \text{init}_a[k])$ 
7: end for
8:  $\text{median} \leftarrow \text{median}(s)$ 
9: if  $\text{median} > 0$  then
10:  for  $k = 1 \rightarrow K$  do
11:     $\text{init}_a[k] \leftarrow \text{init}_a[k] \times \text{median}$ 
12:  end for

```

13: **end if**
14: return *init_a*
15: {Using this initialization lowers the number of iterations in `optimize_asym_dirichlet()` by at least a factor of 2-3}

ALGORITHM 7: *asym_alpha_lhood*

1: input: *alpha*, *alpha_ss*, *D* and *K*
2: *sum_alpha* \leftarrow *sum(alpha)*
3: *lhood* \leftarrow $\ln \Gamma(\text{sum_alpha})$
4: **for** $k = 1 \rightarrow K$ **do**
5: *lhood*+ = (*alpha*[*k*] - 1) \times *alpha_ss*[*k*]/*D* - $\ln \Gamma(\text{alpha}[k])$
6: **end for**
7: return *lhood*

ALGORITHM 8: *compute_H.inverse_g*

1: input: *alpha_guess*, *g*, *K*
2: output: *hg* - vector containing Hessian inverse \times gradient for each component
3: *sum_a* \leftarrow *sum(alpha_guess)*
4: *q* \leftarrow *alpha_guess*
5: **for** $k = 1 \rightarrow K$ **do**
6: *q*[*k*] \leftarrow $1.0 / (-\text{trigamma}(\text{alpha_guess}[k]))$
7: **end for**
8: *z* \leftarrow *trigamma(sum_a)*
9: *sum_q* \leftarrow *sum(q)*
10: *sum_gq* \leftarrow 0
11: **for** $k = 1 \rightarrow K$ **do**
12: *sum_gq*+ = *g*[*k*] \times *q*[*k*]
13: **end for**
14: *b* \leftarrow *sum_gq*/(1.0/*z* + *sum_q*)
15: **for** $k = 1 \rightarrow K$ **do**
16: *hg*[*k*] \leftarrow (*g*[*k*] - *b*) \times *q*[*k*]
17: **end for**
18: return *hg* {for equations concerning derivations, see [Blei et al., 2003]}

ALGORITHM 9: *optimize_asym_dirichlet*

1: input: *alpha_ss*, *alpha_ss_exp*, *alpha_ss_exp_square*, *D* and *K*
2: output: **a** - which is the optimized α
3: *a* \leftarrow *initialize_guesses(alpha_ss_exp, alpha_ss_exp_square, D, K)*
4: initialize *K*-dimensional vectors *g*, *hg* and *a_minus_hg* to 0
5: *old_lh* \leftarrow *D* \times *asym_alpha_lhood(a, alpha_ss, D, K)*
6: $\epsilon \leftarrow$ *DBL_EPSILON*
7: $\lambda \leftarrow$ 0.1
8: *max_iter* \leftarrow 100
9: **for** *iter* = 1 \rightarrow *max_iter* **do**
10: *old_a* \leftarrow *a*; *sum_a* \leftarrow *sum(a)*
11: **if** *sum_a* == 0 **then**
12: **break**

```

13: end if
14: for  $k = 1 \rightarrow K$  do
15:    $g[k] \leftarrow \Psi(\text{sum}_a) - \Psi(a[k]) + \alpha_{ss}[k]/D$ 
16: end for
17:  $\text{break\_flag} \leftarrow \text{false}$ 
18: while true do
19:    $hg \leftarrow \text{compute\_H\_inverse\_g}(a, g, K)$ 
20:   if  $hg < a$  for all dimensions then
21:     for  $k = 1 \rightarrow K$  do
22:        $a\_minus\_hg[k] \leftarrow (a[k] - hg[k])$ 
23:     end for
24:      $\text{new\_lh} \leftarrow D \times \text{asym\_alpha\_lhhood}($ 
        $a\_minus\_hg, \alpha_{ss}, D, K)$ 
25:     if  $\text{new\_lh} > \text{old\_lh}$  then
26:        $\text{old\_lh} \leftarrow \text{new\_lh}$ 
27:        $a \leftarrow a\_minus\_hg$ 
28:        $\lambda \leftarrow \lambda/C$  { $C$  was set to 10 in this paper}
29:       break
30:     end if
31:   end if
32:    $\lambda \leftarrow \lambda \times C$  { $C$  was set to 10 in this paper}
33:   if  $\lambda > \text{Large\_Constant}$  then
34:      $\text{break\_flag} \leftarrow \text{true};$  break;
35:   end if
36: end while
37: if  $\text{break\_flag}$  then
38:    $\text{new\_lh} \leftarrow \text{old\_lh}$ 
39:   break
40: end if
41: for  $k = 1 \rightarrow K$  do
42:   if  $a[k] < \epsilon$  then
43:      $a[k] \leftarrow \epsilon$ 
44:   end if
45: end for
46:  $\text{max\_abs\_a\_minus\_olda} \leftarrow \text{max}(\text{abs}(a - \text{old}_a))$ 
47: if  $\text{max\_abs\_a\_minus\_olda} < 10^{-10}$  then
48:   break
49: end if
50: end for
51: return  $a$ 

```

6.8.5 Important low level features from videos

6.8.5.1 HOG3D FEATURES

The low level HOG3D features for our YouCook videos are obtained using the open source software provided by Klaeser et al. [Kläser et al., 2008]⁸. The codebooks are computed in a standard way using the K-Means algorithm⁹. We outline the main steps in Algo. 8:

Algorithm 8 Codebook computation: a general outline

- 1: Compute HOG3D features from frames sampled from the videos in the training set. Listings 6.1 and 6.2 show the Matlab code for extracting the feature vectors corresponding to a sample training video using the method described in [Kläser et al., 2008].
 - 2: Compute K-Means over the HOG3D features of the training dataset (see Listing 6.3). A maximum of 10,000 feature vectors (each a 300-dimensional descriptor) for each video is obtained and concatenated for all videos in the training set to form the initial dataset for K-Means clustering. After this, clustering is performed for a maximum of 100 iterations and the codebook vectors are saved.
 - 3: A K-dimensional histogram of HOG3D features are computed for each video (training or test) using the codebook obtained in Step 2 and its HOG3D feature descriptor (see Listing 6.4).
-

After the first step in Algo. 8, we obtain large feature files for each video where each line in output of the HOG3D feature extractor corresponds to a frame of the video and consists of a 308-dimensional vector in the following format:

```
<x> <y> <frame> <x-norm.> <y-norm.> <t-norm.> <xy-scale> <t-scale>  
<descriptor>
```

The “descriptor” is the actual 300-dimensional action descriptor that is ultimately used in codebook computation. The stride for the spatial scale is set to nine and that for the temporal scale is set to five. We spatially rescale the video such that the larger of its width and height is 160 pixels.

There is also a [xy/t]-max-scale option that controls at which spatial/temporal scale the sampling is stopped. We set this option’s value to one to reduce computational time.

6.8.5.2 COLOR HISTOGRAM FEATURES

The low level color histogram features are computed in an efficient way from the frames obtained by using the ffmpeg command:

```
ffmpeg -i <movie-file> -y <output-directory>/frame%05d.jpg
```

Listing 6.5 shows the Java code for extracting color histogram for an image using an user specified number of bins for each of the red, green and blue components. We set the number of such bins for each component to be eight for a total of 512-dimensional color histogram for each frame of the video. Histograms from all such frames are concatenated to form the color histogram for the entire video. The text file containing concatenated features for a video are stored on disk in g-zipped format. For use in topic models, we use the average of the values of the color combinations (i.e. the bins) from all of these histograms as the color histogram descriptor of a single video.

⁸http://lear.inrialpes.fr/people/klaeser/software_3d_video_descriptor

⁹<http://crvc.ucf.edu/source/K-Means>

When we use the color histogram features in the topic model, we only use the bins that contain values within the 15th to 85th percentile of the values in the bins. The higher values need to be removed since they often increase the magnitude of the slopes around the fixed points in the fixed point iterations of the local (video document specific) E-step in variational Bayes optimization framework to be more than one and thus leading to non-convergence [Conte and Boor, 1980] and possible degeneration of the proportions of a few topic components. The lower values are removed to offset the bias in removing the higher band. The rest of the bin values are normalized to lie in [0, 255].

6.8.6 Code snippets

Listing 6.1: Driver for obtaining HOG3D histogram

```

cookingVidPaths = ['videos/0002.mp4']; % add more here      1
tmpPath='ffprobe_tmp.txt';                                2
featureDir='feature';                                     3
                                                            4
c = cellstr(cookingVidPaths);                              5
nVids = size(c);                                          6
mkdir(featureDir);                                       7
for i = 1: nVids                                          8
    videoPath = c{i};                                     9
    HOG3DExtractor(videoPath, tmpPath, featureDir);       10
end                                                       11

```

Listing 6.2: Extracting HOG3D features for a video using the static executable in [Kläser et al., 2008]

```

function xyscale = HOG3DExtractor(videoPath, tmpPath, featureDir) 1
% videoPath: path to the input video                             2
% tmpPath: output of "ffprob -show_streams videoPath > tmpPath". 3
% featureDir: Directory that you want to save feature.         4
%                                                                 5
addpath('.');                                                  6
% extract width and height                                       7
[~, name, ~] = fileparts(videoPath);                             8
fprintf('Processing video %s... \n', videoPath);                 9
system(sprintf('ffprobe -show_streams %s > %s', videoPath, tmpPath)); 10
[~, grep_width] = system(sprintf('grep "width=*" %s', tmpPath)); 11
[~, grep_height] = system(sprintf('grep "height=*" %s', tmpPath)); 12
width = sscanf(grep_width, 'width=%d');                         13
height = sscanf(grep_height, 'height=%d');                     14
max_w_h = max(width, height);                                    15
xyscale = -1;                                                  16
if(any(size(max_w_h) == 0))                                     17
    fprintf('!!! error in ffprobe of %s. !!!\n', name);         18
    return;                                                      19
end                                                            20
xy_srtride = 9;                                               21
t_stride = 5;                                                 22
% compute rescaling ratio                                       23
xyscale = 160 / max_w_h;                                        24
featurePath = sprintf('%s/%s.txt', featureDir, name);          25

```



```

system(sprintf('./extractFeatures_static --simg %f --xy-stride %d --xy-max-scale 1 --t 26
-stride %d --t-max-scale 1 %s > %s', xyscale, xy_srtride, t_stride, videoPath,
featurePath));
featureInfo = dir(featurePath);
if( ~exist(featurePath, 'file') || featureInfo.bytes == 0)
    fprintf('!!! error in feature extraction of file %s. !!!\n', name);
    return;
end
fprintf('video %s was processed successfully.\n\n', videoPath);
ffprobe_file = tmpPath
delete(ffprobe_file);
rmpath('.');

```

Listing 6.3: Computing HOG3D codebook vectors using K-Means

```

cookingHOG3DPaths = ['/data/Cooking/features/HOG3D/tmp/0002.txt']; % add more here
c = cellstr(cookingHOG3DPaths);
nVids = size(c);
D = [];
nFrames = 10000;
for i = 1: nVids
    HOG3DPath = c{i};
    fprintf('loading file %s\n',HOG3DPath);
    A = load(HOG3DPath);
    P = A(:,9:308)';
    [row, col] = size(P);
    if ( col > nFrames )
        interval = floor(col/nFrames);
    else
        interval = 1;
    end
    index = [0:interval:col-1]+1;
    D = [D P(:,index)];
end
nClusters = 1000;
[CX, sse] = vgg_kmeans(D, nClusters, 'maxiters', 100);
dlmwrite('/data/Cooking/features/HOG3D/1000CW_dictionary.txt', CX');

```

Listing 6.4: Quantizing a video using a pre-computed HOG3D K-means codebook

```

%
% videoPath: path to the input video
% tmpPath: output of "ffprob -show_streams videoPath > tmpPath".
% featureDir: Directory that you want to save feature.
% codebookPath: address of the coreword
% wordDir: directory to save the quantized feature
% histDir: directort to save the global histogram of HOG3D
%
function xyscale = quantizeVideo(videoPath, tmpPath, featureDir, codebookFile, wordDir
, histDir)

```

```

addpath('/sfw/HOG3D/VGG_KMeans');
10
11
12
scanFormat = '%d %d %d %f %f %f %d %d';
13
for i = 1 : 300
14
    if i ~= 300
15
        scanFormat = [scanFormat, '%f'];
16
    else
17
        scanFormat = [scanFormat, '%f'];
18
    end
19
end
20
21
% read codeword dictionary
22
clusterCen = load(codebookFile);
23
24
% extract width and height
25
[pathstr, name, ext] = fileparts(videoPath);
26
fprintf('Processing video %s... \n', videoPath);
27
system(sprintf('ffprobe -show_streams %s > %s', videoPath, tmpPath));
28
[status grep_width] = system(sprintf('grep "width=*" %s', tmpPath));
29
[status grep_height] = system(sprintf('grep "height=*" %s', tmpPath));
30
31
width = sscanf(grep_width, 'width=%d');
32
height = sscanf(grep_height, 'height=%d');
33
34
max_w_h = max(width, height);
35
xyscale = -1;
36
if(any (size(max_w_h) == 0))
37
    fprintf('!!! error in ffprobe of %s. !!!\n', name);
38
    return;
39
end
40
41
xy_srtride = 9;
42
t_stride = 5;
43
% compute rescaling ratio
44
xyscale = 160 / max_w_h;
45
46
featureFile = sprintf('s/%s.txt', featureDir, name);
47
if exist(featureFile, 'file')
48
    system(sprintf('rm -f %s', featureFile));
49
end
50
system(sprintf('sfw/HOG3D/hog3dcode/extractFeatures_static --simg %f --xy-stride %d
--xy-max-scale 1 --t-stride %d --t-max-scale 1 %s > %s', xyscale, xy_srtride,
t_stride, videoPath, featureFile));
51
52
ftrInfo = dir(featureFile);
53
if( ~exist(featureFile, 'file') || ftrInfo.bytes == 0)
54
    fprintf('!!! error in feature extraction of file %s. !!!\n', name);
55
    return;
56
end
57
58
% read features, save their indeces i.e. Quantize the extracted features using the pre
-computed words
59
fid = fopen(featureFile);
60

```

```

A = fscanf(fid, scanFormat, [308 100000]);           61
word2 = cell(100,1);                               62
i = 1;                                              63
while ~any(size(A)==0)                             64
    [ind d2] = vgg_nearest_neighbour(A(9:end,:), clusterCen'); 65
    word2{i} = [A(1:3, :)]' ind];                 66
    i = i+1;                                       67
    A = fscanf(fid, scanFormat, [308 100000]);     68
end                                                69
words = cat(1, word2{:});                          70
fclose(fid);                                       71

system(sprintf('rm -f %s', featureFile));          72
                                                73
fprintf('saving the words.\n');                    74
% compute histogram from the quantized features    75
hist = zeros(size(clusterCen, 1), 1);             76
for i=1:size(words,1)                             77
    hist(words(i,4)) = hist(words(i,4)) + 1;      78
end                                                79
save_word_name = sprintf('%s/%s.txt', wordDir, name); 80
dlmwrite(save_word_name, words);                  81
                                                82
% This is what we are interested in ultimately    83
save_hist_name = sprintf('%s/%s.txt', histDir, name); 84
dlmwrite(save_hist_name, hist);                   85
                                                86
fprintf('video %s was processed successfully.\n\n', videoPath); 87
rmpath('/sfw/HOG3D/VGG_KMeans');                 88
                                                89

```

Listing 6.5: Java code for computing color histogram from an image

```

public static int[] binnedImageHistogram(BufferedImage input, int nBins) 1
{                                                                 2
    int[] hist = new int[nBins*nBins*nBins];                    3
    int interval = 256/nBins; int nBinsSquared = nBins*nBins;  4
    int W = input.getWidth(); int H = input.getHeight();       5
    for(int i = 0; i < W; ++i) {                                6
        for(int j = 0; j < H; ++j) {                            7
            int red = new Color(input.getRGB (i, j)).getRed(); // 0 to 255  8
            int green = new Color(input.getRGB (i, j)).getGreen(); // 0 to 255  9
            int blue = new Color(input.getRGB (i, j)).getBlue(); // 0 to 255 10
                                                                 11
            int rBin = red/interval;                             12
            int bBin = blue/interval;                             13
            int gBin = green/interval;                             14
            int coord = rBin*nBinsSquared + gBin*nBins + bBin;  15
                                                                 16
            hist[coord]++;                                       17
        }                                                         18
    }                                                             19
    return hist;                                                20
}                                                                 21

```

Chapter 7

Conclusion and Talking Points

“A decent thesis makes a slight impact. A good thesis transforms you. Anything more tries to transform mankind through you.”
- Pradipto Das, after completing his thesis.

Up to this point, the materials covered in this thesis have primarily focused on three important questions:

- [i] How do we create *faceted* latent topics i.e. topics conditional on different types of word annotations which are influenced by document level meta data? This question is very important since often times fine grained supervised models such as those producing automatic content annotations and meta data work at a document or sentence level. On the other hand, unsupervised corpus centric models of topical analysis work at a “global” corpus level and are competitive rather than collaborative with the finer grained local models.
- [ii] How can we apply topic models and discourse analysis through rhetorical parsing to create bullet list summaries?
- [iii] Is the summarization problem intrinsically important? In other words, when we look at the world around us, *do we speak all that we see?*

We now emphasize some talking points which can be a basis for related research within the foreseeable horizon. Touching upon the first point, the issue of automatically deciding the number of topics using non-parametric Bayes has been an active area of research in the context of the LDA model [Gershman et al., 2012, Wang and Blei, 2009, Graber and Blei, 2009, McAuliffe et al., 2006, Blei and Jordan, 2004]. Non-parametric extensions of multimodal LDA has received much less research focus [Yakhnenko and Honavar, 2009]. Hybrid models like the Tag²LDA models described in this thesis will benefit much from non-parametric extensions to decide on the number of topics automatically and is indeed an important research direction. Further, online tag-topic models with infinite vocabulary similar in spirit to [Zhai and Boyd-Graber, 2013], integrating basic ideas of tag-topic models to large scale deep belief networks [Bartlett et al., 2012] and their connection to transfer learning [Pan and Yang, 2010] all give rise to very interesting research areas.

On the problem of bullet list summarization, there is ample research opportunity on automatically identifying upon rhetorical relations in a semi-supervised manner. Additionally, ranking of bullet lists incorporating diversity [Lin and Bilmes, 2012, Kulesza and Taskar, 2012] to reduce redundancy is an obvious avenue of research.

On the problem of video analytics, particularly with the capability of generating lingual descriptions from videos, there is a natural direction in analyzing the *social structure of videos* along the following premises:

- [i] What is the central theme of the video and who is/are the main subject/s of the video? This answers questions about role discovery of objects or concepts *within* videos.
- [ii] What is more interesting is to understand the intentions of the person shooting the video. A person rarely shoots a video of something that he/she is not interested/curious about. If we have several videos that a person has shot over a period of time, can we understand what general things he is interested on during that time (this is answered in part by the first point). How do those interests subside over time and what new interests arise? Are there similar people sharing similar intentions? How do we define a distribution over such intention space with the mass/density being how many people share the same intention? What lies in the tail of such distribution?
- [iii] How do we evaluate such an intention space automatically?

The applications of the models and ideas developed in the course of this thesis can have profound consequences in the way we interact with our surroundings on a day to day basis both on the visual as well as on the textual front.

Bibliography

- [Abu-Mostafa et al., 2012] Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning From Data*. AMLBook. 33, 34, 35
- [Ahmed et al., 2009] Ahmed, A., Xing, E. P., Cohen, W. W., and Murphy, R. F. (2009). Structured correspondence topic models for mining captioned figures in biological literature. In *Proc of the SIGKDD Conference*, pages 39–48. 188
- [Andrzejewski et al., 2009] Andrzejewski, D., Zhu, X., and Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML*, pages 25–32, New York, NY, USA. ACM. 66, 69
- [Arora et al., 2013] Arora, S., Ge, R., Halpern, Y., Mimno, D. M., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 28(2), pages 280–288. JMLR: W&CP. 63, 91, 115
- [Arora et al., 2012] Arora, S., Ge, R., and Moitra, A. (2012). Learning topic models – going beyond svd. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, FOCS '12*, pages 1–10, Washington, DC, USA. IEEE Computer Society. 63
- [Atreya and Elkan, 2011] Atreya, A. and Elkan, C. (2011). Latent semantic indexing (lsi) fails for trec collections. *SIGKDD Explor. Newsl.*, 12(2):5–10. 10
- [Banko and Vanderwende, 2004] Banko, M. and Vanderwende, L. (2004). Using n-grams to understand the nature of summaries. In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Short Papers*, pages 1–4, Boston, Massachusetts, USA. Association for Computational Linguistics. 148
- [Bao et al., 2009] Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y. (2009). Joint emotion-topic modeling for social affective text mining. *IEEE International Conference on Data Mining*, 0:699–704. 110
- [Barbu et al., 2012] Barbu, A., Bridge, A., Burchill, Z., Coroian, D., Dickinson, S. J., Fidler, S., Michaux, A., Mussman, S., Narayanaswamy, S., Salvi, D., Schmidt, L., Shangquan, J., Siskind, J. M., Waggoner, J. W., Wang, S., Wei, J., Yin, Y., and Zhang, Z. (2012). Video in sentences out. In *UAI*. 206

- [Bartlett et al., 2012] Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors (2012). *Large Scale Distributed Deep Networks*. 233
- [Barzilay and Lapata, 2005a] Barzilay, R. and Lapata, M. (2005a). Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 141–148. Association for Computational Linguistics. 83, 108, 109, 122
- [Barzilay and Lapata, 2005b] Barzilay, R. and Lapata, M. (2005b). Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 141–148, Stroudsburg, PA, USA. Association for Computational Linguistics. 150
- [Beal, 2003] Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London. 43, 46, 56, 61, 86, 115, 191
- [Belz and Reiter, 2006] Belz, A. and Reiter, E. (2006). Comparing automatic and human evaluation of nlg systems. In *EACL*. 213
- [Berg et al., 2012] Berg, A. C., Berg, T. L., III, H. D., Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Sood, A., Stratos, K., and Yamaguchi, K. (2012). Understanding and predicting importance in images. In *CVPR*. 206, 211
- [Berger et al., 1996] Berger, A. L., Pietra, S. A. D., and Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *COMPUTATIONAL LINGUISTICS*, 22:39–71. 20
- [Berger, 1985] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York. 27
- [Bilinski and Bremond, 2011] Bilinski, P. and Bremond, F. (2011). Evaluation of local descriptors for action recognition in videos. In *Proceedings of ICCV Conference*. 196, 212
- [Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 1, 44, 46, 66, 69, 73
- [Blei and Lafferty, 2005] Blei, D. and Lafferty, J. (2005). Correlated topic models. In *Advances in Neural Information Processing Systems*. 92, 188
- [Blei and McAuliffe, 2008] Blei, D. and McAuliffe, J. (2008). Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20. 77
- [Blei, 2004] Blei, D. M. (2004). *Probabilistic Models of Text and Images*. PhD thesis, University of California, Berkeley. 9, 10
- [Blei and Jordan, 2003] Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th ACM SIGIR Conference*, pages 127–134. xv, 14, 111, 114, 187, 189, 190, 192, 205, 206, 207, 218

- [Blei and Jordan, 2004] Blei, D. M. and Jordan, M. I. (2004). Variational methods for the dirichlet process. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM. 233
- [Blei and Lafferty, 2006] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 113–120, New York, NY, USA. ACM. 2
- [Blei and Lafferty, 2009] Blei, D. M. and Lafferty, J. D. (2009). Visualizing Topics with Multi-Word Expressions. <http://arxiv.org/abs/0907.1013>. 7
- [Blei and Mcauliffe, 2007] Blei, D. M. and Mcauliffe, J. D. (2007). Supervised topic models. In *NIPS*, volume 21. 15, 111, 113, 114, 121, 123
- [Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022. 3, 7, 10, 11, 17, 31, 35, 42, 43, 61, 67, 68, 75, 86, 87, 92, 101, 111, 113, 115, 121, 138, 141, 147, 149, 151, 152, 154, 185, 187, 192, 196, 207, 225, 226
- [Blitzer et al., 2007] Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th ACL Conference*, pages 440–447, Prague, CZ. 109, 121
- [Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA. 52
- [Boyd-Graber, 2010] Boyd-Graber, J. (2010). Personal communications. 154
- [Bradley and Lang, 1999] Bradley, M. and Lang, P. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. 109, 124
- [Brennan et al., 1987] Brennan, S. E., Friedman, M. W., and Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics, ACL '87*, pages 155–162, Stroudsburg, PA, USA. Association for Computational Linguistics. 80
- [Buttcher et al., 2010] Buttcher, S., Clarke, C. L., and Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press. 1, 6, 10, 97
- [Cao and Fei-Fei, 2007] Cao, L. and Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*. 205
- [Carbonell and Goldstein, 1998] Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98*, pages 335–336, New York, NY, USA. ACM. 169
- [Casella and Berger, 2001] Casella, G. and Berger, R. (2001). *Statistical Inference*. Duxbury Resource Center. 23, 29, 30, 31, 32, 33, 37

- [Celikyilmaz and Hakkani-Tür, 2011] Celikyilmaz, A. and Hakkani-Tür, D. (2011). Discovery of topically coherent sentences for extractive summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11*, pages 491–499, Stroudsburg, PA, USA. Association for Computational Linguistics. 104, 141, 147, 181
- [Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. 163, 202
- [Chang et al., 2009] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*. 35, 63, 113, 154, 158
- [Chen et al., 2009] Chen, D., Tang, J., Yao, L., Li, J., and Zhou, L. (2009). Query-focused summarization by combining topic model and affinity propagation. In *Proceedings of the Joint International Conferences on Advances in Data and Web Management, APWeb/WAIM '09*, pages 174–185, Berlin, Heidelberg. Springer-Verlag. 77, 79, 147
- [Conroy et al., 2010] Conroy, J. M., Schlesinger, J. D., Rankel, P., and O’Leary, D. P. (2010). Guiding classy toward more responsive summaries. In *Proceedings of the Third Text Analysis Conference (TAC 2010) – Guided Summarization Track*. 140, 148, 170, 181
- [Conroy et al., 2011] Conroy, J. M., Schlesinger, J. D., Rankel, P., and O’Leary, D. P. (2011). Classy 2011 at tac: Guided and multi-lingual summaries and evaluation metrics. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011) – Guided Summarization Track*. 140, 148, 170, 175
- [Conte and Boor, 1980] Conte, S. D. and Boor, C. W. D. (1980). *Elementary Numerical Analysis: An Algorithmic Approach*. McGraw-Hill Higher Education, 3rd edition. 25, 68, 229
- [Cover and Thomas, 2006] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience. 21, 67
- [Dang, 2006a] Dang, H. T. (2006a). Duc 2005: evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering, SumQA '06*, pages 48–55. ACL. 108
- [Dang, 2006b] Dang, H. T. (2006b). Duc 2005: evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 48–55. ACL. 142
- [Darling, 2010] Darling, W. M. (2010). Multi-document summarization from first principles – university of guelph. In *Proceedings of the Third Text Analysis Conference (TAC 2011) – Guided Summarization Track*, Gaithersburg, Maryland, USA. 148
- [Das and Srihari, 2009] Das, P. and Srihari, R. (2009). Learning to summarize using coherence. In *NIPS '09 Workshop: Applications of topic models: Text and Beyond*. 17, 77, 108

- [Das et al., 2011] Das, P., Srihari, R., and Fu, Y. (2011). Simultaneous joint and conditional modeling of documents tagged from two perspectives. In *Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11*, pages 1353–1362, New York, NY, USA. ACM. 3, 18, 141, 143, 144, 146, 147, 149, 151, 154, 170
- [Das and Srihari, 2011] Das, P. and Srihari, R. K. (2011). Global and local models for multidocument summarization. In *Text Analysis Conference (TAC) Workshop on Summarization*. 3, 18
- [Das and Srihari, 2013] Das, P. and Srihari, R. K. (2013). Using tag-topic models and rhetorical structure trees to generate bulleted list summaries. In *Submission*. 18
- [Das et al., 2013a] Das, P., Srihari, R. K., and Corso, J. J. (2013a). Translating related words to videos and back through latent topics. In *Proceedings of the 6th ACM WSDM Conference*. 4, 18
- [Das et al., 2013b] Das, P., Xu, C., Doell, R. F., and Corso, J. J. (2013b). A thousand frames in just a few words: generating lingual descriptions of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE CVPR Conference*. 4, 18
- [Daumé and Marcu, 2006] Daumé, III, H. and Marcu, D. (2006). Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44*, pages 305–312, Stroudsburg, PA, USA. Association for Computational Linguistics. 141, 147
- [Daumé III and Marcu, 2006] Daumé III, H. and Marcu, D. (2006). Bayesian query-focused summarization. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Sydney, Australia. 79, 93
- [Delort and Alfonseca, 2012] Delort, J.-Y. and Alfonseca, E. (2012). Dualsum: A topic-model for update summarization. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, Avignon, France. 141, 147, 148, 155, 178
- [Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38. 39, 44
- [Douze et al., 2009] Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., and Schmid, C. (2009). Evaluation of gist descriptors for web-scale image search. In *Proceedings of the International CIVR Conference*, pages 19:1–19:8. 197
- [DUC, 2007] DUC (2005, 2006, 2007). Publications for different tracks in the document understanding conference. <http://www-nlpir.nist.gov/projects/duc/pubs.html>. 78
- [DUC, 2008] DUC (2008). Publications for different tracks in the text analysis conference. <http://www.nist.gov/tac/publications/index.html>. 78
- [Edmundson, 1968] Edmundson, H. (1968). New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 1. 4, 75
- [Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *IJCV*. 208

- [Farhadi et al., 2010] Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: generating sentences from images. In *ECCV*. 206, 210
- [Fellbaum, 1998] Fellbaum, C., editor (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London. 82, 133
- [Feller, 1968] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley. 22, 42
- [Felzenszwalb et al., 2010] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE PAMI*, 32(9). 3, 205, 209, 213
- [Feng et al., 2004] Feng, S. L., Manmatha, R., and Lavrenko, V. (2004). Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the IEEE CVPR Conference*, pages 1002–1009. 188
- [Feng and Lapata, 2010a] Feng, Y. and Lapata, M. (2010a). How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th ACL Conference*, pages 1239–1249. 109
- [Feng and Lapata, 2010b] Feng, Y. and Lapata, M. (2010b). Topic models for image annotation and text illustration. In *NAACL HLT*. 205
- [Fubini, 1958] Fubini, G. (1958). *Sugli integrali multipli*, volume 2. Opere scelte, Cremonese. 39
- [Ganesan et al., 2010] Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 10)*. 3
- [Ganesan et al., 2012] Ganesan, K., Zhai, C., and Viegas, E. (2012). Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the World Wide Web Conference*, pages 869–878. 3
- [Gatt and Reiter, 2009] Gatt, A. and Reiter, E. (2009). Simplenlg: a realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation, ENLG '09*, pages 90–93, Stroudsburg, PA, USA. Association for Computational Linguistics. 148, 176
- [Gelman et al., 2003] Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. 73
- [Genest et al., 2009] Genest, P., Lapalme, G., and Yousfi-Monod (2009). Hextac: the creation of a manual extractive run. In *Proceedings of the Second Text Analysis Conference (TAC 2009) – Guided Summarization Track*. 18, 139
- [Genest and Lapalme, 2011] Genest, P.-E. and Lapalme, G. (2011). Generated abstracts for tac 2011 – university of montreal. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011) – Guided Summarization Track*, Gaithersburg, Maryland, USA. 148, 176, 181

- [Gershman et al., 2012] Gershman, S., Hoffman, M. D., and Blei, D. M. (2012). Nonparametric variational inference. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. 233
- [Ghahramani and Jordan, 1997] Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden markov models. *Mach. Learn.*, 29(2-3):245–273. 65
- [Gillik and Favre, 2009] Gillik, D. and Favre, B. (2009). A scalable global model for summarization. In *NAACL/HLT 2009 Workshop on Integer Linear Programming for Natural Language Processing*. 78
- [Girolami and Kabán, 2003] Girolami, M. and Kabán, A. (2003). On an equivalence between plsi and lda. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03*, pages 433–434, New York, NY, USA. ACM. 10
- [Graber and Blei, 2009] Graber, J. B. and Blei, D. (2009). Syntactic topic models. In *Advances in Neural Information Processing Systems*, volume 21. 77, 79, 233
- [Griffiths and Steyvers, 2004] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235. 9, 31, 69, 70, 71, 72, 114
- [Grosz et al., 1995] Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. In *Computational Linguistics*, volume 21, pages 203–225. 75, 76, 77, 79, 82, 132, 144
- [Guillaumin et al., 2009] Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of ICCV Conference*, pages 309–316. 205
- [Gupta et al., 2012] Gupta, A., Verma, Y., and Jawahar, C. V. (2012). Choosing linguistics over vision to describe images. In *AAAI*. 204
- [Haghighi and Vanderwende, 2009] Haghighi, A. and Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09*, pages 362–370, Stroudsburg, PA, USA. Association for Computational Linguistics. 18, 141, 147
- [Hansen and Larsen, 1996] Hansen, L. K. and Larsen, J. (1996). Unsupervised learning and generalization. In *Proceedings of the IEEE International Conference on Neural Networks 1996, Washington DC*, pages 25–30. 35
- [Harville, 2008] Harville, D. (2008). *Matrix Algebra From a Statistician's Perspective*. Springer, corrected edition. 218
- [Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition. 28
- [Hernault et al., 2010] Hernault, H., Prendinger, H., duVerle, D., and Ishizuka, M. (2010). Hilda: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1:1–33. 146

- [Herstein, 1964] Herstein, I. (1964). *Topics in algebra*. Blaisdell book in the pure and applied sciences. Blaisdell Pub. Co. 55
- [Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM. 17, 62, 69
- [Hospedales et al., 2009] Hospedales, T. M., Gong, S., and Xiang, T. (2009). A markov clustering topic model for mining behaviour in video. In *Proceedings of the ICCV Conference*, pages 1165–1172. 188, 206
- [Hovy et al., 2005] Hovy, E., Lin, C.-Y., and Zhou, L. (2005). A be-based multi-document summarizer with query interpretation. In *Proceedings of the Document Understanding Conference*. 78, 93
- [J et al., 2005] J, J., Pingali, P., and Varma, V. (2005). A relevance-based language modeling approach to duc 2005. <http://duc.nist.gov/pubs.html#2005>. 78, 93
- [Jaakkola, 2000a] Jaakkola, T. S. (2000a). Tutorial on variational approximation methods. In *In Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press. 59
- [Jaakkola, 2000b] Jaakkola, T. S. (2000b). Tutorial on variational approximation methods. In *In Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press. 66
- [Jordan, 2004] Jordan, M. I. (2004). *Learning in graphical models*. MIT Press. 1
- [Jurafsky and Martin, 2000] Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 1 edition. neue Auflage kommt im Frhjahr 2008. 1
- [Kadanoff, 2009] Kadanoff, L. P. (2009). More is the same; phase transitions and mean field theories. Comments: 25 pages, 6 figures. 54
- [Khan et al., 2011] Khan, M. U. G., Zhang, L., and Gotoh, Y. (2011). Towards coherent natural language description of video streams. In *ICCV Workshops*. 206, 208, 209
- [Kipper, 2005] Kipper, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania. <http://repository.upenn.edu/dissertations/AAI3179808/>. 82
- [Kläser et al., 2008] Kläser, A., Marszalek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of BMVC Conference*. 185, 196, 206, 207, 212, 228, 229
- [Klein and Manning, 2003] Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st ACL Conference*, pages 423–430. 202
- [Kuettel et al., 2012] Kuettel, D., Guillaumin, M., and Ferrari, V. (2012). Segmentation propagation in imagenet. In *ECCV*. 205
- [Kulesza and Taskar, 2012] Kulesza, A. and Taskar, B. (2012). Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3). 233

- [Kulkarni et al., 2011] Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *CVPR, CVPR*. 206, 213
- [Kvam and Vidakovic, 2007] Kvam, P. H. and Vidakovic, B. (2007). *Nonparametric Statistics with Applications to Science and Engineering (Wiley Series in Probability and Statistics)*. Wiley-Interscience. 146, 165
- [Landauer et al., 1998] Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284. 10
- [Lang, 1995] Lang, K. (1995). Newsweeder: Learning to filter netnews. 35
- [Lavrenko et al., 2004] Lavrenko, V., Manmatha, R., and Jeon, J. (2004). A model for learning the semantics of pictures. In *Neural Information Processing Systems*. 188
- [Lee, 1997] Lee, J. H. (1997). Analysis of multiple evidence combination. In *20 th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 97)*, pages 267–276. 172
- [Li et al., 2010a] Li, L., Su, H., Xing, E. P., and Fei, L. F. (2010a). Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*. 3
- [Li et al., 2011] Li, L., Zhou, K., Xue, G.-R., Zha, H., and Yu, Y. (2011). Video summarization via transferrable structured learning. In *Proceedings of the WWW Conference*, pages 287–296. 185, 188
- [Li et al., 2010b] Li, L.-J., Su, H., Xing, E. P., and Fei-fei, L. (2010b). Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Proceedings of NIPS Conference*. 185, 187, 188, 190, 196, 205, 222
- [Li et al., 2005] Li, W., Li, W., Li, B., Chen, Q., and Wu, M. (2005). The hong kong polytechnic university at duc 2005. <http://duc.nist.gov/pubs.html#2005>. 78, 93
- [Li and McCallum, 2006] Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning, ICML '06*, pages 577–584, New York, NY, USA. ACM. 69, 104, 130
- [Liang and Klein, 2008] Liang, P. and Klein, D. (2008). Analyzing the errors of unsupervised learning. In *ACL*, pages 879–887. 35
- [Lin and Hovy, 2000] Lin, C.-Y. and Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics - Volume 1, COLING '00*, pages 495–501, Stroudsburg, PA, USA. Association for Computational Linguistics. 141, 148
- [Lin and Hovy, 2003] Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03*, pages 71–78, Stroudsburg, PA, USA. Association for Computational Linguistics. 4, 35, 93, 131, 159, 185, 213, 223

- [Lin and Bilmes, 2012] Lin, H. and Bilmes, J. (2012). Learning mixtures of submodular shells with application to document summarization. In *Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, USA. AUAI. 2, 233
- [Long et al., 2009] Long, C., Huang, M., Zhu, X., and Li, M. (2009). Multi-document summarization by information distance. *IEEE International Conference on Data Mining*, 0:866–871. 78
- [Lu et al., 2010] Lu, C., Hu, X., Chen, X., Park, J. R., He, T., and Li, Z. (2010). The topic-perspective model for social tagging systems. In *Proceedings of the 16th ACM SIGKDD Conference*, pages 683–692. 113
- [Luhn, 1958] Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165. 1, 4, 75
- [Maaten and Hinton, 2008] Maaten, L. V. D. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *JMLR*, 9(Nov):2579–2605. xv, 197
- [Makadia et al., 2008] Makadia, A., Pavlovic, V., and Kumar, S. (2008). A new baseline for image annotation. In *Proceedings of the ECCV Conference*, pages 316–329. 3, 185, 187, 205
- [Mann and Thompson, 1988] Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281. 3, 79, 142, 145
- [Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. 35
- [Marcu, 1999] Marcu, D. (1999). Discourse trees are good indicators of importance in text. In *Advances in Automatic Text Summarization*, pages 123–136. The MIT Press. 18, 142, 149, 164
- [Marcu, 2000a] Marcu, D. (2000a). The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics*, 26:395–448. 145
- [Marcu, 2000b] Marcu, D. (2000b). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, Cambridge, MA ; London. 78
- [McAuliffe et al., 2006] McAuliffe, J. D., Blei, D. M., and Jordan, M. I. (2006). Nonparametric empirical bayes for the dirichlet process mixture model. *Statistics and Computing*, 16(1):5–14. 233
- [McDonald, 2007] McDonald, R. T. (2007). A study of global inference algorithms in multi-document summarization. In *ECIR*, volume 4425 of *Lecture Notes in Computer Science*, pages 557–564. Springer. 4, 75, 77, 78, 149
- [Mei et al., 2007a] Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007a). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, pages 171–180. 77
- [Mei et al., 2007b] Mei, Q., Shen, X., and Zhai, C. (2007b). Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '07*, pages 490–499, New York, NY, USA. ACM. 7

- [Mimno et al., 2009] Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *EMNLP*. 69
- [Minka, 2009] Minka, T. P. (2009). Estimating a Dirichlet distribution. Technical report, MIT. 154, 225
- [Nallapati and Cohen, 2008] Nallapati, R. and Cohen, W. (2008). Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In *International Conference for Weblogs and Social Media*. 77
- [Nasios and Bors, 2006] Nasios, N. and Bors, A. (2006). Variational learning for gaussian mixture models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(4):849–862. 191, 201
- [Natarajan et al., 2011] Natarajan, P. et al. (2011). BBN VISER. In *TRECVID MED*. 202
- [Neal, 2000] Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265. 71
- [Nenkova and Louis, 2008] Nenkova, A. and Louis, A. (2008). Can you summarize this? identifying correlates of input difficulty for generic multi-document summarization. In *Proceedings of Association for Computational Linguistics(ACL-08): HLT*. 94
- [Nenkova and Passonneau, 2004] Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Main Proceedings*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics. 94, 131, 159, 185
- [Nenkova et al., 2006a] Nenkova, A., Vanderwende, L., and McKeown, K. (2006a). A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *SIGIR*. 6, 143, 169
- [Nenkova et al., 2006b] Nenkova, A., Vanderwende, L., and McKeown, K. (2006b). A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 573–580, New York, NY, USA. ACM. 75
- [Newman et al., 2010] Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Stroudsburg, PA, USA. Association for Computational Linguistics. 112
- [Newman, 2005] Newman, M. E. J. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351. 166
- [Nishimoto et al., 2011] Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646. xi, 15
- [Nocedal and Wright, 2006] Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York, 2nd edition. 25, 52

- [Oliva and Torralba, 2001] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International J. Comput. Vision*, 42(3):145–175. 185, 196, 199
- [Oliva and Torralba, 2006] Oliva, A. and Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, 155(1):23–36. 197
- [Pan and Yang, 2010] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359. 233
- [Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL*. 213
- [Parisi, 1988] Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley. 54, 191
- [Pedersen et al., 2004] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::similarity - measuring the relatedness of concepts. In *AAAI*, pages 1024–1025. 112, 126
- [Pereira et al., 2011] Pereira, F., Detre, G., and Botvinick, M. (2011). Generating text from functional brain images. *Frontiers in Human Neuroscience*, 5:72. 14
- [Perera et al., 2012] Perera, A. et al. (2012). TRECVID 2012 GENIE: multimedia event detection and recounting. In *TRECVID Workshop*, Gaithersburg, Maryland, USA. 202
- [Press et al., 2007] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition. 73
- [Putthividhya et al., 2010] Putthividhya, D., Attias, H. T., and Nagarajan, S. S. (2010). Topic regression multi-modal latent dirichlet allocation for image annotation. In *Proceedings of CVPR Conference*, pages 3408–3415. 188, 205
- [Radev et al., 2004] Radev, D. R., Jing, H., Styś, M. g., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919–938. 141, 148, 159, 170
- [Ramage et al., 2010] Ramage, D., Dumais, S. T., and Liebling, D. J. (2010). Characterizing microblogs with topic models. In *Proceeding of the International Conference of Weblogs and Social Media*. 3
- [Ramage et al., 2009a] Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009a). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 EMNLP Conference*, pages 248–256, Singapore. 113
- [Ramage et al., 2009b] Ramage, D., Heymann, P., Manning, C. D., and Garcia-Molina, H. (2009b). Clustering the tagged web. In *Proceedings of the WSDM Conference*, pages 54–63. 13, 14, 110, 111, 113, 123, 187, 190
- [Rashtchian et al., 2010] Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon’s mechanical turk. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. 213

- [Robert and Casella, 2005] Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 69
- [Rohrbach et al., 2012] Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M., and Schiele, B. (2012). Script data for attribute-based recognition of composite activities. In *ECCV*. 212
- [Rosen-Zvi et al., 2004] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence, UAI '04*, pages 487–494, Arlington, Virginia, United States. AUAI Press. 69
- [Sadanand and Corso, 2012] Sadanand, S. and Corso, J. J. (2012). Action bank: A high-level representation of activity in video. In *IEEE Computer Vision and Pattern Recognition Conference*. 206
- [Sadeghi and Farhadi, 2011] Sadeghi, M. A. and Farhadi, A. (2011). Recognition using visual phrases. In *CVPR*. 209
- [Schilder and Kondadadi, 2008] Schilder, F. and Kondadadi, R. (2008). Fastsum: fast and accurate query-based multi-document summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, HLT-Short '08*, pages 205–208, Stroudsburg, PA, USA. Association for Computational Linguistics. 140, 148
- [Shachter, 1998] Shachter, R. D. (1998). Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams. In *In Uncertainty in Artificial Intelligence*, pages 480–487. Morgan Kaufmann. xii, 11, 44, 56, 65, 72, 116
- [Si and Sun, 2009] Si, X. and Sun, M. (2009). Tag-LDA for Scalable Real-time Tag Recommendation. *Journal of Computational Information Systems*. 110, 113, 123
- [Sibun, 1993] Sibun, P. (1993). Domain structure, rhetorical structure, and text structure. ---- 79
- [Sontag and Jaakkola, 2007] Sontag, D. and Jaakkola, T. (2007). New outer bounds on the marginal polytope. In *NIPS*. 50
- [Sontag and Roy, 2011] Sontag, D. and Roy, D. (2011). Complexity of inference in latent dirichlet allocation. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 1008–1016. 63
- [Soricut and Marcu, 2003] Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03*, pages 149–156, Stroudsburg, PA, USA. Association for Computational Linguistics. 145, 146, 164, 170
- [Spiliopoulou and Storkey, 2012] Spiliopoulou, A. and Storkey, A. J. (2012). A topic model for melodic sequences. In *ICML*. 69
- [Srihari et al., 2007] Srihari, R., Xu, L., and Saxena, T. (2007). Use of ranked cross document evidence trails for hypothesis generation. In *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining*, pages 677–686, San Jose, CA. 78, 93, 94, 126

- [Srihari, 1991] Srihari, R. K. (1991). Piction: A system that uses captions to label human faces in newspaper photographs. In *AAAI*. 185
- [Stuart et al., 1999] Stuart, A., Ord, J., and Kendall, M. (1999). *Kendall's advanced theory of statistics: Classical inference and the linear model*. Number v. 2 in Kendall's Advanced Theory of Statistics: Classical Inference and the Linear Model. Charles Griffin & Company limited. 32
- [Tang et al., 2009] Tang, J., Yao, L., and Chen, D. (2009). Multi-topic based query-oriented summarization. In *SDM*, pages 1147–1158. SIAM. 147
- [Titov and McDonald, 2008] Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 111–120, New York, NY, USA. ACM. 3
- [Torralba and Efros, 2011] Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR*. 204
- [Torresani et al., 2010] Torresani, L., Szummer, M., and Fitzgibbon, A. (2010). Efficient object category recognition using classemes. In *Proc of ECCV Conference*, pages 776–789. 188, 196
- [van de Sande et al., 2010] van de Sande, K. E. A., Gevers, T., and Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE PAMI*, 32(9). 207, 212
- [Vanderwende et al., 2007] Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. (2007). Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manage.*, 43(6):1606–1618. 143
- [Varma et al., 2010] Varma, V., Bysani, P., Reddy, K., Reddy, V., Kovelamudi, S., Vaddepally, S. R., Nanduri, R., N, K. K., Gsk, S., and Pingali, P. (2010). Iiit hyderabad in guided summarization and knowledge base population. In *Proceedings of the Fourth Text Analysis Conference (TAC 2010) – Guided Summarization Track*, Gaithersburg, Maryland, USA. 140
- [Varma et al., 2011] Varma, V., Kovelamudi, S., Sood, A., Gupta, J., Jain, H., Priyatam, N., Mogadala, A., and Reddy, S. (2011). Iiit hyderabad in summarization and knowledge base population at tac 2011. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011) – Guided Summarization Track*, Gaithersburg, Maryland, USA. 140
- [Vondrick et al., 2010a] Vondrick, C., Ramanan, D., and Patterson, D. (2010a). Efficiently Scaling Up Video Annotation with Crowdsourced Marketplaces. In *Proceedings of ECCV Conference*. 185
- [Vondrick et al., 2010b] Vondrick, C., Ramanan, D., and Patterson, D. (2010b). Efficiently scaling up video annotation with crowdsourced marketplaces. In *Proc. of the European Conference on Computer Vision*. 209
- [W. and M., 2009] W., Y. and M., G. (2009). Human action recognition by semi-latent topic models. *IEEE Transactions on PAMI*, 31(10):1762–1774. 188, 206
- [Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA. 22, 43, 45, 49, 50, 51, 52, 55, 57, 59, 152, 191, 198, 206

- [Wallach, 2006] Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning, ICML '06*, pages 977–984, New York, NY, USA. ACM. 148
- [Wallach, 2008] Wallach, H. M. (2008). *Structured Topic Models for Language*. PhD thesis, University of Cambridge. 97
- [Wallach et al., 2009] Wallach, H. M., Mimno, D., and McCallum, A. (2009). Rethinking LDA: Why priors matter. In *NIPS*. 14, 113, 149, 154, 155, 206
- [Wang and Blei, 2009] Wang, C. and Blei, D. M. (2009). Variational inference for the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 1990–1998. 233
- [Wang et al., 2008] Wang, C., Blei, D. M., and Heckerman, D. (2008). Continuous time dynamic topic models. In *Proceedings of Uncertainty in Artificial Intelligence*. 2
- [Wang et al., 2009] Wang, C., Blei, D. M., and Li, F.-F. (2009). Simultaneous image classification and annotation. In *CVPR*. 205
- [Wanke et al., 2010] Wanke, J., Ulges, A., Lampert, C. H., and Breuel, T. M. (2010). Topic models for semantics-preserving video compression. In *Proceedings of the international MIR Conference*, pages 275–284. 188, 206
- [Wei et al., 2010] Wei, F., Li, W., Lu, Q., and He, Y. (2010). A document-sensitive graph model for multi-document summarization. *Knowledge and Information Systems*, 22(2):245–259. 77
- [Welch et al., 2010] Welch, M. J., Cho, J., and Chang, W. (2010). Generating advertising keywords from video content. In *CIKM*. 183
- [Xu et al., 2013] Xu, X., Shimada, A., and Taniguchi, R.-i. (2013). Correlated topic model for image annotation. In *Frontiers of Computer Vision, (FCV), 2013 19th Korea-Japan Joint Workshop on*, pages 201–208. 188
- [Yakhnenko and Honavar, 2009] Yakhnenko, O. and Honavar, V. (2009). Multi-modal hierarchical dirichlet process model for predicting image annotation and image-object label correspondence. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, pages 281–294. SIAM. 233
- [Yang et al., 2011] Yang, Y., Teo, C. L., Daumé III, H., and Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proceedings of EMNLP Conference*. 187, 204, 206, 213
- [Yano et al., 2009] Yano, T., Cohen, W. W., and Smith, N. A. (2009). Predicting response to political blog posts with topic models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 477–485, Morristown, NJ, USA. Association for Computational Linguistics. 77
- [Ye et al., 2005] Ye, S., Chua, T.-S., Kan, M.-Y., and Qiu, L. (2005). Nus at duc 2005: Understanding documents via concept links. <http://duc.nist.gov/pubs.html#2005>. 78, 93

- [Yih et al., 2007] Yih, W.-T., Goodman, J., Vanderwende, L., and Suzuki, H. (2007). Multi-document summarization by maximizing informative content-words. *Proceedings of IJCAI*. 108, 148
- [Zeng et al., 2011] Zeng, J., Cheung, W. K., and Liu, J. (2011). Learning topic models by belief propagation. *CoRR*, abs/1109.3437. 45
- [Zhai and Boyd-Graber, 2013] Zhai, K. and Boyd-Graber, J. (2013). Online topic models with infinite vocabulary. In *International Conference on Machine Learning*. 233
- [Zhai et al., 2012] Zhai, K., Boyd-Graber, J., Asadi, N., and Alkhouja, M. (2012). Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *ACM International Conference on World Wide Web*. 73
- [Zhang et al., 1995] Zhang, H. J., Low, C. Y., Smoliar, S. W., and Wu, J. H. (1995). Video parsing, retrieval and browsing: an integrated and content-based solution. In *Proceedings of the third ACM Multimedia Conference*, pages 15–24. 188, 196
- [Zhu et al., 2006] Zhu, X., Blei, D., and Lafferty, J. (2006). Taglda: Bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin, Madison. 13, 17, 73, 84, 110, 111, 113, 119, 122, 123, 141, 149, 152, 168

Index

- Algorithms
 - VBEM inner loop: Tag²LDA family of models (E step), 119
 - VBEM outer loop: Tag²LDA family of models, 119
 - VBEM outer loop: Tag²LDA family of models (M step), 119
- Annotation
 - Document Level, 108
 - Image, 185
 - Video
 - Difficulties, 185
 - Word Level, 108
- Attentional state, 75
- Bayes estimators, 36
- Bayes risk, 37
- Bayesian vs. Frequentist, 38
- Centering Theory, 79
 - Inference load, 76
- Centers
 - Backward-looking, 80
 - Forward-looking, 80
 - of utterances, 75
 - utterance, 81
- Conjugate priors, 31
- Corr-METag²LDA, 117
 - ELBO, 117
 - ELBO expression, 133
 - Latent variable inference, 117
 - Maximum likelihood parameter estimation, 119
 - Maximum likelihood parameter estimation derivation, 136
 - Parameter regularization, 119
 - Variational inference derivations, 134
- Corr-MMGLDA, 14, 190
 - ELBO, 194
 - Latent variable inference, 194
 - Maximum likelihood parameter estimation, 195
 - Normal-Wishart prior updates, 195
 - Sensitivity to independent data scaling, 197
 - Sufficient statistics, 195
 - Word prediction formula, 196
- Corr-MMLDA, 110
- Correspondence - Multinomial Multinomial LDA, *see also* Corr-MMLDA
- Correspondence-Multinomial Multinomial Gaussian LDA, *see* Corr-MMGLDA
- Dataset
 - TAC 2010, 142
 - TAC 2011, 142
- Datasets
 - Amazon reviews, 109
 - DUC 2005, 89, 109
 - TAC 2008, 89
 - TAC 2009, 89
 - TRECVID MED 11, 186
 - Wikipedia, 109
- De Finetti's theorem, 42
- DL Tags, 108
- Document Level tags, *see* DL Tags
- ELBO, 43
 - Annotation anomalies from tag-topic models, 156
 - Effect of the feature sets in DL perspective, 129
 - Sentence log likelihoods, 158
- EM, 39
 - Constrained optimization: The approximate case, 45
 - Lower bound improvement proof, 40
 - Missing data example, 40
 - Unconstrained optimization: The exact case, 44
- Evidence Lower BOUND, *see* ELBO
- Expectation Maximization, *see also* EM
- Exponential family

- minimal, 51
- Exponential Family Distributions, 20
 - Cumulant function, 22
 - Derivatives of log partition function, 23
 - Log partition function, 22
 - Reparameterization, 22
 - Sufficient statistics, 22
- Gaussian Mixture Model
 - visualization, 26
- Grammatical and semantic roles, 76
- GSR, *see* Grammatical and semantic roles
- GSR transition, 76
- GSRs, *see* Grammatical and semantic roles
- GSRt, *see* GSR transition, 82
- Independent and identically distributed, i.i.d, 27
- Latent Dirichlet Allocation, *see* LDA, 11, 61
- Latent Semantic Analysis, 10
- LDA
 - Fixed point iteration for updating variational document topic proportion parameter, 67
 - Amount of training data needed to learn the model parameters, 63
 - Computing ML estimates is NP-Hard, 63
 - Example topics, 6
 - Exponential family representation, 64
 - Gibbs sampling, 72
 - Incorporating word level annotations, 13
 - Mean field approximation, 65
 - Sample topics from patent and legal documents about rockets and propulsion, 7
 - Supervised, 110
 - Topic-Word matrix
 - Anchor words, 63
 - Separability, 63
 - Why does it work?, 62
- Learning To Summarize model, 87
- LeToS, *see also* Learning To Summarize model, 87
 - Latent variable inference and parameter estimation, 89
 - Lower bound expression, 102
 - Maximum likelihood parameter estimation, 105
 - shortcomings, 87, 104
 - Variational inference, 103
- Likelihood
 - Lower bound, 43
 - Maximum likelihood, 27
 - Asymptotics, 32
 - Maximum entropy duality, 28
 - Maximum likelihood Estimators
 - Asymptotic efficiency, 33
 - Consistency, 32
 - Principle, 30
 - Literature review
 - multi-document summarization, 77
 - Multidocument summarization, 147
 - Topic models incorporating domain knowledge, 113
 - Video to text summarization, 188
 - Loss function optimality, 36
 - Relation to Type 1 and Type 2 errors, 38
 - LSA, *see* Latent Semantic Analysis
- Mean field
 - Convexity, 50
 - Factorization, 55
 - Hidden variable and parameter factorization, 55
 - METag²LDA model, 116
 - Non-convexity, 59, 198
 - Optimization, 59
 - Cartoon illustrations, 60
 - Procedure, 57
 - Tractability, 55
- Mean parameters, 49
 - Forward mapping, 51
 - Inference, 50
- METag²LDA, 13
 - ELBO, 117
 - Latent variable inference, 117
 - Maximum likelihood parameter estimation, 118
 - Parameter regularization, 119
- MMGLDA, 190
 - ELBO, 192
 - Latent variable inference, 193
 - Maximum likelihood parameter estimation, 195
 - Normal-Wishart prior updates, 195
 - Sufficient statistics, 195
 - Word prediction formula, 196
- MMLDA, 110

- Multimedia Topic Models, *see* MMGLDA, Cor-MMGLDA
 - Important derivations, 216
 - Prediction ELBO on heldout test data which is dissimilar to training set, 199
 - Prediction ELBO on heldout test data which is similar to training set, 198
 - Test ELBO on heldout test data which is dissimilar to training set, 197
 - Test ELBO on heldout test data which is similar to training set, 197
- Multinomial Multinomial Gaussian LDA, *see* MMGLDA
- Multinomial Multinomial LDA, *see also* MMLDA
- Pronoun resolution, 80
- Rhetorical Structure trees, *see* RS trees, 145
- Sampling
 - Dirichlet, 73
 - Gibbs, 69
 - Imputation Posterior (IP) algorithm, 70
 - Multinomial, 73
- Sparse Coherence Flow, 82, 83
- Sufficient statistics, 29
 - Example, 30
 - Factorization theorem, 29
- Summarization
 - Text multidocument summarization, *see also* Text multidocument summarization, 139
 - Video to text
 - Analogy to text multidocument summarization, 185
 - Importance in search, 185
 - Low level features, 196
 - Low level features: Action descriptors—HOG3D, 196
 - Low level features: Color histogram, 196
 - Low level features: Image summary descriptors—GIST, 197
 - Low level features: Object bank, 196
 - Natural language generation, 202
 - Recall oriented evaluation, 187
 - ROUGE evaluation, 201
 - Vocabulary intrusion, 185
- Summarization problem
 - hardness, 4
 - significance, 4
- Summary
 - bullet list, 176
 - Chapter 3, 97
 - Chapter 4, 132
 - Chapter 5, 181
 - Chapter 6, 215
 - Chapters 1 and 2, 17
- Supervised Learning
 - Amount of training data, 33
 - VC Dimension, 34
- t-SNA, *see* t-statistic based Stochastic Neighbor Embedding
- t-statistic based Stochastic Neighbor Embedding, 197
- Tag²LDA family of models, 114
 - ELBO
 - Effect of the feature sets in DL perspective, 129
- TagLDA, 13, 110
 - ELBO
 - Effect of the feature sets in DL perspective, 129
 - Parameter regularization, 119
- Text multidocument summarization, 139
 - Collaborative dual perspectives in text documents, 143
 - Comparing systems, 159
 - Data preparation
 - TAC newswire collections, 150
 - Event classification, 163
 - Global and local models, 143
 - Guided summarization, 139
 - Local document set specific models, 162
 - Local models
 - Importance of nouns and verbs, 140
 - Recovering query from nouns and verbs, 144
 - Recall oriented evaluation, 170
 - Rhetorical structure trees, *see* RS trees
 - ROUGE evaluation
 - TAC 2010A/2011A: baselines and human summaries, 170
 - TAC 2010A/2011A: Topic model baselines, 172
 - TAC 2010A: Proposed models, 172
 - TAC 2011A: Proposed models, 175

- RS tree representation of sentences, 145
- RS trees, 145
 - Sentence compression, 164
- RST
 - Elementary discourse units, 146
- Tag topic models, *see also* Chapter 4, 149
 - ELBO, 158
 - Importance of asymmetric Dirichlet: α , 155
- Topic models
 - Entity relation discovery, 128
 - Gaussian-Wishart priors over real valued observations, 191
 - Summarization power, 130
- Topic proportion prior α
 - Asymmetric
 - Optimization, 225
 - Asymmetric Dirichlet, 154
- Topics
 - Example, 7
 - Faceted, 10
 - Wikipedia, 111
 - Short documents, 155
 - Text topics from Corr-MMGLDA, 197
 - Translating related words to video frames, 199
- Unsupervised Learning
 - Amount of training data, 35
 - Meta models, 35
- UTM, *see also* Utterance Topic Model, 84
 - Effect of coarse coherence encoding, 91
 - Latent variable inference, 86
 - Lower bound expression, 98
 - Maximum likelihood parameter estimation, 100
 - Parameter estimation, 87
 - Variational inference, 99
- Utterance Topic Model, 84
- Variational Bayes
 - KL divergence, 46
- Video to text
 - Cleaning an appliance example, 16
 - Event classification, 202
 - Metal crafts project example, 16
 - Skateboarding example, 183
 - Woodworking project example, 183
- Wikipedia
 - Generative models for a typical page with embedded multimedia, 9
 - Higher values of the document level α prior for correspondence LDA models, 130
 - Measuring relevancy of caption words to manually assigned document labels, 126
 - Special export URL, 121
 - Typical page with embedded images, 8
- WL annotation, 108
- Word level annotation, *see* WL annotation