ABSTRACT

Title of dissertation: CONNECTING DOCUMENTS, WORDS,
AND LANGUAGES USING TOPIC MODELS

Weiwei Yang, Doctor of Philosophy, 2019

Dissertation directed by: Professor Jordan Boyd-Graber
Department of Computer Science
College of Information Studies
Language Science Center
Institute for Advanced Computer Studies

Professor Philip Resnik
Department of Linguistics
Institute for Advanced Computer Studies

Topic models discover latent topics in documents and summarize documents at a high level. To improve topic models' topic quality and extrinsic performance, external knowledge is often incorporated as part of the generative story. One form of external knowledge is weighted text links that indicate similarity or relatedness between the connected objects. This dissertation 1) uncovers the latent structures in *observed* weighted links and integrates them into topic modeling, and 2) learns *latent* weighted links from other external knowledge to improve topic modeling.

We consider incorporating links at three different levels: documents, words, and topics. We first look at binary document links, e.g., citation links of papers. Document links indicate topic similarity of the connected documents. Past methods model the document links separately, ignoring the entire link density. We instead

uncover latent document blocks in which documents are densely connected and tend to talk about similar topics. We introduce LBH-RTM, a relational topic model with lexical weights, block priors, and hinge loss. It extracts informative topic priors from the document blocks for documents' topic generation. It predicts unseen document links with block and lexical features and hinge loss, in addition to topical features. It outperforms past methods in link prediction and gives more coherent topics.

Like document links, words are also linked, but usually with real-valued weights. Word links are known as word associations and indicate the semantic relatedness of the connected words. They provide more information about word relationships in addition to the co-occurrence patterns in the training corpora. To extract and incorporate the knowledge in word associations, we introduce methods to find the most salient word pairs. The methods organize the words in a tree structure, which serves as a prior (i.e., tree prior) for tree LDA. The methods are straightforward but effective, yielding more coherent topics than vanilla LDA, and slightly improving the extrinsic classification performance.

Weighted topic links are different. Topics are latent, so it is difficult to obtain ground-truth topic links, but *learned* weighted topic links could bridge the topics across languages. We introduce a multilingual topic model (MTM) that assumes each language has its own topic distributions over the words *only* in that language and learns weighted topic links based on word translations and words' topic distributions. It does not force the topic spaces of different languages to be aligned and is more robust than previous MTMs that do. It outperforms past MTMs in classification while still giving coherent topics on less comparable and smaller corpora.

Connecting Documents, Words, and Languages Using Topic Models

by

Weiwei Yang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Philip Resnik, Co-Chair/Co-Advisor
Professor Jordan Boyd-Graber, Co-Chair/Co-Advisor
Professor Douglas Oard, Dean's Representative
Professor Marine Carpuat, Member
Professor Max Leiserson, Member
Professor Mark Dredze, External Member

Acknowledgments

My five-year Ph.D. journey is about to end. It is an unforgettable experience that I will cherish forever. During this journey, I feel extremely fortunate to meet so many wonderful people, without whom this dissertation would never be possible. I hope the following words can express my gratefulness.

First and foremost, I would like to thank my amazing co-advisors, Philip Resnik and Jordan Boyd-Graber, for their support, advice, and guidance throughout my Ph.D. study. Whenever I encounter a problem, they are always there ready to help. During the past five years, they gave me numerous advice in research (from the big picture to technical details), career (from internship to full-time job hunting), and daily life. I will never forget the time we together work through research problems, write papers, and share the pleasure of paper acceptance. I will also remember the valuable advice they give me when I face crucial choices in my life like internships and full-time jobs.

I would also like to thank my dissertation committee members, Douglas Oard, Marine Carpuat, Max Leiserson, and Mark Dredze, for taking their invaluable time reviewing my dissertation. Their insightful and constructive comments and questions greatly help me better present my Ph.D. research.

I am very fortunate to have great collaborators, Shudong Hao, Yoshinari Fujinuma, Mans Hulden, Ling Liu, Mozhi Zhang, Michelle Yuan, and Ting Hua, in the DARPA LORELEI project. We work closely and overcome many difficulties in research and project evaluations. I would particularly thank Shudong, Mozhi, and

Michelle for generously sharing the datasets they collected with me for my research. Many thanks also go to the Defense Advanced Research Projects Agency (DARPA) and Raytheon BBN for supporting my Ph.D research for four years, via the Low Resource Languages for Emergent Incidents (LORELEI) project (HR0011-15-C-0113). I would also like to thank the National Science Foundation (NSF) Award 1211153 (SoCS: Collaborative Research: Data Driven, Computational Models for Discovery and Analysis of Framing) and the UMD Computer Science Department for funding me during my first year.

The CLIP Lab is full of smart and nice people. I am very grateful to Viet-An Nguyen for guiding me through the equations at my early stage of Ph.D. along with Philip and Jordan. I will not forget the meals I had together with Chen Zhao during my last year of Ph.D., during which we encourage each other on research and life. I learned a lot from the discussions with Shi Feng, Fenfei Guo, He He, Hua He, Lingzi Hong, Xing Niu, and Jinfeng Rao. I also enjoyed interacting with Hadi Amiri, Amittai Axelrod, Joe Barrow, Ahmed Elgohary, Allyson Ettinger, Ning Gao, Mohit Iyyer, Khanh Nguyen, Thang Nguyen, Denis Peskov, Sudha Rao, Rashmi Sankepally, Amr Sharaf, Han-chin Shing, Alison Smith, Jo Shoemaker, Yogarshi Vyas, and Yulu Wang. I would also like to acknowledge the help and support from UMIACS staff. Joe Webster and Bahar Azami maintain the CLIP's computation resources. They did a fantastic job and provided a stable and efficient computing and storage infrastructure for us to run all kinds of experiments.

I spent three wonderful summers at three different companies. FiscalNote VP of Research and CLIP alumnus Vlad Eidelman offered me the opportunity to try the life at a start-up. I was given large freedom for choosing projects and received a lot of support from Vlad and other Data Science Team members, Jervis Pinto and Brian Grom. Another CLIP alumnus Ferhan Ture offered me an internship at Comcast Labs DC where for the first time, I applied deep learning in NLP. Another unforgettable summer was spent at Google where I worked with my manager Yundi Qian on ranking YouTube video comments. In addition to the great mentoring from Yundi, I appreciate the effort from Rob Abbott and Tom Schumm for reviewing more than ten thousand lines of my code. I also enjoyed working with Deepanjan Kundu, Fangxiaoyu Feng, Harrison Ruiz, Nolan Walsh, Nikhil Dandekar, Ramki Krishnan, and Eileen Long at YouTube Comments Team.

Many thanks go to my dear friends who make my life more interesting: Mingliang Chen, Ruofei Du, Mingfei Gao, Zebao Gao, Huijing Gong, Xingfeng He, Varun Kumar, Hao Li, Xue Li, Teng Long, Xiaoxu Meng, Xuetong Sun, Hong Wei, Zikai Wen, Qian Wu, Zheng Xu, Zheng Yan, Xi Yi, Ruichi Yu, Zehua Zeng, Zhenpeng Zhao, and Hao Zhou. It is nice to know every one of you.

Finally, I would like to express my deepest appreciation to my family. I thank my mum for endless support, no matter when my paper gets accepted or when my experiments are not positive. She made every effort to support my study, research, career, and life and sacrificed a lot. I can never pay back her unconditional love, but I promise I will do whatever I can to let her enjoy the life. I would also like to give my best wishes to my grandpa who is now at the age of 92 and gives me endless care. I sincerely wish my grandpa to be in good health and long life.

Table of Contents

Acknowledgements ii

List of Tables viii

List of Figures x

1 Introduction 1
- 1.1 Topic Modeling with Document Network . . . 7
- 1.2 Topic Modeling with Word Associations . . . 9
- 1.3 Topic Model for Learning Weighted Topic Links . . . 10
- 1.4 Additional Contributions . . . 12

2 Background 14
- 2.1 LDA Introduction . . . 15
  - 2.1.1 Posterior Inference . . . 17
- 2.2 Topic Model Evaluation . . . 20
- 2.3 Topic Model Extensions . . . 23
  - 2.3.1 Relaxing Current Assumptions . . . 24
  - 2.3.2 Incorporating External Knowledge . . . 27
    - 2.3.2.1 Upstream Models . . . 28
    - 2.3.2.2 Downstream Models . . . 30

3 Topic Modeling with Document Networks 36
- 3.1 Dealing with Links . . . 38
  - 3.1.1 Strongly Connected Components . . . 39
  - 3.1.2 Weighted Stochastic Block Model . . . 40
  - 3.1.3 Relational Topic Model . . . 42
- 3.2 Discriminative Topic Model with Block Prior and Features . . . 44
  - 3.2.1 LDA with Block Priors (BP-LDA) . . . 45
  - 3.2.2 More Features for Link Generation in RTM . . . 46
  - 3.2.3 Aggregated Model . . . 49
- 3.3 Posterior Inference . . . 52
  - 3.3.1 Sampling Block Assignments . . . 53

3.3.2 Sampling Topic Assignments 54
3.3.3 Parameter Optimization 56
3.4 Experimental Results 57
3.4.1 Link Prediction Results 58
3.4.2 Illustrative Example 61
3.4.3 Topic Quality Results 63
3.4.4 Block Analysis 66
3.5 Summary 67

4 Topic Modeling with Word Associations 71
4.1 Tree LDA: LDA with Tree Priors 73
4.1.1 Posterior Inference 76
4.2 Tree Prior Construction from Word Association Scores 77
4.2.1 Two-Level (2LV) 78
4.2.2 Hierarchical Agglomerative Clustering (HAC) 79
4.2.3 HAC with Leaf Duplication (HAC-LD) 80
4.3 Experimental Results 81
4.3.1 Perplexity 82
4.3.2 Topic Coherence 83
4.3.3 Extrinsic Classification 87
4.3.4 Learned Trees 89
4.4 Summary 90

5 Topic Model for Learning Weighted Topic Links 92
5.1 Multilingual Topic Model for Connecting Cross-Lingual Topics 96
5.2 Posterior Inference 101
5.2.1 E-step: Topic Assignment Sampling 101
5.2.2 M-step: Parameter Optimization 103
5.3 Experimental Results 104
5.3.1 Classification with Topic Posteriors 105
5.3.1.1 Classification Results 108
5.3.2 Learned Topics 108
5.3.3 Learned Topic Links 110
5.3.4 Topic Coherence on Less Comparable Corpora 112
5.3.5 Topic Coherence with Various Target Language Corpora Sizes 116
5.4 Summary 119

6 Conclusion and Future Work 121
6.1 Summary of Contributions 122
6.2 Directions for Future Work 124
6.2.1 Primary Limitations and Solutions 124
6.2.2 Other Future Directions 127

A Derivation of the Posterior Inference for LBH-RTM 130

A.1 Sampling Block Assignments 130

A.1.1 Undirected Links 130

A.1.2 Directed Links 132

A.2 Sampling Topic Assignments 134

A.2.1 Sigmoid Loss 135

A.2.2 Hinge Loss 136

A.3 Optimizing Parameters 138

A.3.1 Sigmoid Loss 138

A.3.2 Hinge Loss 139

A.4 Sampling Process 141

B Derivation of the Posterior Inference for the Multilingual Topic Model 142

Bibliography 145

List of Tables

2.1 Five example topics obtained from the TED talk corpus using LDA. Each topic is represented by the top fifteen words with the highest probabilities in that topic. Topic categories are obtained manually. . . 16

2.2 It is easy to find out an "intruder" in a coherent topic, such as "apple" in Topic 1, but hard in an incoherent one as Topic 2 (Chang et al., 2009). 22

3.1 The predictive link rank (PLR) of a document d is the average rank of actually linked documents with d. 59

3.2 Predictive link rank results. The performance improves over RTM when we incrementally add WSBM prior (BS-RTM), lexical weights (LBS-RTM), and hinge loss (LBH-RTM). 60

3.3 PLR of the citation link between example documents K (Koplon and Sontag, 1997) and A (Albertini and Sontag, 1992) (described in Section 3.4.2) 63

3.4 Topic coherence of models on Cora and WebKB, evaluated by Fisher's exact test (Upton, 1992, FET) and log-likelihood ratio (Moore, 2004; Dunning, 1993, LLR). WSBM priors and hinge loss benefit the topic coherence, while lexical weights hurt a little bit. 64

3.5 Three topics' top ten words given by various models. Words with general meanings are in *red and italic*. Words with specific meanings are in **blue and bold**. Generally, LBH-RTM assigns higher weights to specific words and lower weights to general words. 65

3.6 Statistics of Blocks 1 (Learning Theory) and 2 (Bayesian Nets) which are of different topics but linked by two edges. SCC merges the two blocks, while WSBM is robust and identifies two. 66

4.1 Corpora Statistics. 82

4.2 The average perplexity results on the test sets by various models. Tree names indicate the word association score and tree prior construction algorithm. LDA gives the lowest perplexity, because tLDA models have constraint from the tree priors and sacrifice the perplexity. . . . 83

4.3 We sort topics into thirds by Kullback-Leibler divergence (Kullback and Leibler, 1951, KLD): low, medium, and high divergence between vanilla LDA and tLDA. Unique coherent words are in **blue and bold**. Unique incoherent words are in *red and italic*. tLDA brings in more topic-relevant words. 86
4.4 Accuracies of topical classification on 20NewsGroups and sentiment analysis on Amazon reviews. Although not significantly improving the performance, tLDA topics at least do not hurt. 88

5.1 Statistics of the bilingual corpora used in classification experiments with inferred topic posteriors. For Wikipedia, the task is to classify each document into one of six categories. For LORELEI, the goal is to distinguish the need of *evacuation* from other need types. 105
5.2 Our MTM outperforms all the baseline models in both intra-lingual (IN) and cross-lingual (CR) evaluations in F1 scores. Connecting the top linked topics is a better way for topic transformation. 107
5.3 The topics of Moviesgiven by models. For each Chinese topic given by our MTM, the top three English topics and their link weights are also given, while the link weights to all English topics sum up to one. 109
5.4 Topics are linked because they have overlap in topical words. Although explicit word translations can help identify related topics, our MTM can also infer the topic relations beyond word translations, e.g., ZH-14 and EN-16 which have no overlap in words. 110
5.5 Statistics of the corpora for topic coherence evaluation. 30% of the documents in the partially comparable (PACO) corpora have direct translations in the other language, while no documents in the incomparable (INCO) corpora have direct translations. 113

List of Figures

1.1 An example of the three-level hierarchy in a topic model: high-level documents, low-level words, and mid-level latent topics. A document is a multinomial distribution over latent topics. A topic is a multinomial distribution over words. Topic names are summarized manually. 2

1.2 A positive Amazon review for an SD card with five stars. 4

1.3 A negative Amazon review for an SD card with one star. 4

1.4 WSBM identifies blocks in a network as denoted by colors and dashed boxes. Each block is a subset of nodes (denoted by circles) that are densely connected with each other but sparsely connected with the nodes in other blocks. We integrate WSBM into topic models and extract the blocks' topic patterns for better modeling documents' topics and predicting document links. 9

2.1 The Graphical Model of LDA. 15

2.2 A portion of the hierarchy learned by hierarchical LDA on abstracts of the Journal of ACM (Blei et al., 2007). Coarse-grained words are closer to the root, while fine-grained words are at leaves. 24

2.3 The first two principal components of word embeddings and topic/concept vectors in Gaussian LDA (Das et al., 2015, GLDA)/latent concept topic model (Hu and Tsujii, 2016, LCTM). 26

2.4 An example Markov random topic field in which each document propagates its topic distributions to other ones via links (Daumé III, 2009). 27

2.5 An upstream topic model. Topics are conditioned on external knowledge. 28

2.6 An upstream topic model in computer vision (Fei-Fei and Perona, 2005). It uses the image categories as the external knowledge and generates the patches' themes conditioned on them. Some notations are adapted for consistency. 30

2.7 A general graphical model of downstream topic models. Note that the external knowledge encoded in the posterior regularizer Ψ conditions on topic assignments $\boldsymbol{z}$ and/or tokens $\boldsymbol{w}$. 31

3.1 An example of strongly connected components. Every pair of nodes in the same component can reach each other via the nodes *only* in that component. The figure is adapted from `https://commons.wikimedia.org/w/index.php?curid=647584` . . . 39

3.2 The graphical model of weighted stochastic block model (Aicher et al., 2014, WSBM). . . . 41

3.3 SCC (Sharir, 1981) can be distracted by spurious links connecting two groups, while WSBM (Aicher et al., 2014) maintains the distinction. . . . 42

3.4 A two-document segment of relational topic model (Chang and Blei, 2010, RTM). . . . 43

3.5 Graphical Model of BP-LDA. . . . 45

3.6 A two-document segment of RTM with features denoted by grayscale. The document link $B_{d,d'}$ is observed and should be in gray, but we keep it in white background to avoid confusion. . . . 47

3.7 The graphical model of LBH-RTM for two documents, in which a weighted stochastic block model is integrated (γ, $\boldsymbol{\mu}$, $\mathbf{y}$, a, b, $\boldsymbol{\Omega}$, and $\mathbf{A}$) to identify latent document blocks. Each document's topic distribution has an informative prior $\boldsymbol{\pi}$ extracted from the block topic distributions. The model predicts links between documents ($\mathbf{B}$) based on topics ($\mathbf{z}$), words ($\mathbf{w}$), and inter-block link rates ($\boldsymbol{\Omega}$), using a max-margin objective. . . . 49

3.8 Titles and abstracts of Neural Network Papers K (Koplon and Sontag, 1997, upper) and A (Albertini and Sontag, 1992, lower). Paper K cites Paper A. . . . 62

3.9 Topic proportions given by various models on our two illustrative documents (K and A, described in described in Section 3.4.2). As the model gets more sophisticated, the Neural Network topic proportion gets higher and finally dominates the two documents when LBH-RTM is applied. Though LCH-RTM gives the highest proportion to the Neural Network topic, it splits the Neural Network topic into two and does not assign the proportions to the same one. . . . 69

3.10 SCC fails to identify two blocks which are different in topic distributions, because of the two inter-block links. WSBM is robust enough to identify the two. . . . 70

4.1 An example of a tree prior (the tree structure) and gold posterior edge and word probabilities learned by tLDA. Numbers beside the edges denote the probability of moving from the parent node to the child node. A word's probability (i.e., the number below the word) is the product of probabilities moving from the root to the leaf, e.g., $\Pr(\text{orbit}) = 0.61 \times 0.96 \times 0.57 = 0.34$. . . . 74

4.2 A part of a tree prior constructed from synonyms in WordNet. Adapted from Boyd-Graber et al. (2007). . . . 76

4.3 A part of a tree prior constructed from word translations. Adapted from Hu et al. (2014). 77
4.4 A two-level tree example with $N = 2$. The words in the internal nodes (i.e., "sport" and "match" *without* boxes) denote *concepts* and have no effect in tLDA. They are here only for exposition. 78
4.5 An example of building tree priors based on hierarchical agglomerative clustering (Lukasová, 1979, HAC). The construction starts from leaf nodes only, i.e., initial state. Then it repeatedly merges the clusters with the highest association score, as marked by the numbers, until there is only one left. 79
4.6 An example of constructing HAC with leaf duplication (HAC-LD) tree prior for the words "lake", "river", "spring", and "summer" ("Initial State"), whose paired words are shaded in gray and marked with ①. HAC-LD alleviates the problem in HAC that a word with multiple senses can only be assigned to a single cluster close to one of its senses, e.g., the word "spring" which can be either a season or a body of water. 80
4.7 Average PMI scores of the top 10 words in topics given by LDA and tLDA on 20NewsGroups (left) and Amazon reviews (right). Most tLDA topics are more coherent than LDA topics, while the two-level tree priors created on word2vec improve the most. The PMI scores of LCTM are too low to be included: 8.86 ± 0.66 on 20NewsGroups and 6.34 ± 1.21 on Amazon reviews. 85
4.8 Sub-trees for the word "pounds" in two topics, from the 20NewsGroups corpus using a two-level tree prior built on word2vec. "Pounds" is more associated with the sense of British currency in the Politicstopic (upper), while closer to the sense of weight unit in the Healthtopic (lower). High probability *paths* are shaded in blue and high probability *edges* have thicker lines. 89

5.1 Past multilingual topic models (MTMs) learn aligned topics across languages, which is problematic on the corpora with low comparability. Our MTM overcomes it by learning weighted topic links without forcing topic alignment: topic pairs with many word translation pairs have high link weights, e.g., (EN-1, ZH-3), (EN-2, ZH-4), and (EN-3, ZH-5); topic pairs with partial overlap receive lower weights, e.g., (EN-5, ZH-1); a topic is unlinked if there is no corresponding topic in the other language (ZH-2). 94
5.2 Our model uses topic link weight matrices $\boldsymbol{\rho}$'s to transform topics from one language to another. Unlike other models, it allows a topic to linked to another topic or multiple topics. 97
5.3 The graphical model of our multilingual topic model. The topic links $\boldsymbol{\rho}$'s, as instantiated by the function Ψ, encourage topics to encourage word translations to have consistent topic distributions. . . . 99

5.4 Topic coherence performance on PACO dataset with the number of top words in each topic. 114
5.5 Topic coherence performance on INCO dataset with the number of top words in each topic. 115
5.6 The models' performance of topic coherence on PACO dataset when the sizes of target language corpora grow from 10% to 100%, with a step size of 10%. 117
5.7 The models' performance of topic coherence on INCO dataset when the sizes of target language corpora grow from 10% to 100%, with a step size of 10%. 118

Chapter 1: Introduction

Given a collection of documents, people want to understand and summarize it at a high level. More specifically, people are interested in what topics the documents are talking about without reading the documents. Text documents contain a large amount of word co-occurrence patterns, which is perfect for statistical models to identify latent topics automatically, hence *topic models* (Boyd-Graber et al., 2017). Given the document and the words in Figure 1.1, topic models infer latent topics. Each latent topic is a distribution over the words and represented by the dominant words, i.e., the words with the highest probabilities. We can then summarize each topic manually by the dominant words. For instance, the first topic's dominant words are "new", "film", "show", "music", and "movie". We can summarize that it is an Arts topic. In addition to topic distributions over words, topic models also give the document distributions over topics (i.e., documents' topic proportions) as indicated by the colors of topics and words in the document. This helps users understand the documents at a high level and forms a three-level hierarchy: documents, words, and topics (Figure 1.1).

Topic models are useful because they are unsupervised, although they can be extended to supervised models if necessary (McAuliffe and Blei, 2008). They

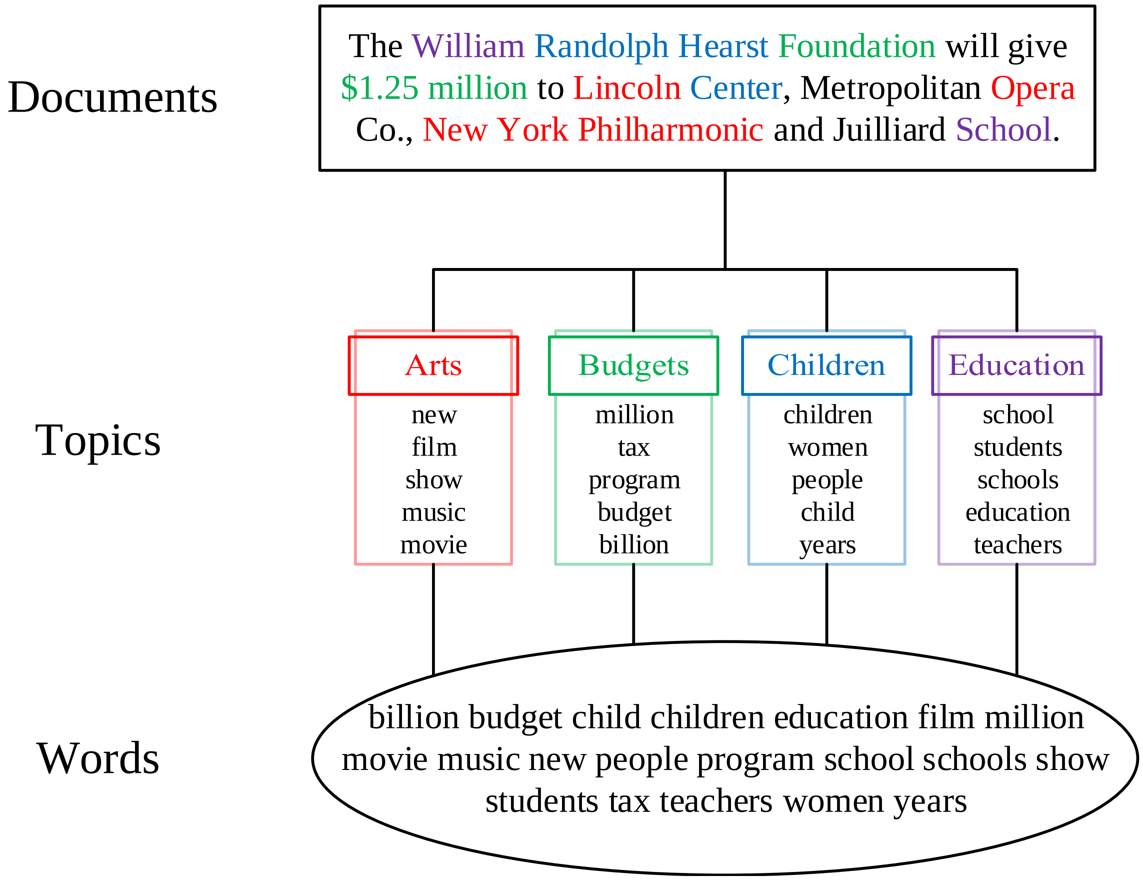

Documents
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School.
Topics
Arts
new
film
show
music
movie
Budgets
million
tax
program
budget
billion
Children
children
women
people
child
years
Education
school
students
schools
education
teachers
Words
billion budget child children education film million movie music new people program school schools show students tax teachers women years

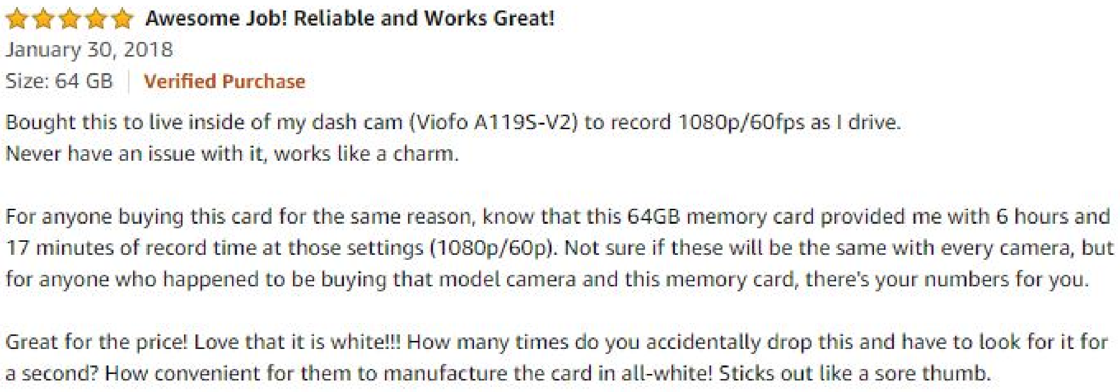
Figure 1.1: An example of the three-level hierarchy in a topic model: high-level documents, low-level words, and mid-level latent topics. A document is a multinomial distribution over latent topics. A topic is a multinomial distribution over words. Topic names are summarized manually.

only require minimal human effort on data preparation (e.g., document collection and preprocessing) and do not need any expensive annotation. In addition to the data, the user only needs to specify the number of topics and the models often give excellent results, although in some cases the model can find the best number of topics by itself (Teh et al., 2006; Blei et al., 2007).

Early topic models are deterministic, applying linear algebra directly on the document-word matrix (Deerwester et al., 1990; Papadimitriou et al., 1998). However, it is often difficult and even awkward to extend the deterministic topic models for adding new components and/or incorporating external knowledge, so topic models soon go probabilistic (Hofmann, 1999). Probabilistic topic models are more powerful than deterministic ones, because they are more flexible to add and/or modify the latent variables and probabilistic distributions, which represent documents, words, and topics, based on the data characteristics and available external knowledge. We will discuss in more depth in Chapter 2.

Among all probabilistic topic models, latent Dirichlet allocation (Blei et al., 2003, LDA) is a powerful and flexible framework. It assumes that the document and words are generated from multinomial distributions of topics and words with Dirichlet priors and yields coherent topics. LDA can be flexibly extended because of its probabilistic nature, so it serves as the base framework for a variety of applications besides the topic discovery (Griffiths and Steyvers, 2004): providing features for ad-hoc information retrieval (Wei and Croft, 2006), disambiguating word senses with WordNet (Boyd-Graber et al., 2007), segmenting multi-party spoken discourse (Purver et al., 2006), modeling user rating profiles for collaborative filtering (Marlin, 2003), and even learning natural scene categories in computer vision (Fei-Fei and Perona, 2005). This dissertation also builds topic models based on LDA.

People extend LDA because they want to add additional constraints and/or guidance to the latent topics so that LDA can take this information into account and produce better topics. Such constraints and guidances are often referred to *external knowledge*, which sometimes comes with the text as metadata or, in other cases, is collected separately. For instance, an Amazon product review often comes with a



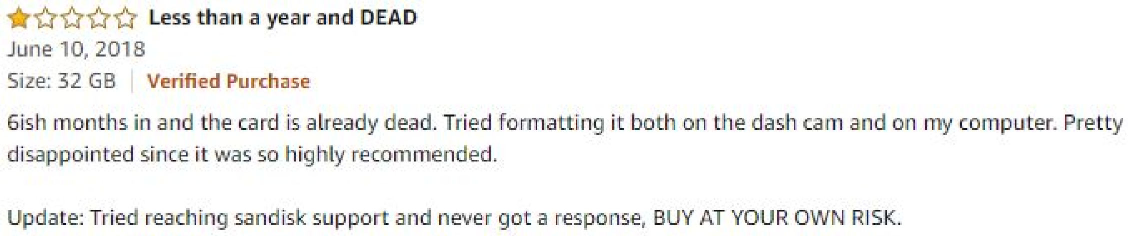
★★★★★ **Awesome Job! Reliable and Works Great!**
January 30, 2018
Size: 64 GB | Verified Purchase

Bought this to live inside of my dash cam (Viofo A119S-V2) to record 1080p/60fps as I drive.
Never have an issue with it, works like a charm.

For anyone buying this card for the same reason, know that this 64GB memory card provided me with 6 hours and 17 minutes of record time at those settings (1080p/60p). Not sure if these will be the same with every camera, but for anyone who happened to be buying that model camera and this memory card, there's your numbers for you.

Great for the price! Love that it is white!!! How many times do you accidentally drop this and have to look for it for a second? How convenient for them to manufacture the card in all-white! Sticks out like a sore thumb.

Figure 1.2: A positive Amazon review for an SD card with five stars.



★☆☆☆☆ **Less than a year and DEAD**
June 10, 2018
Size: 32 GB | Verified Purchase

6ish months in and the card is already dead. Tried formatting it both on the dash cam and on my computer. Pretty disappointed since it was so highly recommended.

Update: Tried reaching sandisk support and never got a response, BUY AT YOUR OWN RISK.

Figure 1.3: A negative Amazon review for an SD card with one star.

rating that indicates the satisfaction of the author towards the product. In positive reviews with five stars, we can usually find words like “love”, “great”, “awesome”, and “reliable” (Figure 1.2). In negative ones, “disappointed”, “bad”, “waste”, and “return” frequently appear (Figure 1.3). This is a useful signal for topic models of what words are likely to appear in positive and negative reviews. Thus, the number of stars can be incorporated and jointly modeled in the generative process (McAuliffe and Blei, 2008). Moreover, if we convert the number of stars to binary labels (e.g., four- and five-star reviews as positive and one- and two-star reviews as negative), we can even apply more advanced techniques in binary classification (Zhu et al., 2012, 2014).

Among all types of external knowledge, *weighted links* are very useful because they indicate some similarity or relatedness between the connected objects. However, it is not easy to model the weighted links and incorporate them into topic modeling, because the links can have different forms and represent different types of similarity or relatedness. For instance, the links can have different densities (e.g., dense versus sparse), amount of information, and types of weights (e.g., integer-valued, real-valued, or non-negative real-valued). This makes a one-for-all model impossible, so we have to develop separate solutions for various types of weighted links. This dissertation studies three distinct types of representative weighted links that correspond to the three-level hierarchy in topic models (documents, words, and topics). Each type of link has its distinct properties and potential applications, and this dissertation introduces methods for incorporating them into topic modeling. The methods are easy to generalize and the ideas behind these methods can be applied to other similar problems involving weighted links.

This dissertation first studies observed *binary-valued* document links and easy-to-obtain *real-valued* word links. Document links indicate topic similarities between the connected documents, e.g., a paper cites another one because they are in the same research area; a Twitter user, if we treat the user's tweets as a document, mentions/retweets/follows another user because of mutual interests. Word links, or word association scores, indicate words' semantic relatedness, e.g., pointwise mutual information (Church and Hanks, 1990, PMI), log-likelihood-ratio (Moore, 2004, LLR), and Fisher's exact test (Upton, 1992, FET). For these observed and/or easy-to-obtain weighted links, this dissertation develops methods that uncover the latent structures in the weighted links and jointly models them with topics. For the unobserved weighted links, such as the links between latent topics, this dissertation studies how to learn them based on available information. For instance, topics in different languages can be connected if they have similar words according to word translations. We run all experiments on open datasets. Although the datasets are relatively small, they are big enough to validate our methods.

Modeling the links with topics can potentially be helpful in several applications/tasks, according to the link types. With a good topic model for document links, we can use it for suggesting potential links, which can help people explore relevant documents that are not yet linked. By adding word links into topic models, in addition to more coherent topics, we can get richer word relationship information from topic models, which is potentially useful for other tasks such as word sense disambiguation. Weighted topic links are particularly helpful for modeling low-resource languages which have limited data. By modeling them along with high-resource languages, the topic patterns on high-resource languages can be transferred to low-resource languages via the topic links, which yields a better topic model on low-resource languages. This will be useful when document analysis is suddenly needed for a low-resource language, such as in the case of disasters: we can quickly obtain a relatively good topic model on the low-resource language with little effort (e.g., finding a dictionary), understand the situation, and send out corresponding rescue resource.

Besides these specific applications, the work in this dissertation has a broader impact to other research areas. This dissertation demonstrates an existing key insight of induction and deduction in a computational manner: when studying the objects we are interested in, it is beneficial to summarize their commonalities and find common patterns from their properties (induction), and apply the summarized patterns to new objects with the same properties (deduction). The commonalities of objects (e.g., documents' topic distributions, words' semantic relatedness, and shared topics across languages) are summarized using statistics and probabilistic distributions, and then applied to understanding new objects with the same properties (e.g., documents in the same block, words in the same subtree, and topics in low-resource languages). The work in this dissertation also benefits other research areas where text is involved. For instance, in computational social science (CSS), people prefer large amounts of data to find interesting patterns (Lazer et al., 2009). Topic models can help them understand the text data at a high level. With the work in this dissertation, CSS people can add more available information to obtain more accurate results (e.g., adding document co-authorship for topic analysis) and perform more analysis (e.g., the culture difference when framing about the same incident).

1.1 Topic Modeling with Document Network

Many documents are organized in networks with binary edges. Scientific papers, including this dissertation, cite other papers because of the relevance in background, methods, and/or datasets. A webpage (e.g., the homepage of a professor) has hyperlinks to other pages because the two pages are related in some way (e.g., the professor's students, publications, and/or courses). Twitter users mention, retweet, and/or follow each other based on mutual interests. All these examples indicate that if two documents are linked, they must share some topics, which could be useful external knowledge for topic modeling.

The *Relational Topic Model* (Chang and Blei, 2010, RTM) jointly models documents' topics and document links. Besides generating the words in documents, it assumes that each binary document link is generated probabilistically from the weighted sum of the Hadamard (element-wise) product of the two documents' posterior topic distributions.

However, RTM ignores the large amount of information in the latent structures of the document network—the link density could split the network into blocks (Figure 1.4). Each block is defined as a subset of documents that are densely connected, but sparsely connected with the ones in other blocks. This allows the model to extract information of every block's topic patterns and use it as informative priors for generating the documents' topics in the blocks.

Thus, Chapter 3 introduces LBH-RTM, which integrates a *weighted stochastic block model* (Aicher et al., 2014, WSBM) for block discovery (Figure 1.4) and then learns the blocks' topic distributions to assist document topic modeling. In contrast to RTM, which uses only topical features for link prediction, LBH-RTM also includes the similarity of documents' word usage and the relationship between the documents' assigned blocks. Moreover, it obtains the document link probability with a max-margin objective function which is more robust than the sigmoid function in binary classification. On both scientific paper abstracts and webpages, it better predicts citations and hyperlinks and gives more coherent topics.

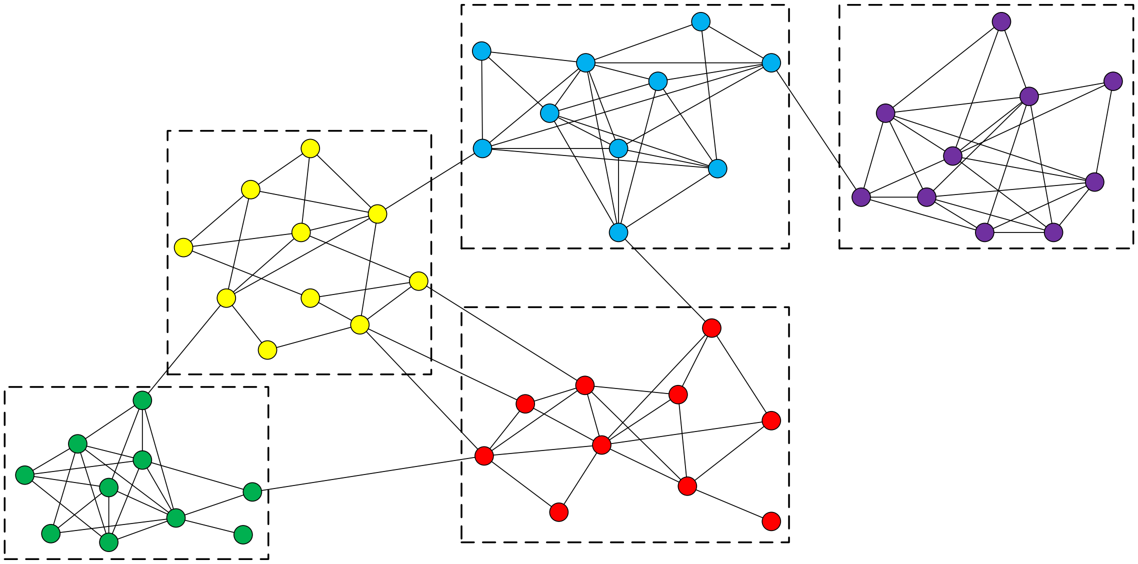


Figure 1.4: WSBM identifies blocks in a network as denoted by colors and dashed boxes. Each block is a subset of nodes (denoted by circles) that are densely connected with each other but sparsely connected with the nodes in other blocks. We integrate WSBM into topic models and extract the blocks' topic patterns for better modeling documents' topics and predicting document links.

1.2 Topic Modeling with Word Associations

Real-valued word association scores link words using traditional statistical methods or more recent word embedding techniques (Mikolov et al., 2013; Pennington et al., 2014). Word associations denote the semantic relatedness of the connected words. Higher association scores denote higher relatedness and more frequent co-occurrences. For example, "science" often co-occurs with "technology", so their association score is high, but "science" is likely to have a low association score with "cat" because they rarely co-occur. Word association scores are easy to obtain and contain a vast amount of information of words' semantic relatedness. Topic models infer latent topics which consist of semantically related words. It is therefore useful to incorporate word association scores into topic modeling.

However, there is redundancy in the word association scores which have a complexity of $O(V^2)$ where V is the size of the vocabulary. Some word association scores have different assumptions from topic models. For instance, word embeddings estimate word associations based on local context windows, while topic models infer topics based on document context. Thus, it is necessary to extract key information and reduce redundancy in word association scores.

In Chapter 4, we introduce three methods to organize the words based on the word association scores in a tree structure, also referred to as tree prior. The methods filter large amounts of redundancy in word association scores and only keep the most salient word links. In a tree prior, words with high association scores are placed in the same small subtree. When tree LDA (Boyd-Graber et al., 2007, tLDA) learns a topic on the tree prior, the probabilities of generating these words are correlated even if the words' term frequencies differ a lot. Experimental results show substantial improvement in topic coherence over LDA on both 20NewsGroups and Amazon review corpora. tLDA also slightly improves the extrinsic classification performance of predicting news documents' categories and positive/negative Amazon reviews.

1.3 Topic Model for Learning Weighted Topic Links

Unlike observed documents and words, topics are latent, so it is difficult to find ground-truth topic links. However, topic links are useful, especially in a multilingual case where topic links connect similar topics based on word semantics across languages. For instance, an English Sports topic with top words "sports", "game", "referee", "champion", and "coach" can be connected with a high weight with a Chinese topic of "运动 (yùn dòng)", "比赛 (bǐ sài)", "裁判 (cái pàn)", "冠军 (guàn jūn)", and "教练 (jiào liàn)", which are direct translations of the English top words, but should not be connected with a topic of "经济 (jīng jì)", "收入 (shōu rù)", "资产 (zī chǎn)", "投资 (tóu zī)", and "股票 (gǔ piào)" in Economy.[1] The weighted topic links can be particularly helpful when modeling low-resource languages in which we have little data to train a good topic model, as they can transfer the well-learned topic patterns from high-resource languages to the low-resource ones and improve the topic model quality on low-resource languages. This can be applied to the case when a disaster takes place at an area where a low-resource language is often used. With the weighted topic links, we can quickly understand the situation from the limited media coverage and social media discussions with the help of high-resource language data and provide assistance needed.

To learn the weighted topic links across languages, we introduce a multilingual topic model (MTM) in Chapter 5. Unlike previous MTMs that require the same numbers of topics or even force the topic spaces to be aligned across languages, our MTM assumes that each language has its own topic distributions over its own words while the numbers of topics do not have to be the same across languages, and *only* connects topics when their dominant words are close in senses based on a word translation dictionary. This keeps the model robust and giving coherent topics when the corpora are less comparable across languages. The topic links are learned by

[1] The English translations of the Chinese Economy words are "economics", "income", "assets", "invest", and "stock".

minimizing the translation pairs' topic distribution distances after transformation by the topic link weights. We validate the model on bilingual classification tasks where we use the topic posteriors of the documents as features. Results show that our MTM substantially outperforms previous MTMs and monolingual LDA both intra- and cross-lingually. Also, our MTM gives coherent topics when the corpora get less comparable or even incomparable, and the corpora sizes get small, in which case previous MTMs sacrifice topic coherence for topic alignment.

1.4 Additional Contributions

Although the methods introduced in this dissertation incorporate and learn weighted links in the text, the intuition and ideas behind the methods apply to more general settings with some extension and/or adaptation. Besides the introduced specific topic models, this dissertation also makes the following contributions to the fields of machine learning and natural language processing:[2]

- This dissertation introduces the idea of uncovering the latent structures in weighted links when jointly modeling links and topics. This applies to other research problems involving joint modeling with weighted links or networks, no matter they are dense or sparse. For dense networks, hierarchical clustering could reduce redundancy while keeping important information (Chapter 4); for sparse ones, identifying small blocks helps to categorize the nodes and

[2] The code for Chapters 3 and 4 is available at `https://github.com/ywwbill/YWWTools`. The work in Chapter 5 is submitted and being reviewed at EMNLP 2019 as of the submission of this dissertation, so the code for Chapter 5 will be added to the repository upon paper acceptance.

facilitate downstream tasks (Chapter 3).

- This dissertation shows the superiority of hinge loss over the conventional sigmoid loss and integrates it with topic modeling in a joint framework. The hinge loss is known for its robustness and good performance. Although we use it for link prediction, it is easy to generalize it to *any* probabilistic model for binary classification tasks.

- This dissertation introduces a novel and robust multilingual topic model from a new angle which does not align topic spaces across languages but instead connects topics only when necessary (Chapter 5). Although the MTM is introduced in a bilingual case, it can be easily extended to multilingual ones. Multilingual knowledge is encoded via a posterior regularizer, and is therefore very flexible to encode the knowledge by any other formulas without changing the model's main structure.

Chapter 2: Background

This chapter introduces topic models with an emphasis on latent Dirichlet allocation (Blei et al., 2003, LDA), including its generative process, posterior inference, and evaluation methods. This chapter also includes some extension methods for LDA that are relevant to our work.

Topic models find latent topics among a set of documents. They are unsupervised, so they do not require expensive annotations but only limited effort of data collection and preprocessing. They infer latent topics and tell people the documents' proportions of topics which serve as high-level summaries of documents. Thus, topic models make it easier for people to analyze extensive collections of unstructured text and reveal insights without reading the documents (Griffiths and Steyvers, 2004; Marwick, 2013; Yang et al., 2011).

Topic models assume that each document d is a distribution of K topics and each topic k is a distribution of V words, denoted by $\boldsymbol{\theta_d}$ and $\boldsymbol{\phi_k}$ respectively. Early topic models like latent semantic analysis (Dumais, 2004, LSA) apply deterministic linear algebra on the document-word matrix M of size $V \times D$ where D is the number of documents. Each cell, $M_{v,d}$, denotes the term frequency of word v in document d. LSA then applies singular value decomposition (SVD) which breaks

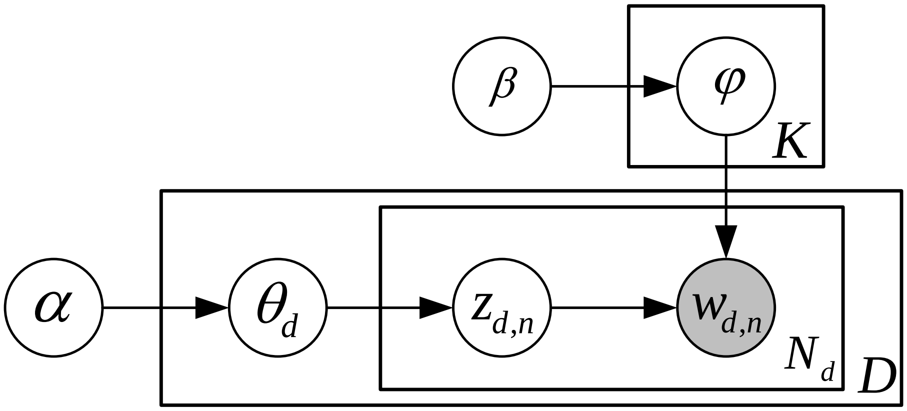


Figure 2.1: The Graphical Model of LDA.

down the matrix M with the number of topics K into the product of three matrices namely 1) $\boldsymbol{\phi}$ of size $V \times K$, 2) $\boldsymbol{\Sigma}$ of size $K \times K$, and 3) $\boldsymbol{\theta}$ of size $K \times D$. Each column of $\boldsymbol{\phi}$ denotes a topic distribution over words and each column of $\boldsymbol{\theta}$ denotes a document distribution over topics.[1]

Unfortunately, due to the deterministic characteristics, LSA is challenging to extend or incorporate external knowledge. With the emergence of Bayesian methods and conjugate priors, which are more flexible than LSA, recent topic models are developed based on probabilistic methods like LDA.

2.1 LDA Introduction

Latent Dirichlet allocation (Blei et al., 2003, LDA) is a probabilistic generative model. It assumes that each document d is a mixture of K topics, denoted by a vector $\boldsymbol{\theta_d}$ of length K. Each latent topic k is a distribution over the vocabulary of size V, denoted by a vector $\boldsymbol{\phi_k}$ of length V. For instance, if we apply LDA on

[1] In LDA, we assume $\boldsymbol{\phi}$ is of size $K \times V$ and each row denotes a topic distribution over words and so for $\boldsymbol{\theta}$.

Topic	Top Words
Country	countries, africa, india, china, country, billion, states, chinese, economy, global, population, united, economic, growth, health
Health	cancer, disease, health, patient, heart, cells, patients, body, blood, care, treatment, hiv, medical, drug, data
Education	school, social, kids, education, children, learn, ideas, community, learning, group, students, game, places, schools, problems
Astronomy	universe, space, earth, light, science, planet, stars, matter, black, physics, mars, theory, sun, dark, billion
Information Technology	data, computer, information, technology, internet, machine, video, web, computers, digital, media, phone, online, robots, software

Table 2.1: Five example topics obtained from the TED talk corpus using LDA. Each topic is represented by the top fifteen words with the highest probabilities in that topic. Topic categories are obtained manually.

English TED talks with fifteen topics, five of them may be similar with the ones in Table 2.1, as represented by the words with highest probability masses in the topics.

To generate a token in document d, LDA first picks a topic k from the document's topic distribution $\boldsymbol{\theta_d}$ and then picks a word from topic k's word distribution $\boldsymbol{\phi_k}$. In the formal description, the generative process of LDA is as follows and corresponds to the graphical model in Figure 2.1.

1. For each topic $k \in \{1, \ldots, K\}$
    (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dirichlet}(\beta)$
2. For each document $d \in \{1, \ldots, D\}$
    (a) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dirichlet}(\alpha)$
    (b) For each token $t_{d,n}$ in document d
        i. Draw a topic $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta_d})$
        ii. Draw a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{z_{d,n}}})$

where α and β are pre-defined hyperparameters of the (conjugate) Dirichlet priors for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ respectively.

2.1.1 Posterior Inference

Once we define the generative process and obtain the documents, the next step is to infer the parameters in latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ that best fit the observed data, i.e., posterior inference. Gibbs sampling is a commonly used method to perform posterior inference (Geman and Geman, 1984; Resnik and Hardisty, 2010). It assumes that every token is assigned to a topic which is randomly chosen during initialization. Then it iteratively updates every token's topic assignment with probabilities calculated based on some statistics excluding the current token. The equations for computing the topics' probabilities are called Gibbs sampling equations which are the core of Gibbs sampling.

To obtain the Gibbs sampling equation, we first define the joint probability of generating the tokens $\mathbf{w}$ and tokens' topic assignments $\mathbf{z}$ with current parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, α, and β:

$$\Pr\left(\mathbf{w}, \mathbf{z} \,|\, \alpha, \beta\right) = \underbrace{\Pr\left(\mathbf{z} \,|\, \alpha\right)}_{\text{Genearting topic assignments.}} \underbrace{\Pr\left(\mathbf{w} \,|\, \mathbf{z}, \beta\right)}_{\text{Generating tokens.}} \tag{2.1}$$

$$= \int \Pr\left(\mathbf{z} \,|\, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{\theta} \,|\, \alpha\right) \mathrm{d}\boldsymbol{\theta} \int \Pr\left(\mathbf{w} \,|\, \mathbf{z}, \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\phi} \,|\, \beta\right) \mathrm{d}\boldsymbol{\phi}, \tag{2.2}$$

where the expansions are based on the definition of the generative process and graphical model.

Then we replace the probabilities in Equation 2.2 with the definitions of Dirichlet and multinomial distributions:

$$
\begin{aligned}
\Pr\left(\mathbf{w}, \mathbf{z} \,|\, \alpha, \beta\right) = &\int \underbrace{\left(\prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{N_{d,k}}\right)}_{\text{Multinomial}} \underbrace{\left(\prod_{d=1}^{D}\frac{\Gamma\left(K\alpha\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha\right)}\prod_{k=1}^{K}\theta_{d,k}^{\alpha-1}\right)}_{\text{Dirichlet}} \mathrm{d}\boldsymbol{\theta} \\
&\int \underbrace{\left(\prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{N_{k,v}}\right)}_{\text{Multinomial}} \underbrace{\left(\prod_{k=1}^{K}\frac{\Gamma\left(V\beta\right)}{\prod_{v=1}^{V}\Gamma\left(\beta\right)}\prod_{v=1}^{V}\phi_{k,v}^{\beta-1}\right)}_{\text{Dirichlet}} \mathrm{d}\boldsymbol{\phi}, \qquad (2.3)
\end{aligned}
$$

where $N_{d,k}$ denotes the number of tokens in document d that are assigned to topic k; $N_{k,v}$ denotes the number of times that word v is assigned to topic k; $\Gamma(\cdot)$ is the Gamma function:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t. \qquad (2.4)$$

Here, we use its property

$$\Gamma(x+1) = x\Gamma(x). \qquad (2.5)$$

We then drop the constants and combine the terms in Equation 2.3:

$$\Pr\left(\mathbf{w}, \mathbf{z} \,|\, \alpha, \beta\right) \propto \int \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{N_{d,k}+\alpha-1}\mathrm{d}\boldsymbol{\theta} \int \prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{N_{k,v}+\beta-1}\mathrm{d}\boldsymbol{\phi}. \qquad (2.6)$$

The elegant property of the conjugacy of Dirichlet and multinomial distributions allows us to integrate out $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ from Equation 2.6, after adding some constants:

$$\Pr\left(\mathbf{w},\mathbf{z}\,|\,\alpha,\beta\right) \tag{2.7}$$

$$\propto \prod_{d=1}^{D} \frac{\prod_{k=1}^{K}\Gamma\left(N_{d,k}+\alpha\right)}{\Gamma\left(N_{d,\cdot}+K\alpha\right)} \underbrace{\int \prod_{d=1}^{D} \frac{\Gamma\left(N_{d,\cdot}+K\alpha\right)}{\prod_{k=1}^{K}\Gamma\left(N_{d,k}+\alpha\right)} \prod_{k=1}^{K} \theta_{d,k}^{N_{d,k}+\alpha-1} \mathrm{d}\boldsymbol{\theta}}_{\text{Dirichlet posterior equals to 1.}}$$

$$\prod_{k=1}^{K} \frac{\prod_{v=1}^{V}\Gamma\left(N_{k,v}+\beta\right)}{\Gamma\left(N_{k,\cdot}+V\beta\right)} \underbrace{\int \prod_{k=1}^{K} \frac{\Gamma\left(N_{k,\cdot}+V\beta\right)}{\prod_{v=1}^{V}\Gamma\left(N_{k,v}+\beta\right)} \prod_{v=1}^{V} \phi_{k,v}^{N_{k,v}+\beta-1} \mathrm{d}\boldsymbol{\phi}}_{\text{Dirichlet posterior equals to 1.}} \tag{2.8}$$

$$\propto \underbrace{\left(\prod_{d=1}^{D} \frac{\prod_{k=1}^{K}\Gamma\left(N_{d,k}+\alpha\right)}{\Gamma\left(N_{d,\cdot}+K\alpha\right)}\right)}_{\text{The denominators are constants.}} \left(\prod_{k=1}^{K} \frac{\prod_{v=1}^{V}\Gamma\left(N_{k,v}+\beta\right)}{\Gamma\left(N_{k,\cdot}+V\beta\right)}\right) \tag{2.9}$$

$$\propto \left(\prod_{d=1}^{D}\prod_{k=1}^{K}\Gamma\left(N_{d,k}+\alpha\right)\right) \left(\prod_{k=1}^{K} \frac{\prod_{v=1}^{V}\Gamma\left(N_{k,v}+\beta\right)}{\Gamma\left(N_{k,\cdot}+V\beta\right)}\right), \tag{2.10}$$

where $\cdot$ denotes marginal counts, i.e., $N_{d,\cdot} = \sum_{k=1}^{K} N_{d,k}$.

Finally, we derive the Gibbs sampling equation for updating $z_{d,n}$, the topic assignment of the n-th token in document d, as the quotient of the joint probabilities including and excluding the token:

$$\Pr\left(z_{d,n}=k\,|\,\mathbf{z}^{-\mathbf{d},\mathbf{n}},\mathbf{w}^{-\mathbf{d},\mathbf{n}},w_{d,n}=v,\alpha,\beta\right) \tag{2.11}$$

$$=\frac{\Pr\left(z_{d,n}=k,\mathbf{z}^{-\mathbf{d},\mathbf{n}},w_{d,n}=v,\mathbf{w}^{-\mathbf{d},\mathbf{n}}\,|\,\alpha,\beta\right)}{\Pr\left(\mathbf{z}^{-\mathbf{d},\mathbf{n}},\mathbf{w}^{-\mathbf{d},\mathbf{n}}\,|\,\alpha,\beta\right)} \tag{2.12}$$

$$\propto \left(\prod_{d'=1}^{D}\prod_{k'=1}^{K} \frac{\Gamma\left(N_{d',k'}+\alpha\right)}{\Gamma\left(N_{d',k'}^{-d,n}+\alpha\right)}\right) \left(\prod_{k'=1}^{K} \frac{\Gamma\left(N_{k',\cdot}^{-d,n}+V\beta\right)}{\Gamma\left(N_{k',\cdot}+V\beta\right)} \prod_{v'=1}^{V} \frac{\Gamma\left(N_{k',v'}+\beta\right)}{\Gamma\left(N_{k',v'}^{-d,n}+\beta\right)}\right) \tag{2.13}$$

$$\propto \frac{\Gamma\left(N_{d,k}+\alpha\right)}{\Gamma\left(N_{d,k}^{-d,n}+\alpha\right)} \frac{\Gamma\left(N_{k,\cdot}^{-d,n}+V\beta\right)}{\Gamma\left(N_{k,\cdot}+V\beta\right)} \frac{\Gamma\left(N_{k,v}+\beta\right)}{\Gamma\left(N_{k,v}^{-d,n}+\beta\right)} \tag{2.14}$$

$$\propto \left(N_{d,k}^{-d,n}+\alpha\right) \frac{N_{k,v}^{-d,n}+\beta}{N_{k,\cdot}^{-d,n}+V\beta}, \tag{2.15}$$

where $^{-d,n}$ denotes the count excluding the n-th token in document d. The final step is based on the property of Gamma function (Equation 2.5) and the differences between the numerators and denominators, e.g., $N_{d,k} = N_{d,k}^{-d,n} + 1$. The overall time complexity of Gibbs sampling for LDA is $O(MKN)$ where M is the number iterations, K is the number of topics, and N is the total number of tokens in the training corpus.

When the posterior inference converges, the values of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are estimated using the final state of topic assignments as

$$\theta_{d,k} = \frac{N_{d,k} + \alpha}{N_{d,\cdot} + K\alpha} \tag{2.16}$$

$$\phi_{k,v} = \frac{N_{k,v} + \beta}{N_{k,\cdot} + V\beta}. \tag{2.17}$$

The core of an LDA model is its $\boldsymbol{\phi}$ matrix, the topic distributions over words. Once we finish training, the $\boldsymbol{\phi}$ matrix of the model is fixed and can be applied on an unseen corpus to infer the topic distributions of new documents. Thus, in the inference process for new documents, we fix $\boldsymbol{\phi}$ and only update the $\boldsymbol{\theta}$, the document distributions over topics:

$$\Pr\left(z_{d,n} = k \,|\, \mathbf{z}^{-\mathbf{d},\mathbf{n}}, \mathbf{w}^{-\mathbf{d},\mathbf{n}}, w_{d,n} = v, \alpha\right) \propto \left(N_{d,k}^{-d,n} + \alpha\right) \phi_{k,v}. \tag{2.18}$$

2.2 Topic Model Evaluation

As many other NLP methods, topic models can be evaluated both extrinsically and intrinsically. Extrinsic evaluation applies the output of a topic model to another task and evaluates the performance of *that* task. For instance, we can take each document's topic posteriors inferred by a topic model as a representation of the document and use it as features for classification. If the topic posteriors are good representations of documents, we can expect good classification performance.

Intrinsic evaluation, on the contrary, does not involve any downstream tasks. It evaluates topic models on one or more key metrics that can be computed independently. A straightforward intrinsic evaluation is to estimate the likelihood of a trained model on an unseen corpus in the same domain of the training data. If a topic model is well trained, the likelihood of generating the unseen corpus in the same area should be high.

It takes two steps to obtain the model's likelihood on a new corpus. The first step is to infer the topic assignments of all tokens in the new documents using Equation 2.18 while ignoring all out-of-vocabulary (OOV) words. Then we compute the log-likelihood of generating the new corpus using the current model as

$$\mathcal{L}\left(\mathbf{w} \,|\, \boldsymbol{\theta}, \boldsymbol{\phi}\right) = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \log \left(\sum_{k=1}^{K} \theta_{d,k} \phi_{k,w_{d,n}} \right), \tag{2.19}$$

which is basically adding up the log likelihood of generating every token using the trained model and the new documents' topic distributions.

However, the scale of the log likelihood depends on the size of the unseen corpus. The log likelihood gets lower as the size of the unseen corpus increases, which makes the log likelihood values incomparable across corpora. Thus, people often use perplexity which normalizes the log-likelihood by the total number of tokens in the unseen corpus to evaluate the model quality:

$$\mathcal{P}\left(\mathbf{w} \,|\, \boldsymbol{\theta}, \boldsymbol{\phi}\right) = \exp\left(- \frac{\mathcal{L}\left(\mathbf{w} \,|\, \boldsymbol{\theta}, \boldsymbol{\phi}\right)}{\sum_{d=1}^{D} N_d} \right). \tag{2.20}$$

The perplexity can also be interpreted as the expected size of vocabulary with uniform word distribution that the model would need to generate a token of the unseen corpus (Heinrich, 2008). In other words, perplexity indicates the number

Topic	Words
1	dog, cat, horse, apple, pig, cow
2	car, teacher, platypugs, agile, blue, Zaire

Table 2.2: It is easy to find out an "intruder" in a coherent topic, such as "apple" in Topic 1, but hard in an incoherent one as Topic 2 (Chang et al., 2009).

of bits the model requires to encode the data. Thus, a lower perplexity denotes a better topic model.

Another intrinsic evaluation, *word intrusion*, mainly focuses on the interpretability or the coherence of the topic words (Chang et al., 2009). It is designed to evaluate manually how well each topic's top words are related to each other. Namely, for each topic, the human evaluators are given the words with the highest probabilities in that topic and an irrelevant "intruder" word chosen elsewhere.

The intuition behind word intrusion is that if the topic words are of good coherence, it is relatively easy for human evaluators to find the "intruder". For instance, in Topic 1 of Table 2.2, human evaluators can easily tell that this topic is about Animals from the words "dog", "cat", "horse", "pig", and "cow". Thus the word "apple" is an "intruder" because it is not an animal. On the contrary, if the topic is incoherent, like Topic 2, it is difficult to find the "intruder" or sometimes, every word looks like an "intruder". So the more "intruders" are found, the better the topic model is.

A distinct disadvantage of word intrusion is that it requires a lot of human effort, so people have developed an automatic alternative (Lau et al., 2014). Specifically, it computes the average word association score of the pair-wise top words of every topic. Given the top N words in a topic, a topic's coherence is measured on a large reference corpus as

$$C(k) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} S(v_{k,i}, v_{k,j}), \tag{2.21}$$

where $v_{k,i}$ denotes the word with the i-th highest weight in topic k; $S(\cdot, \cdot)$ denotes the word association score of the two words on a reference corpus.[2] Then the model's coherence is defined as the average topic coherence

$$C = \frac{1}{K} \sum_{k=1}^{K} C(k). \tag{2.22}$$

This method interprets the topic coherence as the topic's top words' average association scores, which matches the intuition and correlates human evaluations well according to experiments (Lau et al., 2014). According to the experiments, the Pearson correlation coefficient between word intrusion and this method is 0.865 in the domain of news articles. It also requires less human effort in evaluation, so it has been adopted as an intrinsic evaluation metric for topic models by the community, including this dissertation.

2.3 Topic Model Extensions

As introduced at the beginning of this chapter, a significant strength of probabilistic topic models like LDA over past deterministic ones is their flexibility for extensions. This allows people to extend the models for more general purpose and/or include some specific characteristics of the data. To extend a topic model, one could relax current assumptions of documents, topics, and/or words, incorporate external knowledge, or both.

[2] The value of N is pre-defined.

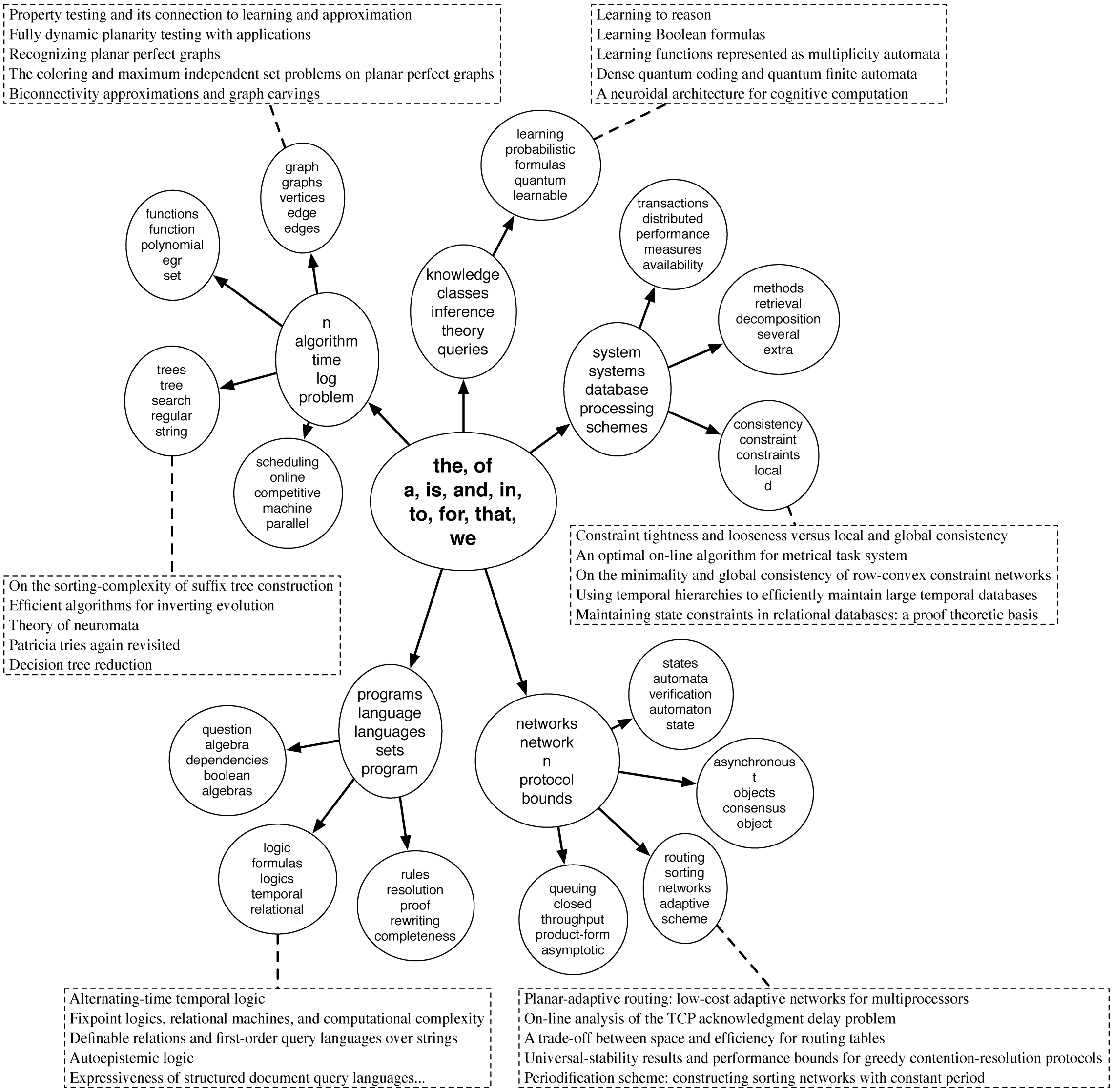

Property testing and its connection to learning and approximation
Fully dynamic planarity testing with applications
Recognizing planar perfect graphs
The coloring and maximum independent set problems on planar perfect graphs
Biconnectivity approximations and graph carvings
Learning to reason
Learning Boolean formulas
Learning functions represented as multiplicity automata
Dense quantum coding and quantum finite automata
A neuroidal architecture for cognitive computation
learning probabilistic formulas quantum learnable
graph graphs vertices edge edges
functions function polynomial egr set
transactions distributed performance measures availability
knowledge classes inference theory queries
methods retrieval decomposition several extra
n algorithm time log problem
system systems database processing schemes
trees tree search regular string
consistency constraint constraints local d
scheduling online competitive machine parallel
the, of a, is, and, in, to, for, that, we
Constraint tightness and looseness versus local and global consistency
An optimal on-line algorithm for metrical task system
On the minimality and global consistency of row-convex constraint networks
Using temporal hierarchies to efficiently maintain large temporal databases
Maintaining state constraints in relational databases: a proof theoretic basis
On the sorting-complexity of suffix tree construction
Efficient algorithms for inverting evolution
Theory of neuromata
Patricia tries again revisited
Decision tree reduction
states automata verification automaton state
programs language languages sets program
networks network n protocol bounds
question algebra dependencies boolean algebras
asynchronous t objects consensus object
logic formulas logics temporal relational
rules resolution proof rewriting completeness
queuing closed throughput product-form asymptotic
routing sorting networks adaptive scheme
Alternating-time temporal logic
Fixpoint logics, relational machines, and computational complexity
Definable relations and first-order query languages over strings
Autoepistemic logic
Expressiveness of structured document query languages...
Planar-adaptive routing: low-cost adaptive networks for multiprocessors
On-line analysis of the TCP acknowledgment delay problem
A trade-off between space and efficiency for routing tables
Universal-stability results and performance bounds for greedy contention-resolution protocols
Periodification scheme: constructing sorting networks with constant period

Figure 2.2: A portion of the hierarchy learned by hierarchical LDA on abstracts of the Journal of ACM (Blei et al., 2007). Coarse-grained words are closer to the root, while fine-grained words are at leaves.

2.3.1 Relaxing Current Assumptions

Relaxing the assumptions directly changes the underlying assumptions of documents, topics, words, and/or their associated distributions and yields a new topic model. For instance, LDA assumes the number of topics and the size of vocabulary are fixed, but with Dirichlet process (Ferguson, 1973, DP), LDA could theoretically have infinite numbers of topics (Teh et al., 2006) and/or vocabulary (Zhai and Boyd-Graber, 2013). Moreover, DP can even help LDA find the number of topics that best fits the data.

We can also change the fundamental structures of topics and words. In "vanilla" LDA, topics are organized in a flat structure. With the help of nested DP, topics can be organized in a hierarchy, which is called hierarchical LDA (Blei et al., 2007, hLDA). Every node in the hierarchy is a topic, i.e., a distribution over words. A child topic such as Algorithm, System, Programs, or Networks emphasizes on a certain area of its father topic (Figure 2.2). This helps to categorize the words with information content—more coarse-grained words, such as "the" and "a", are more likely to be assigned to high-level topics, while more specific words are in low-level topics.

The advantage of DP also applies to hLDA—the topic hierarchy can expand or shrink as needed. hLDA creates new topics when the data does not fit existing topics. It also deletes a topic when there is no token assigned to it.[3]

Recently, with the emergence of word embeddings, representations of words are no longer discrete (Mikolov et al., 2013; Pennington et al., 2014). Instead, words are mapped to a low dimensional continuous semantic space, usually between 100 and 300 dimensions. Thus LDA could be extended to generate such continuous word vectors instead of discrete word types.

Gaussian LDA (Das et al., 2015, GLDA) assumes that a topic (red crosses in Figure 2.3) is a Gaussian distribution in the word embedding space and generates

[3] The topic creation and deletion do not apply to the root topic.

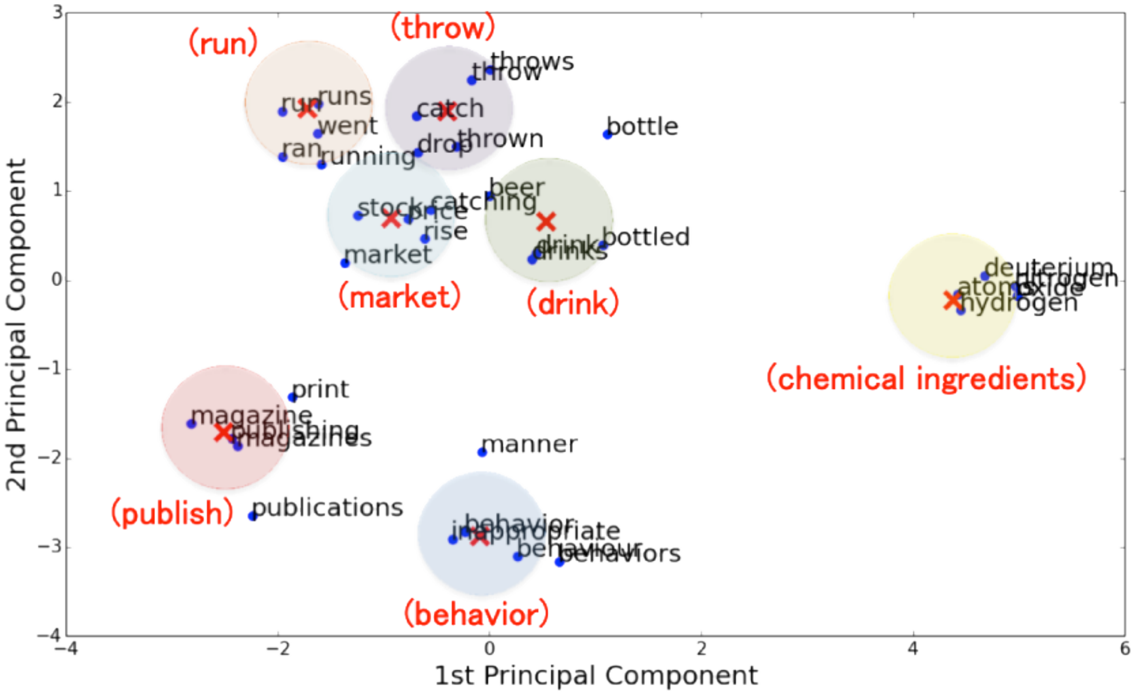


Figure 2.3: The first two principal components of word embeddings and topic/concept vectors in Gaussian LDA (Das et al., 2015, GLDA)/latent concept topic model (Hu and Tsujii, 2016, LCTM).

surrounding word vectors (blue dots). However, some topically related words may be far away in the word embedding space. For instance, the words “neural” and “net” may have very different word embeddings if the corpus has many biological documents, but they are related in the computer science topic. GLDA is not able to put these two words in the same topic, so it is further extended to the latent concept topic model (Hu and Tsujii, 2016, LCTM). LCTM renames the “topics” in GLDA to “concepts” and then defines its own “topics” as multinomial distributions over the “concepts”. Thus the concepts for “neural” and “net” could be assigned to the same machine learning topic.[4]

[4] However, this greatly sacrifices the topic interpretability. See Section 4.3.2.

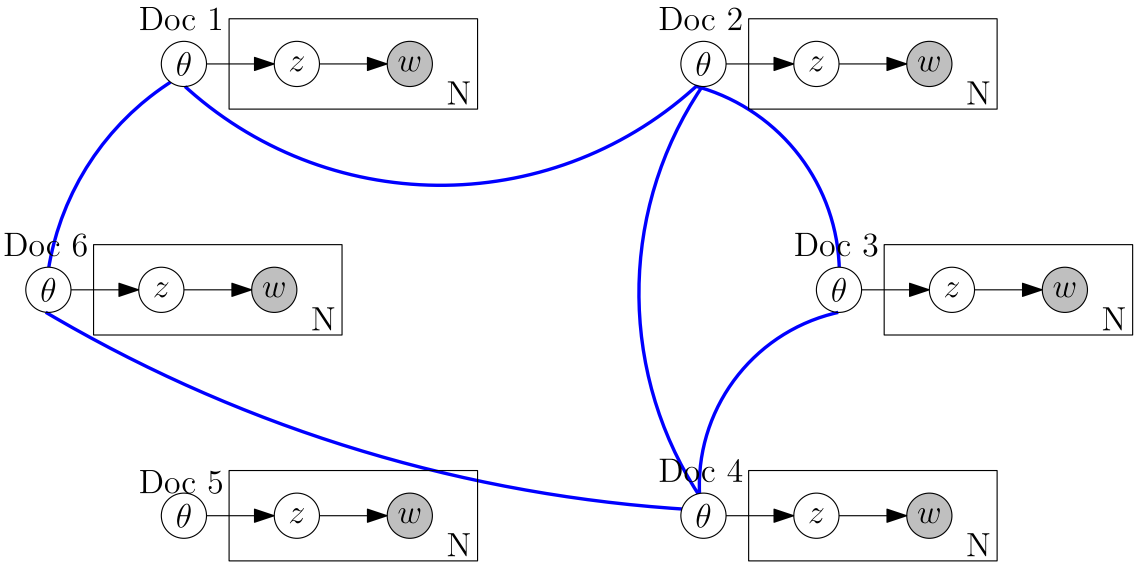


Figure 2.4: An example Markov random topic field in which each document propagates its topic distributions to other ones via links (Daumé III, 2009).

2.3.2 Incorporating External Knowledge

As introduced in Chapter 1, external knowledge, such as document labels and document links, includes valuable extra information for topic models in addition to word co-occurrence patterns. Thus incorporating external knowledge is a straightforward and effective method to improve topic models.

Generally, there are two directions for incorporating external knowledge: upstream and downstream models. The major difference between the two methods is the dependency between external knowledge and topic assignments. Upstream models assume that the topic assignments are conditioned on external knowledge, while downstream models generate external knowledge from topic assignments.

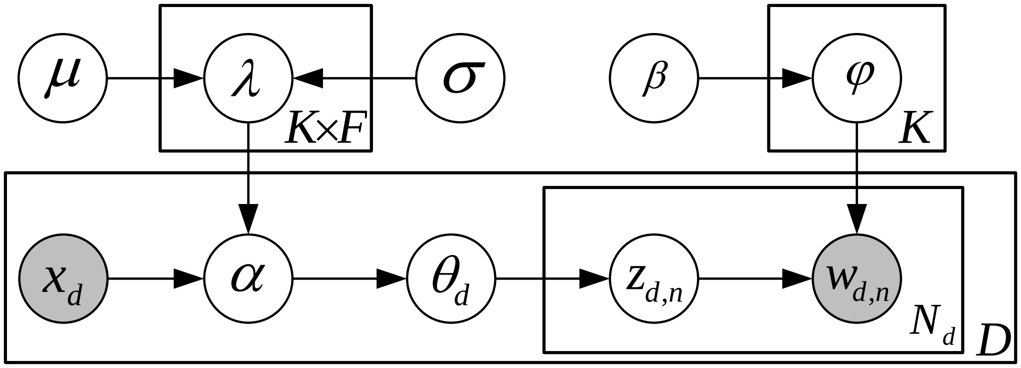


Figure 2.5: An upstream topic model. Topics are conditioned on external knowledge.

2.3.2.1 Upstream Models

In an upstream model, the generation of topic assignments depends on external knowledge. The conditional probability is then written as $\Pr(\mathbf{z} \mid \alpha, \boldsymbol{\theta}, \text{external knowledge})$, using the notations of vanilla LDA (Equation 2.2). For instance, the external knowledge is document links which indicate the connected documents' topic similarities. By using a Markov random topic field (Daumé III, 2009, MRTF) built on document links, a document's topic distribution depends on those of its linked documents (Figure 2.4).

Mimno and McCallum (2012) introduce another upstream topic model which could incorporate arbitrary features for more general settings. It assumes that each document has a feature vector $\mathbf{x}$ and each topic has a weight vector $\boldsymbol{\lambda}$ over the features with a Gaussian prior. The model generates $\boldsymbol{\alpha}$, the Dirichlet prior of document distributions over topics, using the dot product of $\boldsymbol{\lambda}$ and $\mathbf{x}$, which thereby serves as an informative prior of the correlations between topics and document features (Figure 2.5):

1. For each topic $k \in \{1, \ldots, K\}$
    (a) Draw feature weight $\boldsymbol{\lambda_k} \sim \mathcal{N}(0, \sigma^2 I)$
    (b) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dirichlet}(\beta)$
2. For each document $d \in \{1, \ldots, D\}$
    (a) For each topic k let $\alpha_{d,k} = \exp\left(\mathbf{x}_\mathbf{d}^\top \boldsymbol{\lambda_k}\right)$
    (b) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dirichlet}(\boldsymbol{\alpha_d})$
    (c) For each token $t_{d,n}$ in document d
        i. Draw a topic $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta_d})$
        ii. Draw a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{z_{d,n}}})$

In this upstream model, external knowledge (i.e., the features) connects documents and topics via the document feature vector $\mathbf{x}$ and the topic feature weights $\boldsymbol{\lambda}$. The dot product of $\mathbf{x}$ and $\boldsymbol{\lambda}$ is a pre-estimation of the document's tendency towards topics, so it is assigned to $\boldsymbol{\alpha_d}$ and the topic assignments are conditioned on it.

Such an extension can even be applied to computer vision. Fei-Fei and Perona (2005) treat an *image* as a document and each *patch* in the image as a token in the document. A patch's topic is one of K intermediate *themes*, e.g., $\underline{\text{Foliage}}$, $\underline{\text{Water}}$, and $\underline{\text{Sky}}$. Each image is also associated with one of C high-level categories (e.g., coast, highway, or streets) which serve as the supervision. Fei-Fei and Perona (2005) introduce an upstream vision topic model which generates the themes conditioned on the image's category (Figure 2.6):

1. For each image $i \in \{1, \ldots, I\}$

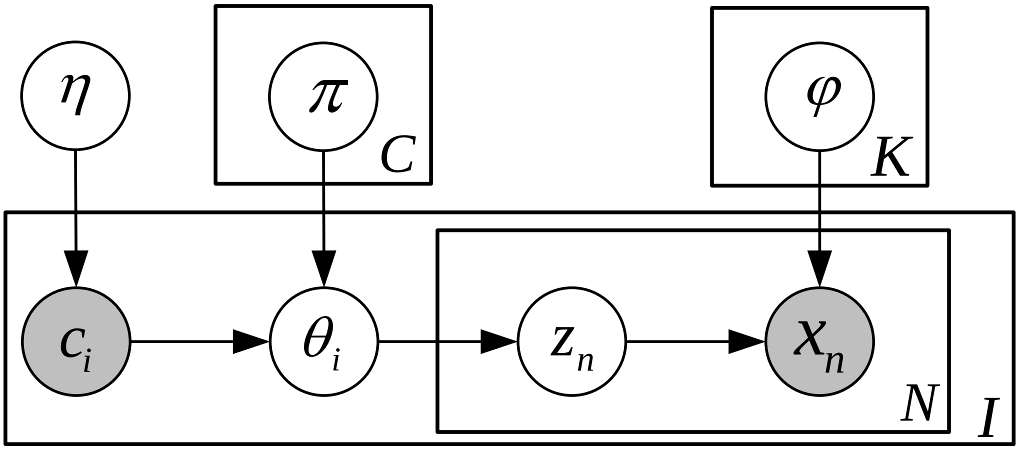


Figure 2.6: An upstream topic model in computer vision (Fei-Fei and Perona, 2005). It uses the image categories as the external knowledge and generates the patches' themes conditioned on them. Some notations are adapted for consistency.

(a) Draw a category $c_i \sim \text{Multinomial}(\boldsymbol{\eta})$

(b) Draw theme distribution $\boldsymbol{\theta_i} \sim \text{Dirichlet}(\boldsymbol{\pi_{c_i}})$

(c) For each of the N patches

i. Draw a theme $z_n \sim \text{Multinomial}(\boldsymbol{\theta_i})$

ii. Draw a patch $x_n \sim \text{Multinomial}(\boldsymbol{\phi_{z_n}})$

This model is similar to the one developed by Mimno and McCallum (2012). It encodes the category knowledge in the informative prior $\boldsymbol{\pi}$ and generates the image's theme (topic) distribution $\boldsymbol{\theta}$ conditioned on the corresponding row of $\boldsymbol{\pi}$ according to the image category assignment.

2.3.2.2 Downstream Models

Downstream models, on the other hand, generates external knowledge based on topic assignments, i.e., $\Pr(\text{external knowledge} \,|\, \mathbf{z})$. As Yang et al. (2015b) point out,

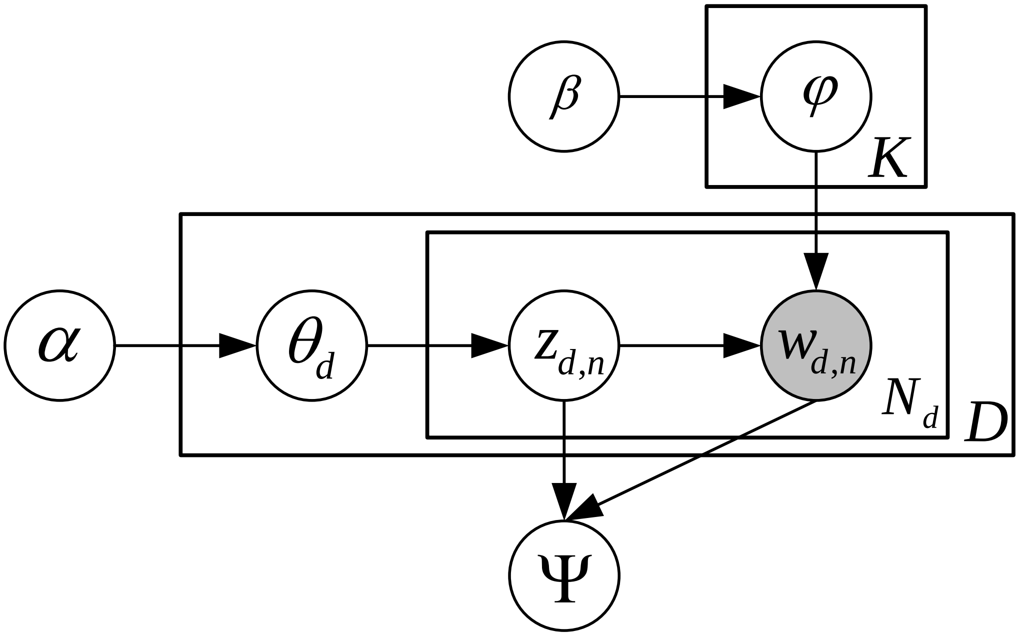


Figure 2.7: A general graphical model of downstream topic models. Note that the external knowledge encoded in the posterior regularizer Ψ conditions on topic assignments $\boldsymbol{z}$ and/or tokens $\boldsymbol{w}$.

each prior knowledge m in the knowledge set M can be represented by a potential function $f_m(\mathbf{z}, \mathbf{w}, \mathbf{d})$ of topic assignments $\mathbf{z}$, words $\mathbf{w}$, and/or documents $\mathbf{d}$. Higher value of $f_m(\mathbf{z}, \mathbf{w}, \mathbf{d})$ indicates better consistency with m at the current state. Finally, all prior knowledge is encoded into a posterior regularizer Ψ as

$$\Psi\left(\mathbf{z}, \mathbf{w}, M\right) = \prod_{m \in M} \exp\left(f_m(\mathbf{z}, \mathbf{w}, \mathbf{d})\right). \tag{2.23}$$

A general downstream topic model first generates tokens and topic assignments following vanilla LDA, and then generates the external knowledge encoded in the posterior regularizer Ψ from the documents, topic assignments, and/or words (Figure 2.7):

1. For each topic $k \in \{1, \ldots, K\}$

   (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dirichlet}(\beta)$

2. For each document $d \in \{1, \dots, D\}$

   (a) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dirichlet}(\alpha)$

   (b) For each token $t_{d,n}$ in document d

      i. Draw a topic $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta_d})$

      ii. Draw a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{z_{d,n}}})$

3. Draw external knowledge $\Psi(\mathbf{z}, \mathbf{w})$

With knowledge potential functions $f_m(\mathbf{z}, \mathbf{w}, \mathbf{d})$ and posterior regularizers Ψ, it is very flexible to incorporate various types of external knowledge. For instance, if each document has a real-valued label y_d (e.g., an Amazon review is associated with an integer rating from one to five), we assume the knowledge potential function of each document d is the log-likelihood of drawing the label from a Gaussian distribution as

$$f(\mathbf{z}, \mathbf{w}, d) = \log \mathcal{N}\left(y_d \,\middle|\, \sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}}, \rho^2 \right), \tag{2.24}$$

where $\boldsymbol{\eta}$ is a weight vector to be optimized (introduced in Equations 2.35 and 2.36 later in this chapter) and ρ is a pre-defined variance, and then we get supervised LDA (McAuliffe and Blei, 2008, sLDA) with a posterior regularizer:

$$\Psi = \exp\left(\sum_{d=1}^{D} f(\mathbf{z}, \mathbf{w}, d) \right) = \prod_{d=1}^{D} \mathcal{N}\left(y_d \,\middle|\, \sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}}, \rho^2 \right). \tag{2.25}$$

If the document label y_d is a binary value with one denoting positive sentiment and zero denoting negative sentiment, the knowledge potential function is

$$f(\mathbf{z}, \mathbf{w}, d) = \log \left(y_d \sigma \left(\sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}} \right) + (1 - y_d) \left(1 - \sigma \left(\sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}} \right) \right) \right), \tag{2.26}$$

where $\sigma(\cdot)$ is a sigmoid function that $\sigma(x) = 1/(1 + \exp(-x))$, and the posterior regularizer Ψ is

$$\Psi = \prod_{d:y_d=1} \sigma\left(\sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}}\right) \prod_{d':y_{d'}=0} \left(1 - \sigma\left(\sum_{k=1}^{K} \eta_k \frac{N_{d',k}}{N_{d',\cdot}}\right)\right). \tag{2.27}$$

Similarly, if binary document links $y_{d,d'}$ are provided, we get the relational topic model (Chang and Blei, 2010, RTM). The knowledge potential function of a link between documents d and d' is

$$f(\mathbf{z}, \mathbf{w}, d, d') = \log \sigma\left(\sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}}\right), \tag{2.28}$$

which leads to the posterior regularizer of

$$\Psi = \prod_{(d,d'):y_{d,d'}=1} \sigma\left(\sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}}\right). \tag{2.29}$$

The knowledge potential function $f_m(\mathbf{z}, \mathbf{w}, \mathbf{d})$ could also be a metric derived from observed evidence. For instance, each word w has a must-link set M_w^m, which contains the words highly correlated with w, and a cannot-link set M_w^c with the words not correlated with w. A knowledge potential function could be derived to encourage highly correlated words to be assigned to the same topic and uncorrelated words not to be assigned to the same topic (Yang et al., 2015b):

$$f(\mathbf{z}, w) = \log \prod_{u \in M_w^m} \max(\lambda, N_{k,u}) \prod_{v \in M_w^c} \frac{1}{\max(\lambda, N_{k,v})}, \tag{2.30}$$

where λ is a smoothing hyperparameter. Then the posterior regularizer Ψ is

$$\Psi = \prod_{w} \prod_{u \in M_w^m} \max(\lambda, N_{k,u}) \prod_{v \in M_w^c} \frac{1}{\max(\lambda, N_{k,v})}. \tag{2.31}$$

The posterior inference of downstream models can be derived following the steps in Section 2.1.1. The value of the posterior regularizer without the current token remains a constant and can be dropped. Only the posterior regularizer with the current token is included in the final Gibbs sampling equation:

$$\begin{aligned} &\Pr(z_{d,n} = k \,|\, \mathbf{z}^{-\mathbf{d},\mathbf{n}}, w_{d,n} = v, \mathbf{w}^{-\mathbf{d},\mathbf{n}}, \alpha, \beta) \\ &\propto \underbrace{\left(N_{d,k}^{-d,n} + \alpha\right) \frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta}}_{\text{LDA Sampling}} \underbrace{\Psi\left(z_{d,n} = k, \mathbf{z}^{-\mathbf{d},\mathbf{n}}, \mathbf{w}, M\right)}_{\text{Posterior Regularizer}}, \end{aligned} \tag{2.32}$$

where the first two terms are the same with vanilla LDA and the third term is the posterior regularizer which encodes external knowledge. In this formulation, the potential functions shape Gibbs sampling inference: topic assignments are more likely when they are consistent with the external knowledge included in the potential functions. This brings significant flexibility in the expression of the potential function $f_m(\mathbf{z}, \mathbf{w}, \mathbf{d})$. The expression is not restricted to probabilistic distributions, exponential family, or conjugacy. It can be expressed flexibly using the combinations of *any* of the values from topic assignments, words, and/or documents. Moreover, changing the expressions of potential functions does not change the main structure of the topic model or require full re-derivation of the Gibbs sampling equation. Hence this allows more flexible experimentation to find the best formulation.

If the posterior regularizer has some variables (e.g., the weight vector $\boldsymbol{\eta}$ in sLDA) to be optimized, the posterior inference should be made by stochastic EM which consists of an E-step and an M-step in each iteration (Celeux, 1985). If we take sLDA as an example, the E-step updates the topic assignments using Gibbs sampling while keeping $\boldsymbol{\eta}$ fixed:

$$\begin{aligned} &\Pr\left(z_{d,n} = k \,|\, \mathbf{z^{-d,n}}, \mathbf{w^{-d,n}}, w_{d,n} = v, \alpha, \beta\right) \\ &\propto \left(N_{d,k}^{-d,n} + \alpha\right) \frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta} \exp\left(-\frac{\left(y_d - \sum_{k'=1}^{K} \eta_{k'} \frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}} - \frac{\eta_k}{N_{d,\cdot}}\right)^2}{2\rho^2}\right), \end{aligned} \tag{2.33}$$

The M-step optimizes $\boldsymbol{\eta}$ to maximize the likelihood of generating external knowledge. In the optimization, we usually add a Gaussian prior $\mathcal{N}(\mu, \sigma^2)$ on each value of $\boldsymbol{\eta}$ to prevent overfitting:

$$\mathcal{L}(\boldsymbol{\eta}) \propto \log \underbrace{\prod_{d=1}^{D} \exp\left(-\frac{\left(y_d - \sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}}\right)^2}{2\rho^2}\right)}_{\text{Likelihood of generating labels}} \underbrace{\prod_{k=1}^{K} \exp\left(-\frac{(\eta_k - \mu)^2}{2\sigma^2}\right)}_{\text{Priors for } \boldsymbol{\eta}} \tag{2.34}$$

$$\propto -\sum_{d=1}^{D} \frac{\left(y_d - \sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}}\right)^2}{2\rho^2} - \sum_{k=1}^{K} \frac{(\eta_k - \mu)^2}{2\sigma^2}. \tag{2.35}$$

This objective function is maximized using L-BFGS and partial derivatives with respect to every η_k (Liu and Nocedal, 1989):

$$\frac{\partial \mathcal{L}(\boldsymbol{\eta})}{\partial \eta_k} \propto \sum_{d=1}^{D} \frac{y_d - \sum_{k'=1}^{K} \eta_{k'} \frac{N_{d,k'}}{N_{d,\cdot}}}{\rho^2} \frac{N_{d,k}}{N_{d,\cdot}} - \frac{\eta_k - \mu}{\sigma^2}. \tag{2.36}$$

Chapter 3: Topic Modeling with Document Networks

As described in Chapter 1, weighted links in the text contain rich information about the objects they connect. With the topic model extension methods introduced in Section 2.3, we are now able to incorporate weighted text links into topic modeling. In this chapter, we focus on binary-valued document links that indicate the topic similarities of the connected documents.

Documents often appear within a network structure with binary-weight edges: social media users have mentions, retweets, and follower relationships; Web pages have hyperlinks; scientific papers have citations. The phenomenon of *homophily* indicates that network structure interacts with the topics in the text, in that documents linked in a network are more likely to have similar topic distributions (McPherson et al., 2001). For instance, a citation link between two papers suggests that they are about a related field; a hyperlink between two professors' academic homepages indicates they may be colleagues or may have collaborated; and a mentioning link between two social media users often indicates common interests. Conversely, if two documents have similar topic distributions, they are likely to have a link between them. For example, the topic model (Blei et al., 2003, LDA) and block detection papers (Holland et al., 1983) are relevant to this dissertation, so we cite them; if two professors are in the same research area and/or work at the same institution, there is a higher chance that they collaborate and link to each other's homepage; if a social media user A finds another user B with shared interests, then A is more likely to mention, retweet, and/or follow B.

Since document links usually imply topical similarity, it is beneficial to incorporate the binary document links into topic models. Thus, we introduce a new joint topic model, based on the *relational topic model* (Chang and Blei, 2010, RTM), that makes fuller use of the rich link structure within a document network, in contrast to the past methods which model text and links separately (Kim and Leskovec, 2012; Liben-Nowell and Kleinberg, 2007; Chaturvedi et al., 2012). Specifically, our model combines the *weighted stochastic block model* (Aicher et al., 2014, WSBM) with topic modeling to identify blocks which consist of subsets of documents that are densely connected. The WSBM categorizes each document in a network probabilistically as belonging to one of L latent blocks, based on its connections with each block. Our model can be viewed as a principled probabilistic extension of Yang et al. (2015a), where we identify blocks in a document network deterministically as *strongly connected components* (Sharir, 1981, SCC) *before* topic modeling. As in that work, we assign a distinct Dirichlet prior to each block to capture its topical commonalities and guide the topic generation of the documents in that block. A linear regression model with a discriminative, max-margin objective function (Zhu et al., 2012, 2014) is jointly trained to reconstruct the binary links, taking into account the features of documents' topic and word distributions (Nguyen et al., 2013), block assignments, and inter-block link rates.

We validate our approach on a scientific paper abstract dataset and a collection of webpages, with citation links and hyperlinks respectively, to predict links among previously unseen documents and from those new documents to training documents. Combining the WSBM with a topic model leads to substantial improvements in link prediction over previous models; it also improves block detection and topic interpretability. The key advantage in combining WSBM compared to using SCC is its flexibility and robustness in the face of noisy links. Our results also lend additional support for using max-margin learning for a downstream supervised topic model (McAuliffe and Blei, 2008), and show that predictions from lexical as well as topic features improve performance (Nguyen et al., 2013).[1]

3.1 Dealing with Links

In this section, we introduce some basic methods to process links. In a general network (i.e., not restricted to text links) where nodes are connected, the link density is not always distributed evenly. A subset of nodes may be densely connected, while sparsely connected with the rest of the nodes. Thus this subset of nodes forms a *block*. The nodes in the same block usually have similar properties. In terms of the documents in the same block, they are likely to have similar topic distributions. Thus it is essential and useful to identify blocks in a network, either deterministically (Section 3.1.1) or probabilistically (Section 3.1.2).

[1] The work done in this chapter has been published in "Birds of a Feather Linked Together: A Discriminative Topic Model using Link-based Priors" (Yang et al., 2015a) and "A Discriminative Topic Model using Document Network Structure" (Yang et al., 2016).

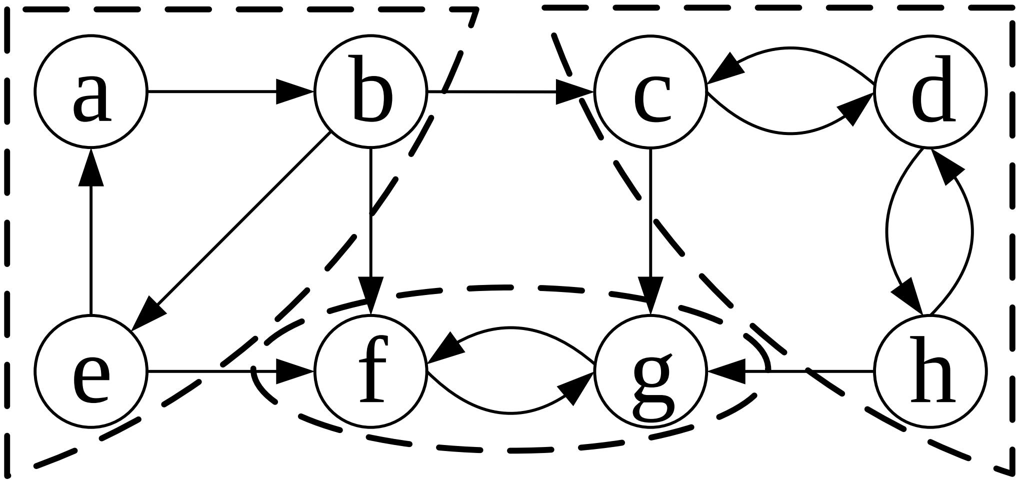


Figure 3.1: An example of strongly connected components. Every pair of nodes in the same component can reach each other via the nodes *only* in that component. The figure is adapted from `https://commons.wikimedia.org/w/index.php?curid=647584`

To incorporate the document links, the *relational topic model* (Chang and Blei, 2010, RTM) is a basic *downstream* model (Section 3.1.3). It encodes every document link in the posterior regularizer and jointly models topics and document links by encouraging connected documents to have similar topic distributions.

3.1.1 Strongly Connected Components

Strongly connected components (Sharir, 1981, SCC) is a deterministic method to identify small clusters or cliques in a network. In each block identified by SCC, every node is reachable from any other nodes in the same component, via path(s) along the nodes in this component *only* (Figure 3.1). Thus in the blocks which it identifies, the nodes are very closely connected and likely to share similar patterns.

SCC identifies blocks using a depth-first search (DFS). It starts from a node that has not been assigned to any blocks and creates a new block with that node.

Then it uses DFS to search for other unassigned nodes that are reachable from and to the new block and includes them into the current block. SCC repeats this procedure until all nodes have block assignments.

The major advantage of the SCC algorithm is its efficiency and the fact that it is non-parametric. It does not require a pre-selection of the number of clusters. Instead, it decides on its own during the process of DFS which has an approximate time complexity of $O(|D| + |E|)$ where $|D|$ denotes the number nodes (documents) and $|E|$ denotes the number of edges. Its major disadvantage is also obvious. SCC only cares whether two nodes are connected, without taking into account the link density of neighboring nodes. Thus it has a high variance in the output: if we make a slight change in the input by adding a link that connects two blocks, the output of SCC will change significantly—it merges the two previously independent blocks into a big one.[2]

3.1.2 Weighted Stochastic Block Model

Weighted stochastic block model (Aicher et al., 2014, WSBM), on the other hand, is a probabilistic generative block detection method. It generalizes the *stochastic block model* (Holland et al., 1983; Wang and Wong, 1987, SBM) and can model nonnegative integer-weight links, instead of binary-weight links.

The graphical model of WSBM is given in Figure 3.2. WSBM assumes that each of the D nodes (documents) belongs to exactly one of L latent blocks. The block assignments are drawn from a multinomial distribution $\boldsymbol{\mu}$ with a Dirichlet

[2] Look ahead to Figure 3.3.

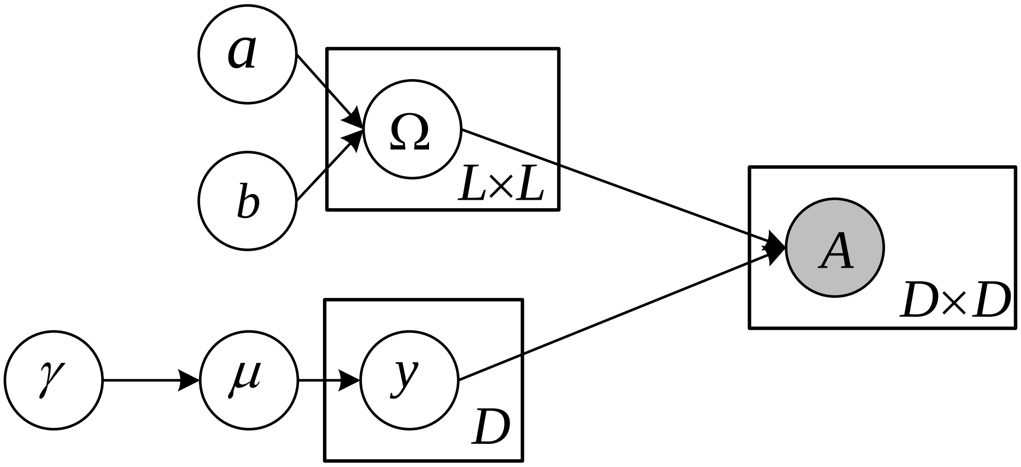


Figure 3.2: The graphical model of weighted stochastic block model (Aicher et al., 2014, WSBM).

prior parameterized by γ. A nonnegative integer-weight link connecting two nodes (documents) in blocks l and l' has a weight generated from a Poisson distribution with parameters $\Omega_{l,l'}$ which has a Gamma prior with parameters a and b. The full generative process is:

1. For each pair of blocks $(l, l') \in \{1, \dots, L\}^2$
   (a) Draw inter-block link rate $\Omega_{l,l'} \sim \text{Gamma}(a, b)$
2. Draw block distribution $\boldsymbol{\mu} \sim \text{Dirichlet}(\gamma)$
3. For each node (document) $d \in \{1, \dots, D\}$
   (a) Draw block assignment $y_d \sim \text{Multinomial}(\boldsymbol{\mu})$
4. For each link $(d, d') \in \{1, \dots, D\}^2$
   (a) Draw link weight $A_{d,d'} \sim \text{Poisson}(\Omega_{y_d, y_{d'}})$

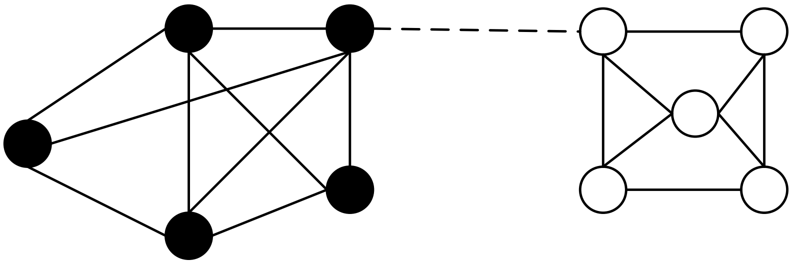


Figure 3.3: SCC (Sharir, 1981) can be distracted by spurious links connecting two groups, while WSBM (Aicher et al., 2014) maintains the distinction.

Unlike SCC that is vulnerable to noisy links, WSBM detects the blocks probabilistically and is more robust. As mentioned in the introduction of SCC (Section 3.1.1), given a graph like Figure 3.3, the existence of the dashed link will significantly change the output of SCC. If the dashed link does not exist, both WSBM and SCC can identify two blocks as denoted by colors. However, if the dashed link does exist, SCC will merge the two blocks and return only one big block that contains all nodes, which contradicts our intuition. In this case, WSBM is robust and still keeps the nodes in two reasonable blocks.

3.1.3 Relational Topic Model

Relational topic model (Chang and Blei, 2010, RTM) is a downstream model (Section 2.3.2.2) that jointly models the topics and document links (Figure 3.4). Although RTM can be described by the general generative process of topic models with posterior regularizers, we give its original generative process to reveal the intuitions better as follows:

1. For each topic $k \in \{1, \ldots, K\}$

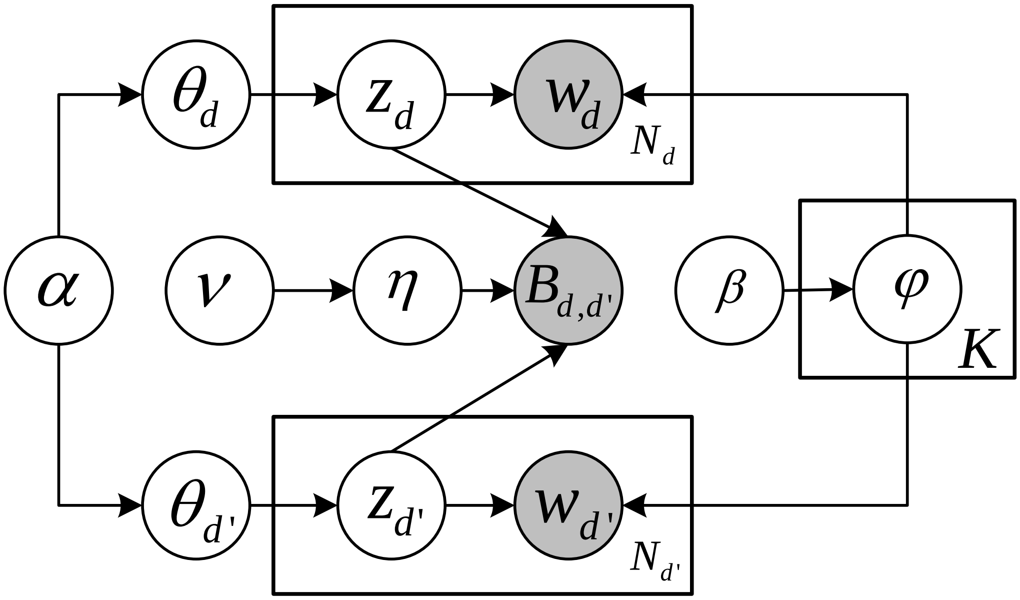


Figure 3.4: A two-document segment of relational topic model (Chang and Blei, 2010, RTM).

(a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dirichlet}(\beta)$

(b) Draw topic regression parameter $\eta_k \sim \mathcal{N}(0, \nu^2)$

2. For each document $d \in \{1, \dots, D\}$

(a) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dirichlet}(\alpha)$

(b) For each token $t_{d,n}$ in document d

i. Draw a topic $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta_d})$

ii. Draw a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{z_{d,n}}})$

3. For each explicit link (d, d')

(a) Draw link weight $B_{d,d'} \sim f(\mathbf{z_d}, \mathbf{z_{d'}}, \boldsymbol{\eta})$

where we use B to denote the document links because as we will introduce later (Section 3.2.3), the links fed to WSBM and topic model are different. Each link (d, d')

that connects documents d and d' is drawn from a link probability function $f(\mathbf{z_d}, \mathbf{z_{d'}}, \boldsymbol{\eta})$ that takes a weight vector and the two documents' topic posteriors:

$$B_{d,d'} \sim f(\mathbf{z_d}, \mathbf{z_{d'}}, \boldsymbol{\eta}) = \sigma\left(\sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}}\right), \tag{3.1}$$

where $\sigma(\cdot)$ is a sigmoid function.

As most downstream topic models, the posterior inference of RTM is based on stochastic EM and consists of an E-step and an M-step (Celeux, 1985). The E-step updates the topic assignments while holding the topic weight vectors:

$$\begin{aligned}&\Pr\left(z_{d,n} = k \,|\, \mathbf{z}^{-\mathbf{d,n}}, \mathbf{w}^{-\mathbf{d,n}}, w_{d,n} = v, \mathbf{B}, \boldsymbol{\eta}, \alpha, \beta\right)\\ &\propto \left(N_{d,k}^{-d,n} + \alpha\right) \frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta} \prod_{d':(d,d')\in\mathbf{B}} \sigma\left(\sum_{k'=1}^{K} \eta_{k'} \frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}} \frac{N_{d',k'}}{N_{d',\cdot}} + \eta_k \frac{1}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}}\right).\end{aligned} \tag{3.2}$$

The M-step optimizes the weight vector $\boldsymbol{\eta}$ to maximize the log-likelihood of generating the links with the topic assignments and Gaussian priors:

$$\mathcal{L}\left(\boldsymbol{\eta} \,|\, \mathbf{z}, \nu\right) = \sum_{(d,d')\in\mathbf{B}} \log\sigma\left(\sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}}\right) - \sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2}, \tag{3.3}$$

with L-BFGS and the partial derivative with respect to every η_k (Liu and Nocedal, 1989):

$$\frac{\partial\mathcal{L}\left(\boldsymbol{\eta} \,|\, \mathbf{z}, \nu\right)}{\partial\eta_k} = \sum_{(d,d')\in\mathbf{B}} \frac{\exp\left(-\sum_{k'=1}^{K} \eta_{k'} \frac{N_{d,k'}}{N_{d,\cdot}} \frac{N_{d',k'}}{N_{d',\cdot}}\right)}{1 + \exp\left(-\sum_{k'=1}^{K} \eta_{k'} \frac{N_{d,k'}}{N_{d,\cdot}} \frac{N_{d',k'}}{N_{d',\cdot}}\right)} \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}} - \frac{\eta_k}{\nu^2}. \tag{3.4}$$

3.2 Discriminative Topic Model with Block Prior and Features

Our model identifies latent document blocks from the document network with a WSBM, extracts topic patterns of each block as informative priors, and uses this

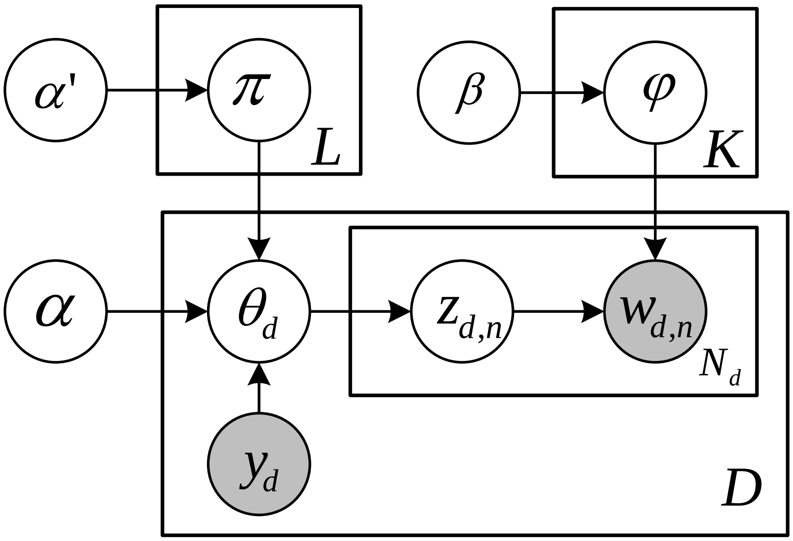


Figure 3.5: Graphical Model of BP-LDA.

information to infer topics and reconstruct the links. For presentation, we decompose it into several key components (Sections 3.2.1 and 3.2.2) and then aggregate (Section 3.2.3).

3.2.1 LDA with Block Priors (BP-LDA)

As argued at the beginning of this chapter, linked documents are likely to have similar topic distributions, which can be generalized to the documents in the same block. Inspired by this intuition and the block assignments we obtain in the previous sections, we want to extract some external knowledge from these blocks. Thus we introduce an LDA with block priors, hence BP-LDA, as shown in Figure 3.5, which has the following generative process:

1. For each topic $k \in \{1, \ldots, K\}$

   (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dirichlet}(\beta)$

2. For each block $l \in \{1, \dots, L\}$

   (a) Draw topic distribution $\boldsymbol{\pi_l} \sim \text{Dirichlet}(\alpha')$

3. For each document $d \in \{1, \dots, D\}$

   (a) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dirichlet}(\alpha \boldsymbol{\pi_{y_d}})$

   (b) For each token $t_{d,n}$ in document d

      i. Draw a topic $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta_d})$

      ii. Draw a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{z_{d,n}}})$

Most of BP-LDA's generative process is similar to vanilla LDA. However, unlike vanilla LDA, which uses an uninformative topic prior (i.e., same α value for all topics), BP-LDA puts a distinct Dirichlet prior $\boldsymbol{\pi}$ on each block to capture that block's topic distribution. Then BP-LDA uses the block's topic patterns as an informative prior (i.e., $\alpha \boldsymbol{\pi_{y_d}}$) which has emphases on some topics, when drawing each document's topic distribution in the block. In other words, a document's topic distribution—i.e., what the document is about—is not just informed by the words present in the document but also by the broader context of its network neighborhood.

3.2.2 More Features for Link Generation in RTM

Building on the relational topic model, we want to generate the links between documents based on more features we have (Chang and Blei, 2010). Specifically, in addition to topic distributions, documents' word distributions (Nguyen et al., 2013) and the link rate of the two documents' assigned blocks are also included

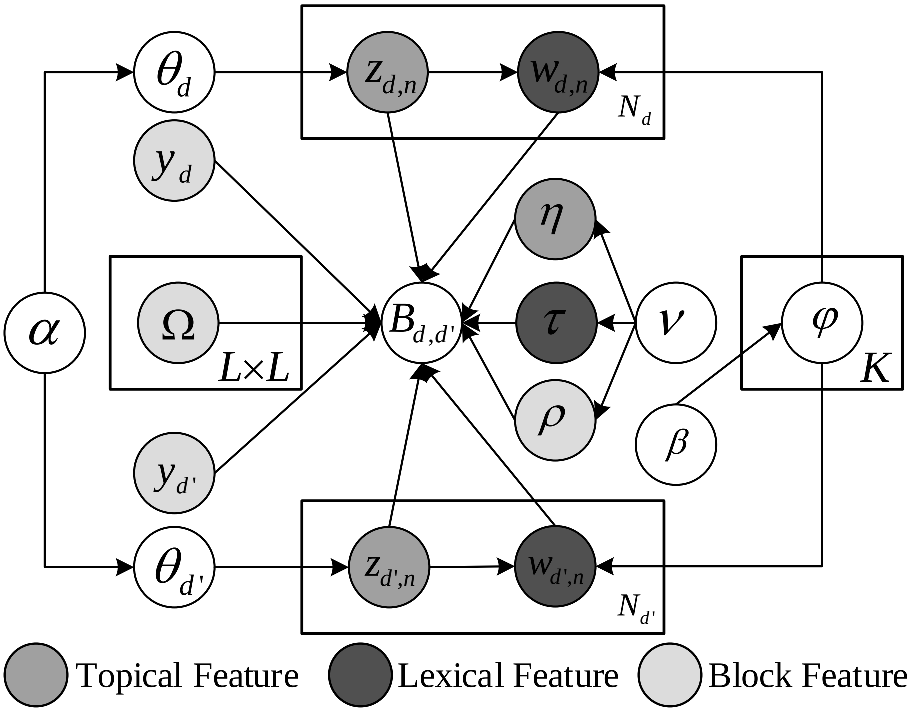


Figure 3.6: A two-document segment of RTM with features denoted by grayscale. The document link $B_{d,d'}$ is observed and should be in gray, but we keep it in white background to avoid confusion.

in the feature set, with the intuition that similar word usage and high inter-block link rate also indicate document similarity and the intent that these additional features improve link generation. RTM involves the relationship between a pair of documents, so it is difficult to show the whole model graphically; therefore Figure 3.6 illustrates with a two-document segment. The generative process is:

1. For each pair of blocks $(l, l') \in \{1, \ldots, L\}^2$

   (a) Draw block regression parameter $\rho_{l,l'} \sim \mathcal{N}(0, \nu^2)$

2. For each topic $k \in \{1, \ldots, K\}$

   (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dirichlet}(\beta)$

   (b) Draw topic regression parameter $\eta_k \sim \mathcal{N}(0, \nu^2)$

3. For each word $v \in \{1, \dots, V\}$

   (a) Draw lexical regression parameter $\tau_v \sim \mathcal{N}(0, \nu^2)$

4. For each document $d \in \{1, \dots, D\}$

   (a) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dirichlet}(\alpha)$

   (b) For each token $t_{d,n}$ in document d

      i. Draw a topic $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta_d})$

      ii. Draw a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{z_{d,n}}})$

5. For each explicit link (d, d')

   (a) Draw link weight $B_{d,d'} \sim f\left(y_d, y_{d'}, \boldsymbol{\Omega}, \mathbf{z_d}, \mathbf{z_{d'}}, \mathbf{w_d}, \mathbf{w_{d'}}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)$

Binary links are generated by a link probability function f which takes the regression value $R_{d,d'}$ of documents d and d' as an argument. Assuming documents d and d' belong to blocks l and l' respectively, $R_{d,d'}$ is

$$R_{d,d'} = \sum_{k=1}^{K} \eta_k \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}} + \sum_{v=1}^{V} \tau_v \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d',v}}{N_{d',\cdot}} + \rho_{l,l'} \Omega_{l,l'}, \tag{3.5}$$

where as Chang and Blei (2010), the two documents' topic and word distribution similarities are captured by the weighted sum of element-wise (Hardamard) product; $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ are the weight vectors and matrix for topic-based, lexical-based and rate-based predictions, respectively.

A common choice of the link probability function f is a sigmoid (Chang and Blei, 2010):

$$f(R_{d,d'}) = \Pr\left(B_{d,d'} = 1 \,|\, R_{d,d'}\right) = \sigma\left(R_{d,d'}\right) = \frac{1}{1 + \exp\left(-R_{d,d'}\right)}. \tag{3.6}$$

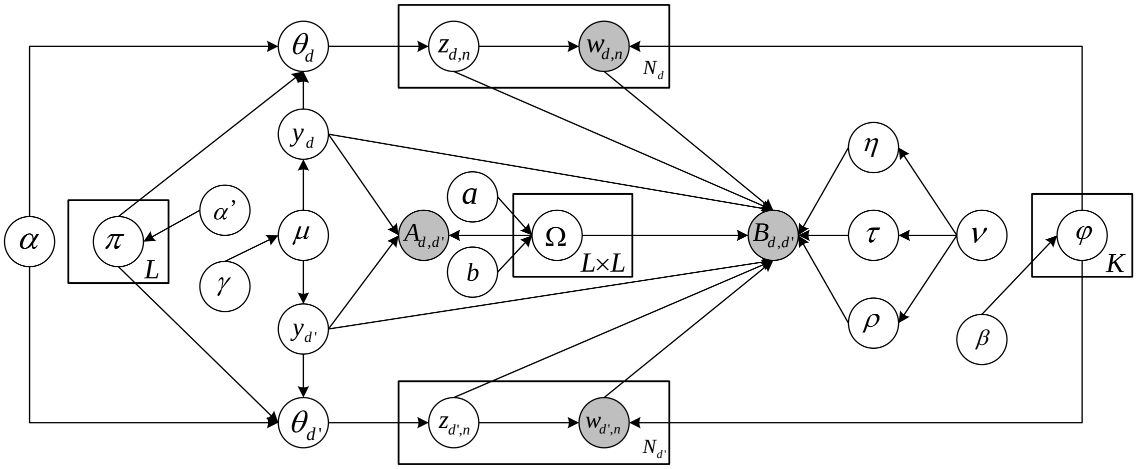


Figure 3.7: The graphical model of LBH-RTM for two documents, in which a weighted stochastic block model is integrated (γ, $\boldsymbol{\mu}$, $\mathbf{y}$, a, b, $\boldsymbol{\Omega}$, and $\mathbf{A}$) to identify latent document blocks. Each document's topic distribution has an informative prior $\boldsymbol{\pi}$ extracted from the block topic distributions. The model predicts links between documents ($\mathbf{B}$) based on topics ($\mathbf{z}$), words ($\mathbf{w}$), and inter-block link rates ($\boldsymbol{\Omega}$), using a max-margin objective.

However, we instead use hinge loss so that RTM can use the max-margin principle, making more effective use of side information when inferring topic assignments (Zhu et al., 2012). Using hinge loss, the probability that documents d and d' are linked is

$$\Pr\left(B_{d,d'} \mid R_{d,d'}\right) = \exp\left(-2\max(0, \zeta_{d,d'})\right), \tag{3.7}$$

where $\zeta_{d,d'} = 1 - B_{d,d'}R_{d,d'}$. Positive and negative link weights are denoted by 1 and -1, respectively, in contrast to sigmoid loss which denotes negative link weights by 0 instead.

3.2.3 Aggregated Model

Finally, we put all the pieces together and introduce LBH-RTM: RTM with lexical weights (L), block priors (B), and hinge loss (H). Its graphical model is given in Figure 3.7.

1. For each pair of blocks $(l, l') \in \{1, \ldots, L\}^2$
    (a) Draw inter-block link rate $\Omega_{l,l'} \sim \text{Gamma}(a, b)$
    (b) Draw block regression parameter $\rho_{l,l'} \sim \mathcal{N}(0, \nu^2)$
2. Draw block distribution $\boldsymbol{\mu} \sim \text{Dirichlet}(\gamma)$
3. For each block $l \in \{1, \ldots, L\}$
    (a) Draw topic distribution $\boldsymbol{\pi_l} \sim \text{Dirichlet}(\alpha')$
4. For each topic $k \in \{1, \ldots, K\}$
    (a) Draw word distribution $\boldsymbol{\phi_k} \sim \text{Dirichlet}(\beta)$
    (b) Draw topic regression parameter $\eta_k \sim \mathcal{N}(0, \nu^2)$
5. For each word $v \in \{1, \ldots, V\}$
    (a) Draw lexical regression parameter $\tau_v \sim \mathcal{N}(0, \nu^2)$
6. For each document $d \in \{1, \ldots, D\}$
    (a) Draw a block $y_d \sim \text{Multinomial}(\boldsymbol{\mu})$
    (b) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dirichlet}(\alpha \boldsymbol{\pi_{y_d}})$
    (c) For each token $t_{d,n}$ in document d
        i. Draw a topic $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta_d})$
        ii. Draw a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{z_{d,n}}})$
7. For each link $(d, d') \in \{1, \ldots, D\}^2$

(a) Draw link weight $A_{d,d'} \sim \text{Poisson}(\Omega_{y_d,y_{d'}})$

8. For each explicit link (d, d')

(a) Draw link weight $B_{d,d'} \sim f(y_d, y_{d'}, \boldsymbol{\Omega}, \mathbf{z_d}, \mathbf{z_{d'}}, \mathbf{w_d}, \mathbf{w_{d'}}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho})$

where the link sets **A** (for block detection) and **B** (for document link replication) are assumed independent in the model, but they can be derived from the same set of links in practice.

Link set **A** is primarily used to find blocks, so it treats all links *deterministically*. In other words, the links observed in the input are considered explicit positive links, while the unobserved links are considered explicit negative links, in contrast to the implicit links in **B**.

In terms of link set **B**, while it adopts all explicit positive links from the input, it *does not deny* the existence of unobserved links, or implicit negative links, because sometimes it makes sense for a link to exist between two unlinked documents, e.g., a good but missing citation for a paper.[3] Thus **B** consists of only explicit positive links. However, to avoid overfitting, we randomly sample some implicit links and add them to **B** as explicit negative links (Gutmann and Hyvärinen, 2012; Mnih and Teh, 2012; Collobert and Weston, 2008).

The general workflow of LBH-RTM is as follows. WSBM detects the latent blocks and documents' block assignments. Then topic priors are extracted from blocks and guide documents' topic generation. Finally, document links are drawn from a max-margin probability function, with topical, lexical, and block features.

[3] This indicates a potential application of document link suggestion for our LBH-RTM.

Algorithm 1 Sampling Process of LBH-RTM

```
1: Sample implicit negative links as explicit ones from a uniform distribution
2: Set every $\lambda_{d,d'} = 1$ and initialize every topic assignment $z_{d,n}$ from a uniform distribution
3: for $m = 1$ to $M$ do
4:     Optimize $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ using L-BFGS (Equation 3.18)
5:     for each document $d = 1$ to $D$ do
6:         Draw block assignment $y_d$ from the multinomial distribution (Equation 3.8)
7:         for each token $n$ in document $d$ do
8:             Draw a topic assignment $z_{d,n}$ from the multinomial distribution (Equation 3.12)
9:         end for
10:        for each document $d'$ which document $d$ explicitly links do
11:            Draw $\lambda_{d,d'}^{-1}$ (and then $\lambda_{d,d'}$) from the inverse Gaussian distribution (Equation 3.26)
12:        end for
13:    end for
14: end for
```

3.3 Posterior Inference

Like other downstream topic models, the posterior inference of LBH-RTM (Algorithm 1) is based on stochastic EM and consists of an E-step of updating topic and block assignments and an M-step of optimizing the weight vectors and matrix (Celeux, 1985).[4] We add an auxiliary variable $\boldsymbol{\lambda}$ for hinge loss (see Section 3.3.2), which is included as part of the E-step. The updating of $\boldsymbol{\lambda}$ is not necessary when using sigmoid loss.

The sampling procedure is an iterative process after initialization (Lines 1 and 2). In each of the M iterations, we first optimize the weight vectors and matrix (Line 4) before updating documents' block assignments (Line 6) and topic assign-

[4] More details about sampling procedures and equations in this chapter, including the sampling and optimization equations using sigmoid loss, are available in Appendix A.

ments (Line 8). When using hinge loss, the auxiliary variable $\boldsymbol{\lambda}$ for every explicit link needs to be updated (Line 11).

3.3.1 Sampling Block Assignments

Block assignment sampling is done by Gibbs sampling, using the block assignments and link statistics based on the link set $\mathbf{A}$, but excluding document d and its related links.[5] The probability that document d is assigned to block l is

$$\Pr\left(y_d = l \,|\, \mathbf{A}^{-\mathbf{d}}, \mathbf{y}^{-\mathbf{d}}, a, b, \gamma\right) \propto \left(N_l^{-d} + \gamma\right) \times$$
$$\prod_{l'} \frac{\left(S_e^{-d}(l,l') + b\right)^{S_w^{-d}(l,l')+a}}{\left(S_e^{-d}(l,l') + b + S_e(d,l')\right)^{S_w^{-d}(l,l')+a+S_w(d,l')}} \prod_{i=0}^{S_w(d,l')-1} \left(S_w^{-d}(l,l') + a + i\right), \quad (3.8)$$

where N_l is the number of documents assigned to block l; $^{-d}$ denotes that the count excludes document d; $S_w(d,l)$ and $S_w(l,l')$ are the sums of link weights from document d to block l and from block l to block l', respectively:

$$S_w(d,l) = \sum_{d':y_{d'}=l} A_{d,d'} \quad (3.9)$$
$$S_w(l,l') = \sum_{d:y_d=l} S_w(d,l'). \quad (3.10)$$

$S_e(d,l)$ is the maximum number of possible links from document d to l, i.e., assuming document d connects to every document in block l, which equals N_l. The maximum number of possible links from block l to l' is $S_e(l,l')$, i.e., assuming every document in block l connects to every document in block l':

$$S_e(l,l') = \begin{cases} N_l \times N_{l'} & l \neq l' \\ \frac{1}{2} N_l (N_l - 1) & l = l'. \end{cases} \quad (3.11)$$

[5] These equations deal with undirected edges, but they can be adapted for directed edges. See Appendix Section A.1.2.

The time complexity of inferring document d's block assignment is $\sum_l S_w(d,l)$ which is the total link weight from/to document d. In our case, we are dealing with binary-valued links, so the total link weight from/to document d equals the *degree* of document d. In the implementation, the values of the power and product terms in Equation 3.8 may exceed the range of (double-precision) float numbers, so it is suggested to compute the logarithmic scores for each block, apply normalization, and finally sample a block.

If we rearrange the terms of Equation 3.8 and put the terms which have $S_w(d,l')$ together, we will find that WSBM considers the document's link density when updating its block assignment: when the value of $S_w(d,l')$ increases, or document d is more densely connected with the documents in block l', the probability of assigning document d to block l decreases exponentially. Thus if document d is more densely connected with the documents in block l and sparsely connected with other blocks, it is (exponentially) more likely to be assigned to block l.

3.3.2 Sampling Topic Assignments

Following Polson and Scott (2011), we introduce an auxiliary variable $\lambda_{d,d'}$ for updating topic assignments when using hinge loss. With $\lambda_{d,d'}$, the conditional probability of assigning $t_{d,n}$, the n-th token in document d, to topic k is

$$\begin{aligned}&\Pr\left(z_{d,n}=k \,|\, \mathbf{z}^{-\mathbf{d},\mathbf{n}}, \mathbf{w}^{-\mathbf{d},\mathbf{n}}, w_{d,n}=v, y_d=l, \boldsymbol{\pi}, \alpha, \alpha', \beta\right)\\ &\propto \left(N_{d,k}^{-d,n}+\alpha\pi_{l,k}^{-d,n}\right)\frac{N_{k,v}^{-d,n}+\beta}{N_{k,\cdot}^{-d,n}+V\beta}\prod_{d':(d,d')\in\mathbf{B}}\exp\left(-\frac{\left(\zeta_{d,d'}+\lambda_{d,d'}\right)^2}{2\lambda_{d,d'}}\right),\end{aligned} \tag{3.12}$$

where $N_{d,k}$ is the number of tokens in document d that are assigned to topic k; $N_{k,v}$ denotes the count of word v assigned to topic k; Marginal counts are denoted by $\cdot$; ${}^{-d,n}$ denotes that the count excludes $t_{d,n}$; d' denotes all documents that have explicit links with document d. The block topic prior $\pi_{l,k}^{-d,n}$ is estimated based on the maximal path assumption (Cowans, 2006; Wallach, 2008):

$$\pi_{l,k}^{-d,n} = \frac{\sum_{d':y_{d'}=l} N_{d',k}^{-d,n} + \alpha'}{\sum_{d':y_{d'}=l} N_{d',\cdot}^{-d,n} + K\alpha'}. \tag{3.13}$$

The link prediction argument $\zeta_{d,d'}$ is

$$\zeta_{d,d'} = 1 - B_{d,d'} \left(R_{d,d'}^{-d,n} + \eta_k \frac{1}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}} \right), \tag{3.14}$$

where

$$R_{d,d'}^{-d,n} = \sum_{k'=1}^{K} \eta_{k'} \frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}} \frac{N_{d',k'}}{N_{d',\cdot}} + \sum_{v=1}^{V} \tau_v \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d',v}}{N_{d',\cdot}} + \rho_{y_d,y_{d'}} \Omega_{y_d,y_{d'}}. \tag{3.15}$$

Looking at the first term of Equation 3.12, the probability of assigning $t_{d,n}$ to topic k depends not only on its own document topic distribution, but also the topic distribution of the block it belongs to, which reflects the theory of homophily. The links also matter: Equation 3.14 gives us the intuition that a topic is more likely to be selected if it could increase the likelihood of links, which forms an interaction between topics and the link graph—the links are guiding the topic sampling while updating topic assignments is maximizing the likelihood of the link graph.

The time complexity of inferring a token's topic assignment in document d is $O(\mathrm{Deg}(d)K(K+V))$ where Deg(d) denotes the degree of document d; the term $(K+V)$ comes from the calculation of $R_{d,d'}^{-d,n}$. However, the documents' word distributions are often sparse, so we can skip the words with zero term frequencies in documents d or d' and achieve a much better complexity than $O(K+V)$.

3.3.3 Parameter Optimization

While topic assignments are updated iteratively in the E-step, the weight vectors and matrix $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ are optimized in the M-step of each global iteration over the whole corpus using L-BFGS (Liu and Nocedal, 1989). It takes the likelihood of generating the link set $\mathbf{B}$ using $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, $\boldsymbol{\rho}$, and current topic and block assignments as the objective function, and optimizes it using the partial derivatives with respect to every weight vector/matrix element.

The log likelihood of generating link set $\mathbf{B}$ using $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, $\boldsymbol{\rho}$, and hinge loss, i.e., the sum of the exponents in Equation 3.12, is

$$\mathcal{L}\left(\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\rho}\,|\,\mathbf{B},\mathbf{z},\mathbf{w},\mathbf{y}\right) \tag{3.16}$$

$$= -\sum_{d,d'}\frac{\left(\zeta_{d,d'}+\lambda_{d,d'}\right)^2}{2\lambda_{d,d'}} - \sum_{k=1}^{K}\frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V}\frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L}\frac{\rho_{l,l'}^2}{2\nu^2} \tag{3.17}$$

$$\propto -\sum_{d,d'}\frac{R_{d,d'}^2 - 2\left(1+\lambda_{d,d'}\right)B_{d,d'}R_{d,d'}}{2\lambda_{d,d'}} - \sum_{k=1}^{K}\frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V}\frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L}\frac{\rho_{l,l'}^2}{2\nu^2}. \tag{3.18}$$

Thus the partial derivatives are

$$\frac{\partial\mathcal{L}\left(\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\rho}\right)}{\partial\eta_k} \propto -\sum_{d,d'}\frac{R_{d,d'}-\left(1+\lambda_{d,d'}\right)B_{d,d'}}{\lambda_{d,d'}}\frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} - \frac{\eta_k}{\nu^2} \tag{3.19}$$

$$\frac{\partial\mathcal{L}\left(\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\rho}\right)}{\partial\tau_v} \propto -\sum_{d,d'}\frac{R_{d,d'}-\left(1+\lambda_{d,d'}\right)B_{d,d'}}{\lambda_{d,d'}}\frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} - \frac{\tau_v}{\nu^2} \tag{3.20}$$

$$\frac{\partial\mathcal{L}\left(\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\rho}\right)}{\partial\rho_{l,l'}} \propto -\sum_{d\in l,d'\in l'}\frac{R_{d,d'}-\left(1+\lambda_{d,d'}\right)B_{d,d'}}{\lambda_{d,d'}}\Omega_{l,l'} - \frac{\rho_{l,l'}}{\nu^2}. \tag{3.21}$$

We also need to update the auxiliary variable $\lambda_{d,d'}$. Since the likelihood of $\lambda_{d,d'}$ follows a generalized inverse Gaussian distribution (Barndorff-Nielsen and Halgreen, 1977; Seshadri, 1997)

$$\begin{aligned}\mathcal{L}\left(\lambda_{d,d'}\,|\,\mathbf{z},\mathbf{w},\mathbf{y},\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\rho}\right)=&\frac{1}{\sqrt{2\pi\lambda_{d,d'}}}\exp\left(-\frac{\left(\zeta_{d,d'}+\lambda_{d,d'}\right)^2}{2\lambda_{d,d'}}\right) &(3.22)\\ \propto&\frac{1}{\sqrt{2\pi\lambda_{d,d'}}}\exp\left(-\frac{\zeta_{d,d'}^2}{2\lambda_{d,d'}}-\frac{\lambda_{d,d'}}{2}\right) &(3.23)\\ =&\mathcal{GIG}\left(\lambda_{d,d'};\frac{1}{2},1,\zeta_{d,d'}^2\right), &(3.24)\end{aligned}$$

where

$$\mathcal{GIG}\left(x;p,a,b\right)=C(p,a,b)x^{p-1}\exp\left(-\frac{1}{2}\left(\frac{b}{x}+ax\right)\right), \tag{3.25}$$

where $C(p,a,b)$ is a normalization constant, so we sample its reciprocal $\lambda_{d,d'}^{-1}$ from an inverse Gaussian distribution and then obtain $\lambda_{d,d'}$ (Chhikara, 1988):

$$\Pr\left(\lambda_{d,d'}^{-1}\,|\,\mathbf{z},\mathbf{w},\mathbf{y},\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\rho}\right)=\mathcal{IG}\left(\lambda_{d,d'}^{-1};\frac{1}{|\zeta_{d,d'}|},1\right), \tag{3.26}$$

where

$$\mathcal{IG}\left(x;a,b\right)=\sqrt{\frac{b}{2\pi x^3}}\exp\left(-\frac{b\left(x-a\right)^2}{2a^2x}\right) \tag{3.27}$$

for $a>0$ and $b>0$.

3.4 Experimental Results

We evaluate our LBH-RTM using two datasets. The first is Cora (McCallum et al., 2000). After removing stopwords and the words that appear in fewer than ten documents, as well as the documents with no words or links, our vocabulary has 1,240 unique word types. The corpus has 2,362 computer science paper abstracts with 4,231 citation links.

The second dataset is WebKB. It is already preprocessed and has 1,703 unique word types in vocabulary. The corpus has 877 web pages with 1,608 hyperlinks.

We treat all links as undirected. Both datasets are split into five folds, each further split into development and test sets with approximately the same size when used for evaluation.

We first introduce LBH-RTM's link prediction performance (Section 3.4.1) and show the model's superiority over others with an illustrative example (Section 3.4.2). Then we evaluate the model's topic coherence both quantitatively and qualitatively (Section 3.4.3). We finally illustrate the robustness of WSBM over SCC with an example (Section 3.4.4).

3.4.1 Link Prediction Results

We evaluate LBH-RTM and its variations on link prediction tasks using *predictive link rank* (PLR) against baseline models. A document's PLR is the average rank of the documents to which it has explicit positive links, among all documents, so lower PLR indicates better link prediction performance, as actually linked documents are ranked higher. For instance, given a query document, we rank and sort all other documents by the link probabilities to the query document as shown in Table 3.1. Among the six candidate documents, documents 2, 3, and 6 have actual links with the query document and their ranks are 2, 3, and 5. Thus the average rank is $(2 + 3 + 5)/3 \approx 3.33$.

Following the experiment setup in Chang and Blei (2010), we train the models

Rank	Doc ID	Link Probability	Actual Link?
1	5	0.90	
2	3	0.85	Yes
3	2	0.82	Yes
4	4	0.70	
5	6	0.63	Yes
6	1	0.50	

Table 3.1: The predictive link rank (PLR) of a document d is the average rank of actually linked documents with d.

on the training set and predict citation links/hyperlinks within held-out documents as well as from held-out documents to training documents. We tune two important parameters—α and negative edge ratio. α controls the strength of the informative prior, while negative edge ratio controls the size of the randomly sampled negative links—it is the ratio of the number of randomly sampled negative links to the number of explicit positive links. These parameters are tuned on the development set and we then apply the trained model which performs the best on the development set to the test set. We also tune the number of blocks for the WSBM and set it to 35 and 20 for Cora and WebKB respectively. The block topic priors $\boldsymbol{\pi}$ are not applied on unseen documents, since we don't have available links.

The cross-validation results are given in Table 3.2, where models are differently equipped with lexical weights (L), WSBM prior (B) versus SCC prior (C), hinge loss (H) versus sigmoid loss (S).[6] Link prediction performance generally improves with incremental application of external knowledge (WSBM prior (BS-RTM) and lexical weights (LBS-RTM)) and more sophisticated learning techniques (hinge loss (LBH-

[6] The values of RTM are different from the result reported by Chang and Blei (2010), because we re-preprocessed the Cora dataset and used different parameters.

Model	Cora	WebKB
RTM (Chang and Blei, 2010)	419.33	141.65
LCH-RTM (Yang et al., 2015a)	459.55	150.32
BS-RTM	391.88	127.25
LBS-RTM	383.25	125.41
LBH-RTM	**360.38**	**111.79**

Table 3.2: Predictive link rank results. The performance improves over RTM when we incrementally add WSBM prior (BS-RTM), lexical weights (LBS-RTM), and hinge loss (LBH-RTM).

RTM)).

The WSBM brings around 6.5% and 10.2% improvement over RTM in PLR on the Cora and WebKB datasets, respectively. This indicates that the latent blocks identified by WSBM are reasonable and consistent with reality. The lexical weights also help link prediction (LBS-RTM), though less for BS-RTM. This is understandable since word distributions are much sparser and do not make as significant a contribution as topic distributions. Finally, hinge loss improves PLR substantially (LBH-RTM), about 14.1% and 21.1% improvement over RTM on the Cora and WebKB datasets respectively, demonstrating the effectiveness of max-margin learning.

The only difference between LCH-RTM and LBH-RTM is the block detection algorithm, i.e., SCC vs. WSBM. However, their link prediction performance is poles apart—LCH-RTM even fails to outperform RTM. This implies that the quality of blocks identified by SCC is not as good as WSBM, which we also illustrate in Section 3.4.4.

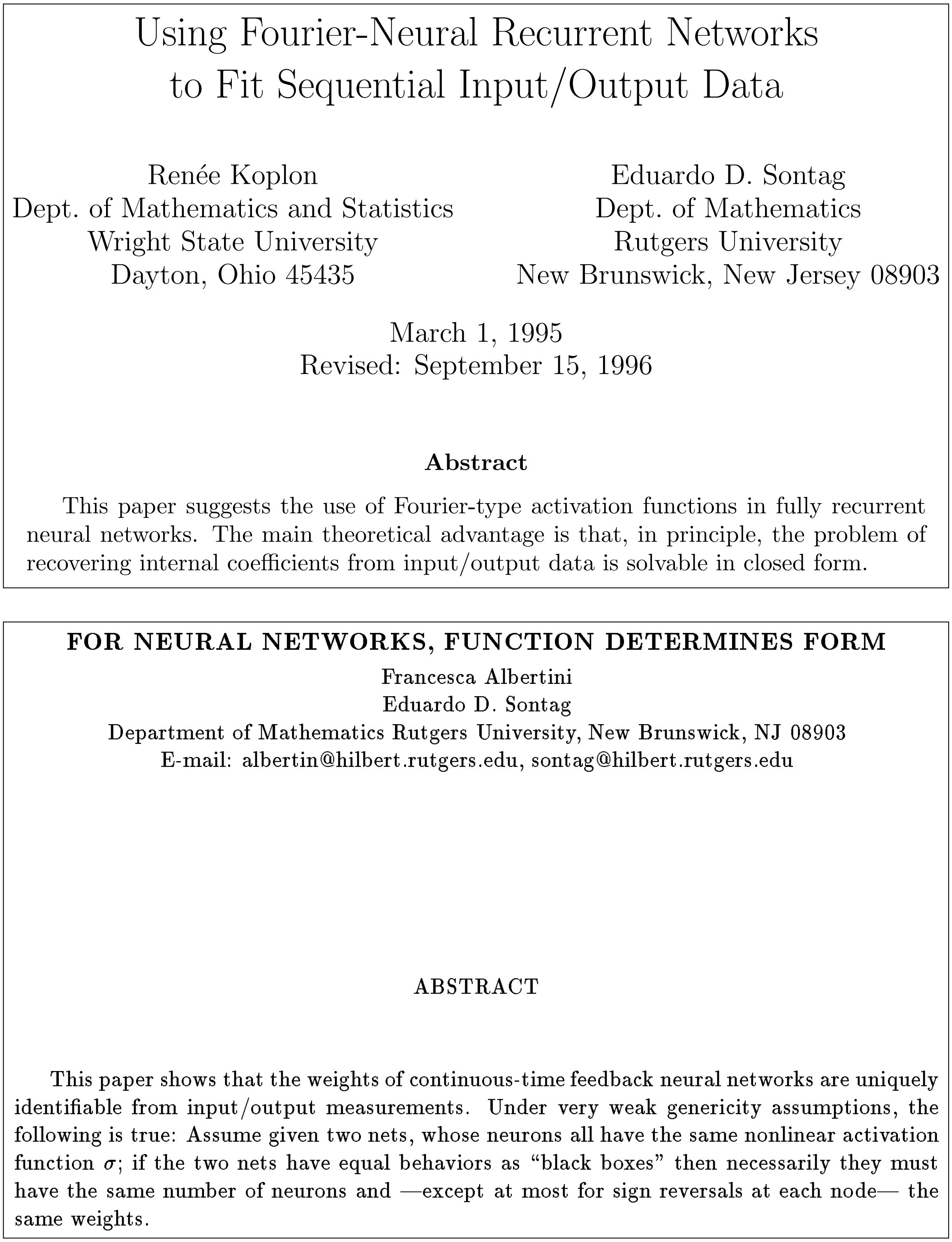
3.4.2 Illustrative Example

We illustrate our model's behavior qualitatively by looking at two abstracts, Koplon and Sontag (1997) and Albertini and Sontag (1992) from the Cora dataset, designated K and A for short.

Paper A shows that two neural networks must have the same number of neurons and the same weights (except sign reversals) if they use the same activation function and have equal input/output behaviors as "black boxes" (Figure 3.8). Paper K studies Fourier-type activation function in recurrent neural networks and its solvability from input/output data. Thus we can easily find that both of them are about the topic of Neural Network. Looking at the words, they both contain words like "neural", "networks", "activation", and "function", which corresponds to the inferred Neural Network topic with words "neural", "network", "train", "learn", "function", "recurrent", etc.

As a ground-truth, there is a citation between K and A. The ranking of this link improves as the model gets more sophisticated (Table 3.3), except LCH-RTM, which is consistent with our PLR results.

In Figure 3.9, we also show the proportions of topics that dominate the two documents according to the various models. Multiple topics are dominating K and A according to RTM (Figure 3.9(a)). As the model gets more sophisticated, the Neural Network topic proportion gets higher (Figures 3.9(c) and 3.9(d)). Finally, only the Neural Network topic dominates the two documents when LBH-RTM is applied (Figure 3.9(e)).



Using Fourier-Neural Recurrent Networks
to Fit Sequential Input/Output Data

Renée Koplon
Dept. of Mathematics and Statistics
Wright State University
Dayton, Ohio 45435

Eduardo D. Sontag
Dept. of Mathematics
Rutgers University
New Brunswick, New Jersey 08903

March 1, 1995
Revised: September 15, 1996

Abstract

This paper suggests the use of Fourier-type activation functions in fully recurrent neural networks. The main theoretical advantage is that, in principle, the problem of recovering internal coefficients from input/output data is solvable in closed form.

FOR NEURAL NETWORKS, FUNCTION DETERMINES FORM

Francesca Albertini
Eduardo D. Sontag
Department of Mathematics Rutgers University, New Brunswick, NJ 08903
E-mail: albertin@hilbert.rutgers.edu, sontag@hilbert.rutgers.edu

ABSTRACT

This paper shows that the weights of continuous-time feedback neural networks are uniquely identifiable from input/output measurements. Under very weak genericity assumptions, the following is true: Assume given two nets, whose neurons all have the same nonlinear activation function σ; if the two nets have equal behaviors as "black boxes" then necessarily they must have the same number of neurons and —except at most for sign reversals at each node— the same weights.

Figure 3.8: Titles and abstracts of Neural Network Papers K (Koplon and Sontag, 1997, upper) and A (Albertini and Sontag, 1992, lower). Paper K cites Paper A.

Model	Link Rank
RTM	1,265
LCH-RTM	1,385
BS-RTM	635
LBS-RTM	132
LBH-RTM	**106**

Table 3.3: PLR of the citation link between example documents K (Koplon and Sontag, 1997) and A (Albertini and Sontag, 1992) (described in Section 3.4.2)

LCH-RTM gives the highest proportions to the Neural Network topic (Figure 3.9(b)). However, the Neural Network topic is split into two topics, and the proportions are not assigned to the same topic, which dramatically brings down the link prediction performance as it is based on the weighted sum of element-wise product. The splitting of the Neural Network topic also happens in RTM (Figures 3.9(a)) and LBS-RTM (Figure 3.9(d)), but they assign proportions to the same topic(s). Further comparing with LBH-RTM, the blocks detected by SCC are not improving the modeling of topics and links—some documents that should be in two different blocks are assigned to the same one, which generates a confusing block prior, as we will show in Section 3.4.4.

3.4.3 Topic Quality Results

We use the automatic coherence detection method to evaluate the topic quality (Lau et al., 2014, Section 2.2). Specifically, for each topic, we pick out the top N words and compute the average association score of each pair of words, based on the held-out documents in development and test sets.

We choose $N = 25$ and use Fisher's exact test (Upton, 1992, FET) and log

Model	FET		LLR	
	Cora	WebKB	Cora	WebKB
RTM	0.1330	0.1312	3.001	6.055
LCH-RTM	0.1418	0.1678	3.071	6.577
BS-RTM	0.1415	0.1950	3.033	6.418
LBS-RTM	0.1342	0.1963	2.984	6.212
LBH-RTM	**0.1453**	**0.2628**	**3.105**	**6.669**

Table 3.4: Topic coherence of models on Cora and WebKB, evaluated by Fisher's exact test (Upton, 1992, FET) and log-likelihood ratio (Moore, 2004; Dunning, 1993, LLR). WSBM priors and hinge loss benefit the topic coherence, while lexical weights hurt a little bit.

likelihood ratio (Moore, 2004; Dunning, 1993, LLR) as the association measures (Table 3.4). The main advantage of these measures is that they are robust even when the reference corpus is not large.

Coherence improves with WSBM and max-margin learning, but drops a little when adding lexical weights except for the FET score on the WebKB dataset, because lexical weights are intended to improve link prediction performance, not topic quality. Topic quality of LBH-RTM is also better than that of LCH-RTM, suggesting that WSBM benefits topic quality more than SCC.

Table 3.5 gives the top ten words in three topics across models. RTM yields topics with more words with general meanings, such as "algorithm", "method", "model", "paper", and "system". Adding WSBM block priors (BS-RTM) mostly lowers the weight of the general words and adds more weight on the words with specific meanings to the topics, e.g., "markov" and "chain" for Markov Chain topic, "visual", "recognit", "imag", and "neural" for DL for CV topic, and "parallel", "execut", "instruct", and "schedul" for Parallel Execution topic. Lexical weights (LBS-RTM) do not necessarily improve the topic coherence. It sometimes gives

Topic	Model	Words
Markov Chain	RTM	*algorithm*, distribut, **markov**, state, converg, **chain**, *method*, sampl, *model*, approxim
	LCH-RTM	estim, distribut, *model*, *method*, sampl, *algorithm*, **chain**, bayesian, **markov**, data
	BS-RTM	distribut, *algorithm*, converg, *method*, bayesian, estim, **chain**, **markov**, sampl, approxim
	LBS-RTM	*model*, distribut, estim, **markov**, *method*, bayesian, sampl, **chain**, function, prior
	LBH-RTM	**chain**, **markov**, distribut, converg, *algorithm*, sampl, *method*, state, sampler, estim
Deep Learning for Computer Vision (DL for CV)	RTM	*model*, object, **visual**, pattern, **recognit**, **imag**, represent, *system*, network, connect
	LCH-RTM	network, *model*, *learn*, **neural**, **visual**, object, pattern, represent, input, structur
	BS-RTM	*model*, object, pattern, **visual**, process, represent, **imag**, **neuron**, dynam, *system*
	LBS-RTM	*model*, network, pattern, **visual**, represent, object, input, **recognit**, **neural**, **neuron**
	LBH-RTM	*model*, object, **visual**, network, **neural**, **imag**, face, **recognit**, **neuron**, human
Parallel Execution	RTM	**parallel**, perform, machin, **execut**, *paper*, processor, *approach*, instruct, implement, *result*
	LCH-RTM	*network*, *learn*, *neural*, *model*, *system*, **parallel**, adapt, *algorithm*, *paper*, gener
	BS-RTM	**parallel**, **execut**, **instruct**, processor, perform, machin, architectur, program, *paper*, *system*
	LBS-RTM	**parallel**, **execut**, processor, perform, **instruct**, machin, **schedul**, implement, *paper*, *present*
	LBH-RTM	**parallel**, perform, **execut**, processor, **instruct** implement, control, **schedul**, **branch**, predict

Table 3.5: Three topics' top ten words given by various models. Words with general meanings are in *red and italic*. Words with specific meanings are in **blue and bold**. Generally, LBH-RTM assigns higher weights to specific words and lower weights to general words.

high weights to general words (Markov Chain topic) and low weights to specific words (Parallel Execution topic). After adding hinge loss (LBH-RTM), the topic quality is the best—it has more specific words with high weights and fewer or even no general words. SCC prior (LCH-RTM), however, sometimes brings the topic to

Block	#Nodes	#Links in the Block	#Links across Blocks
1	42	55	2
2	84	142	

Table 3.6: Statistics of Blocks 1 (Learning Theory) and 2 (Bayesian Nets) which are of different topics but linked by two edges. SCC merges the two blocks, while WSBM is robust and identifies two.

a wrong direction. For instance, in the topic of Parallel Execution, its top words are "network", "learn", "neural", and "model", which obviously should not be in this topic. This is probably due to its vulnerability to sparse links across blocks, as we will discuss in the next section.

3.4.4 Block Analysis

We illustrate the effectiveness of the WSBM over SCC.[7] As we have argued, WSBM can separate two internally densely-connected blocks even if few links are connecting them, while SCC tends to merge them in this case.

As an example, we focus on two blocks in the Cora dataset identified by WSBM, designated Blocks 1 and 2. Some statistics are given in Table 3.6. The two blocks are very sparsely connected, but comparatively quite densely connected inside either block. The two blocks' topic distributions also reveal their differences: abstracts in Block 1 mainly focus on Learning Theory ("learn", "algorithm", "bound", "result", etc.) and MCMC ("markov", "chain", "distribution", "converge", etc.). Abstracts in Block 2, however, have higher weights on Bayesian Networks ("net-

[7] We omit the comparison of WSBM with other models, because this has been done by Aicher et al. (2014). In addition, WSBM is a probabilistic method while SCC is deterministic. They are not comparable quantitatively, so we compare them qualitatively.

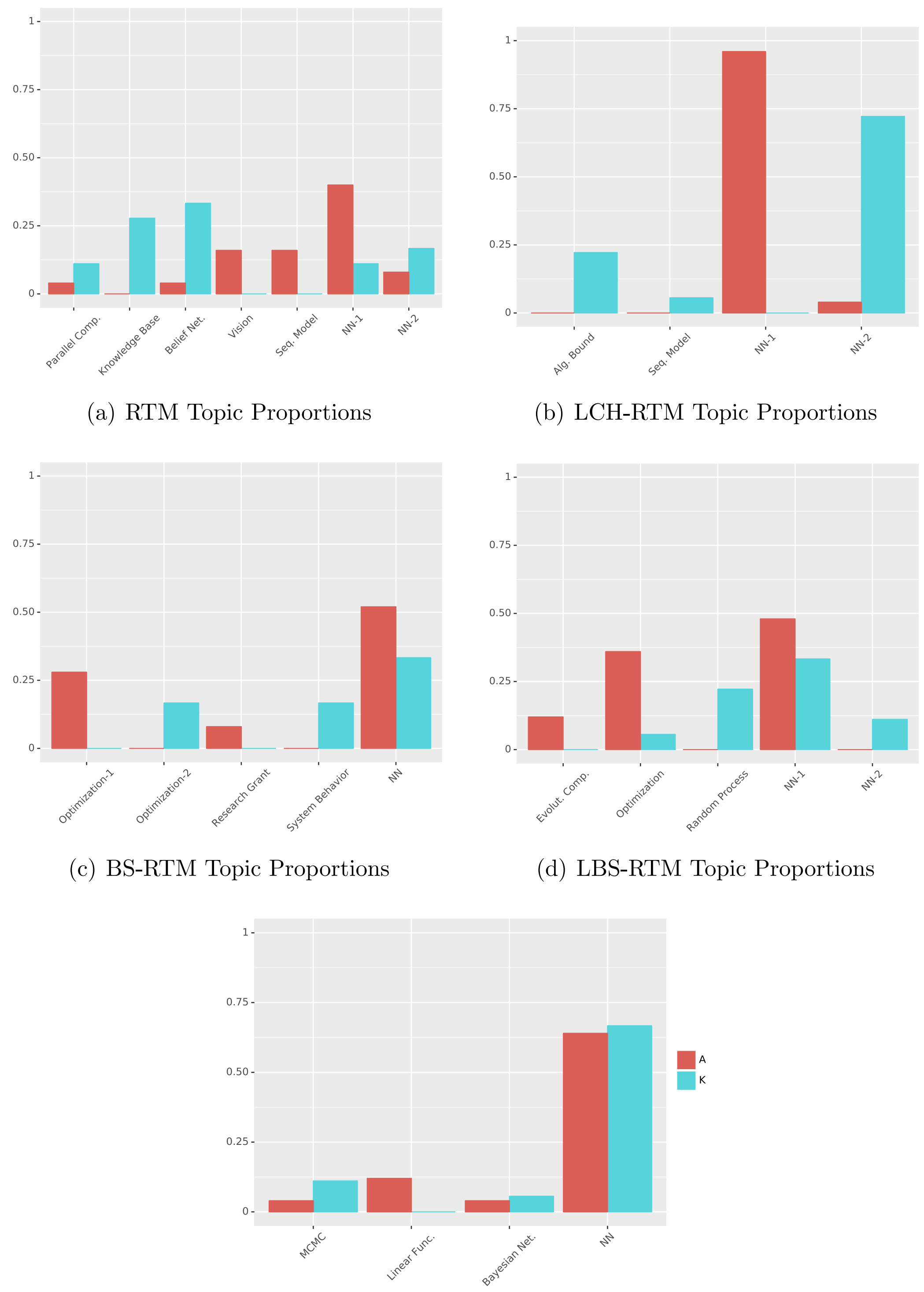
work", "model", "learn", "bayesian", etc.) and Bayesian Estimation ("estimate", "bayesian", "parameter", "analysis", etc.), which differs from Block 1's emphasis. Because of the two inter-block links, SCC merges the two blocks into one, which makes the block topic distribution unclear and misleads the sampler. WSBM, on the other hand, keeps the two blocks separate, which generates a high-quality prior for the sampler.

3.5 Summary

In this chapter, we focus on incorporating binary-valued document links into topic modeling, as they indicate the topic similarities of the two connected documents. We introduce LBH-RTM, a discriminative topic model that jointly models topics and binary document links. It detects latent blocks in the document network probabilistically by a weighted stochastic block model, rather than treating each link separately or via strongly connected-components as in previous models. We assign a separate Dirichlet prior for each block to capture its topic preferences, which serves as an informed prior when inferring documents' topic distributions in that block. We predict links using max-margin learning from documents' topic and word distributions and block assignments.

Our model better captures the connections and content of paper abstracts and web pages, as measured by predictive link rank and/or topic coherence. LBH-RTM yields topics with enhanced coherence, though not all techniques contribute to the improvement. We support our quantitative results with qualitative analysis by examining a pair of example documents and a pair of blocks, highlighting the robustness of WSBM over blocks defined as SCC.

While document links indicate the high-level (topic) similarity between documents, weighted word links provide the basic low-level semantic relatedness between words. Such information is beneficial for topic models to refine the topic words. Thus, in the next chapter, we will explore methods for incorporating weighted word links into topic modeling.



(a) RTM Topic Proportions

(b) LCH-RTM Topic Proportions

(c) BS-RTM Topic Proportions

(d) LBS-RTM Topic Proportions

(e) LBH-RTM Topic Proportions

Figure 3.9: Topic proportions given by various models on our two illustrative documents (K and A, described in described in Section 3.4.2). As the model gets more sophisticated, the Neural Network topic proportion gets higher and finally dominates the two documents when LBH-RTM is applied. Though LCH-RTM gives the highest proportion to the Neural Network topic, it splits the Neural Network topic into two and does not assign the proportions to the same one.

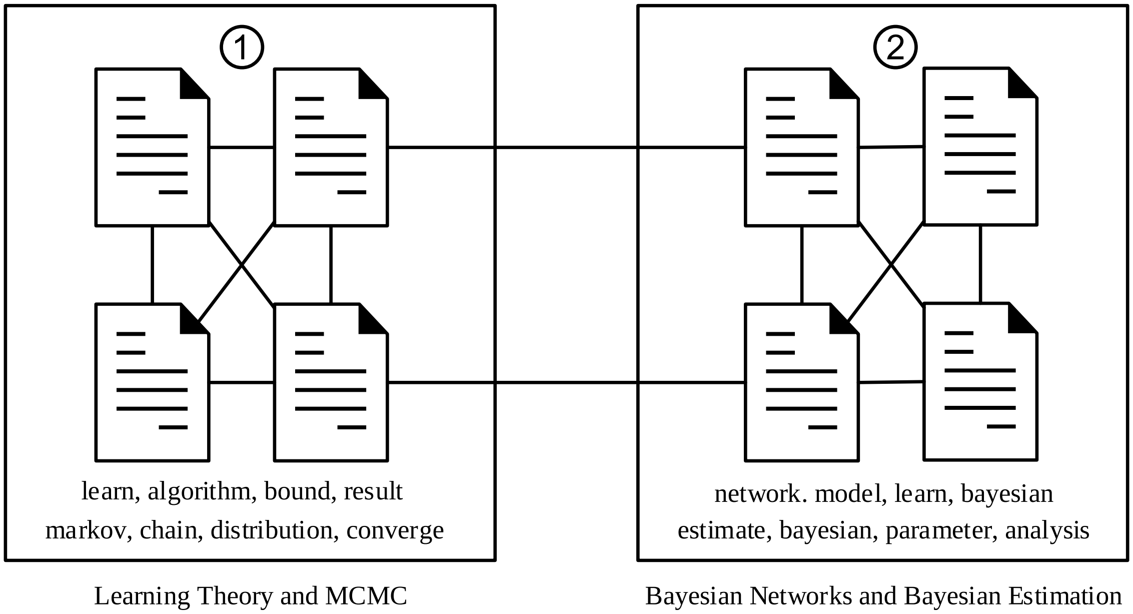

1
2
learn, algorithm, bound, result
markov, chain, distribution, converge
network. model, learn, bayesian
estimate, bayesian, parameter, analysis
Learning Theory and MCMC
Bayesian Networks and Bayesian Estimation

Figure 3.10: SCC fails to identify two blocks which are different in topic distributions, because of the two inter-block links. WSBM is robust enough to identify the two.

Chapter 4: Topic Modeling with Word Associations

In this chapter, we shift our focus from high-level binary document links to low-level weighted word links, or word association scores. Word association scores represent the word relatedness of word pairs and are easy to obtain from a large corpus, mostly based on statistical co-occurrences. Researchers have developed dozens of methods to compute these scores with various emphases, e.g., pointwise mutual information (Church and Hanks, 1990, PMI) when there are sufficient data, Fisher's exact test (Upton, 1992, FET), and log likelihood ratio (Moore, 2004; Dunning, 1993, LLR), as we used in Chapter 3, when data are limited. Recently, with the emergence of word embeddings, words are represented by vectors in a continuous semantic space (Mikolov et al., 2013; Pennington et al., 2014). Thus, the word associations can be evaluated with any methods applicable to vector similarities, e.g., cosine similarity and Euclidean distance.

In topic modeling, word association scores are especially important, because topic models put semantically related words in the same topic. Word association scores not only contain rich information about the vocabulary but also serve as an evaluation metric of topic interpretability (Chang et al., 2009). Nevertheless, most topic models are still trained using methods that optimize likelihood and not taking into account word association scores (McAuliffe and Blei, 2008; Nguyen et al., 2013).

Goodman (1996) introduces a key insight for machine learning models in natural language processing: if you know how performance on a problem is evaluated, it makes more sense to optimize using *that* evaluation metric, rather than others. Goodman applies this insight to parsing algorithms, but it has had an even more substantial impact in machine translation, where the introduction of the fully automatic BLEU metric makes it possible to tune systems using a score correlated with human rankings of machine translation system performance (Papineni et al., 2002).

We take the logical next step suggested by bringing together the insights of Goodman (1996) and Chang et al. (2009), namely incorporating an approximation of human topic interpretability into the topic model optimization process in a way that is effective and more straightforward than previous methods that involve heavy computation with the word association matrix in complex posterior regularizers (Newman et al., 2011). We take advantage of the human-centered evaluation of Chang et al. (2009), which can be reasonably approximated using an automatic metric based on real-valued word associations derived from a large, more general corpus (Lau et al., 2014, Section 2.2). We exploit LDA and its Bayesian formulation by bringing word associations into the picture using a prior—specifically, we dig into the dense external lexical associations to create a tree structure which encodes the most salient word association information and filters out redundancies. We then use *tree LDA* (Boyd-Graber et al., 2007, tLDA), which derives topics using a given tree prior.

We construct tree priors with combinations of two types of word association scores (skip-gram probability (Mikolov et al., 2013) and G2 likelihood ratio (Dunning, 1993)) learned on a large reference corpus and three construction algorithms (two-level method, hierarchical clustering with and without leaf duplication). Then tLDA identifies topics with these tree priors in Amazon reviews and the 20NewsGroups datasets. tLDA topics are more coherent than those given by "vanilla" LDA and the latent concept topic model (Hu and Tsujii, 2016, LCTM), which directly models on word embeddings instead of discrete word types while retaining and often slightly improving topics' extrinsic performance as features for supervised classification. Our approach can be viewed as a form of adaptation, and the flexibility of the tree prior approach—amenable to *any* association score—suggests that there are many directions to pursue beyond the two flavors of associations explored here. For instance, hierarchical word associations (e.g., hypernyms and hyponyms in Figure 4.2) could be encoded in the tree prior (Boyd-Graber et al., 2007); word translation dictionaries (Figure 4.3) are another source of word associations, in which word link weights are binary values—1 if two words are translations of each other and 0 if otherwise (Hu et al., 2014).[1]

4.1 Tree LDA: LDA with Tree Priors

Tree priors organize the vocabulary of a dataset in a tree structure (Figure 4.1), contrasting with introducing topic correlations (Blei and Lafferty, 2007; He et al., 2017). All words are located at the leaf nodes and share ancestor internal nodes

[1] The work done in this chapter has been published in "Adapting Topic Models using Lexical Associations with Tree Priors" (Yang et al., 2017).

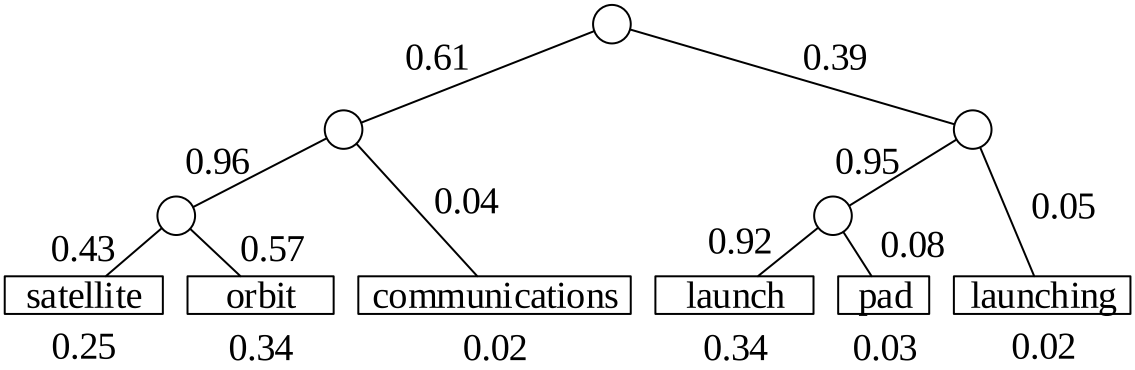


Figure 4.1: An example of a tree prior (the tree structure) and gold posterior edge and word probabilities learned by tLDA. Numbers beside the edges denote the probability of moving from the parent node to the child node. A word's probability (i.e., the number below the word) is the product of probabilities moving from the root to the leaf, e.g., Pr(orbit) $= 0.61 \times 0.96 \times 0.57 = 0.34$.

(circles in Figure 4.1). In our use of tree priors, if two words have a lower association score, their common ancestor node will be closer to the root node, e.g., contrast (orbit, satellite) with (orbit, launch). This encodes the word association information in the hierarchy. Highly semantically related words are organized in the same sub-tree, while less related words are placed in other sub-trees. It also significantly reduces the complexity of storing word associations. To encode V words, if we do not add duplicate leaf nodes, the space complexity of extra nodes and edges in the tree structure is $O(V)$, a contrast to the complexity of $O(V^2)$ for all word association scores which contain a lot of redundancy.

Tree LDA (Boyd-Graber et al., 2007, tLDA) is an LDA extension that creates topics from tree priors. Each topic corresponds to a tree prior with the same hierarchy. In a tree prior (Figure 4.1), an internal node is a multinomial distribution over its child nodes, and tLDA learns the probabilities of moving to them. For example, in one of the learned topics, the root node in Figure 4.1 has probabilities of 0.61 to move to its left child along the left edge and 0.39 to its right child along the right edge. A word can be reached from the root node via a unique path which consists of one or more internal nodes and edges.[2] The probability of a path is the product of probabilities of picking the nodes in the path, e.g., $\Pr(\text{satellite}) = 0.61 \times 0.96 \times 0.43 \approx 0.25$. Thus two paths with shared nodes, e,g., paths to "satellite" and "orbit", have correlated weights in a topic. The more semantically related of the two words (i.e., the farther of their lowest common ancestor to the root node), the more edges they have in common and the more correlated of their weights are in a topic. A topic in tLDA thus can be viewed as a multinomial distribution over the paths from the root to leaves. The generative process of tLDA is:

1. For topics $k \in \{1, \dots, K\}$ and internal nodes n_i
   (a) Draw child distribution $\boldsymbol{\pi_{k,i}} \sim \text{Dirichlet}(\beta)$[3]
2. For each document $d \in \{1, \dots, D\}$
   (a) Draw topic distribution $\boldsymbol{\theta_d} \sim \text{Dirichlet}(\alpha)$
   (b) For each token $t_{d,n}$ in document d
      i. Draw a topic $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta_d})$
      ii. Draw a path $y_{d,n}$ to word $w_{d,n}$ with probability $\prod_{(i,j) \in y_{d,n}} \pi_{z_{d,n},i,j}$

[2] If every word directly connects to the root node, tLDA degenerates to vanilla LDA.

[3] Unlike other tree-based topic models such as Andrzejewski et al. (2009), all Dirichlet hyperparameters are the same for all internal nodes. Regardless of cardinality, all Dirichlet parameters are the same scalar β.

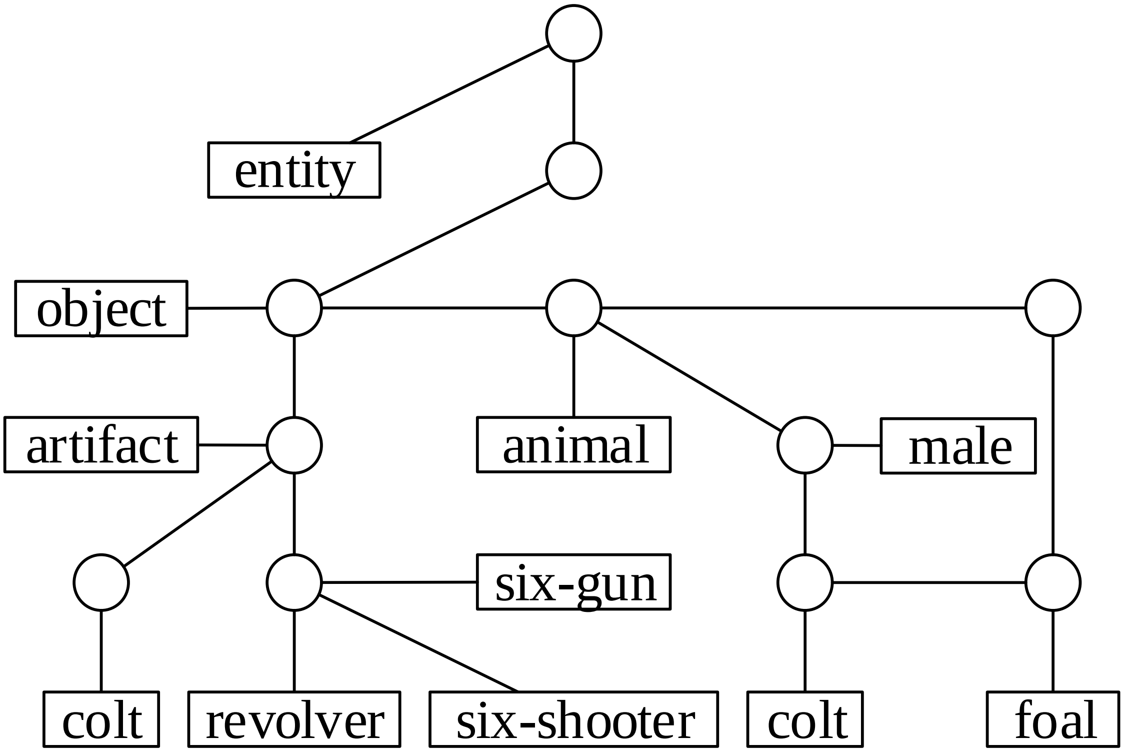


Figure 4.2: A part of a tree prior constructed from synonyms in WordNet. Adapted from Boyd-Graber et al. (2007).

tLDA can perform different tasks using different tree priors. If we encode synonyms from WordNet (Miller, 1995) in the tree prior, tLDA disambiguates word senses (Boyd-Graber et al., 2007). With word translation priors (Figure 4.3), it is a multilingual topic model (Hu et al., 2014).

4.1.1 Posterior Inference

The parameters in a tLDA model are inferred by Gibbs sampling (see Section 2.1), by updating the path assignment and the topic assignment for each token. Namely, for $t_{d,n}$, the n-th token in document d, the probability of assigning a

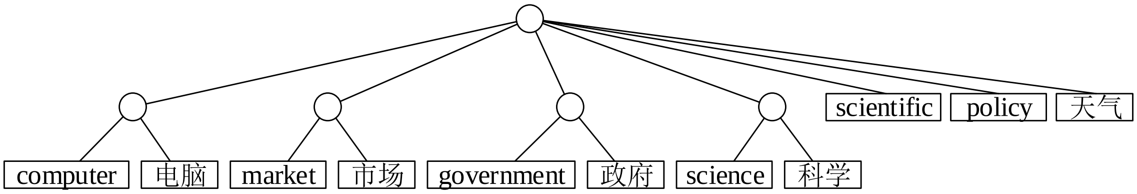


Figure 4.3: A part of a tree prior constructed from word translations. Adapted from Hu et al. (2014).

path $y_{d,n}$ and a topic $z_{d,n}$ is

$$\begin{aligned} &\Pr\left(z_{d,n}=k, y_{d,n}=s \,|\, \mathbf{z^{-d,n}}, \mathbf{y^{-d,n}}, w_{d,n}=v, \mathbf{w^{-d,n}}, \alpha, \beta\right) \\ &\propto \mathbb{1}\left(\Omega(s)=v\right)\left(N_{d,k}^{-d,n}+\alpha\right) \underbrace{\prod_{(i\to j)\in s} \frac{N_{i\to j,k}^{-d,n}+\beta}{\sum_{j'}\left(N_{i\to j',k}^{-d,n}+\beta\right)}}_{\text{Path probability.}}, \end{aligned} \tag{4.1}$$

where $\Omega(s)$ represents the word on the leaf node of path s; $\mathbb{1}(\cdot)$ is an indicator function. If the path s does not leads to word v, it gets a weight of zero. $N_{d,k}^{-d,n}$ denotes the number of tokens assigned to topic k in document d; $N_{i\to j,k}^{-d,n}$ denotes the number of times that edge $i \to j$ is chosen in topic k. $^{-d,n}$ denotes the count excludes $t_{d,n}$. The time complexity of inferring a token v's topic assignment is $O(K|S(v)|\sum_{s\in S(v)}|E(s)|)$ where $S(v)$ is the set of paths that lead to word v and $E(s)$ is the set of edges in path s.

4.2 Tree Prior Construction from Word Association Scores

We introduce three methods to extract information and build tree priors for tLDA from word association scores. The first method creates tree priors flatly by querying each word to word association scores and obtaining the closest words (Section 4.2.1). The other two methods, on the contrary, build tree priors with

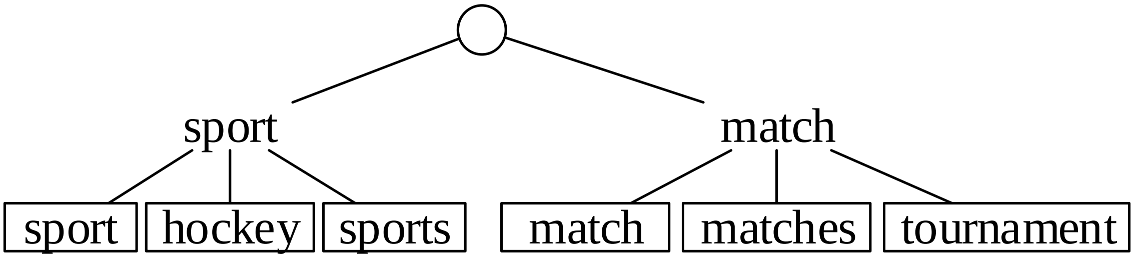


Figure 4.4: A two-level tree example with $N = 2$. The words in the internal nodes (i.e., "sport" and "match" *without* boxes) denote *concepts* and have no effect in tLDA. They are here only for exposition.

hierarchies that encode word association score magnitudes (Sections 4.2.2 and 4.2.3).

4.2.1 Two-Level (2LV)

Word association scores tell us the semantic relatedness of word pairs, so we can use the scores to find the most related words for a query word. For instance, if we query word embeddings with the word "sport", the closest words are "hockey" and "sports" according to cosine similarity.

A two-level tree (Figure 4.4) is constructed based on this intuition straightforwardly.[4] Each non-root internal node, n_i, is a *concept* associated with a word v_i in the vocabulary (e.g., "sport" and "match" without boxes in Figure 4.4), but this fact is not taken into account in posterior inference. Then we query the word association scores with word v_i and sort all other words in descending order of their association scores with v_i. The top N most associated words with v_i are selected as the internal node n_i's child leaf nodes, e.g., "hockey" and "sports" for "sport" if we set $N = 2$. However, a word in the vocabulary may not be selected as a leaf node if it is not among the top N most associated words with any other words. In this case,

[4] The root node is not considered a level.

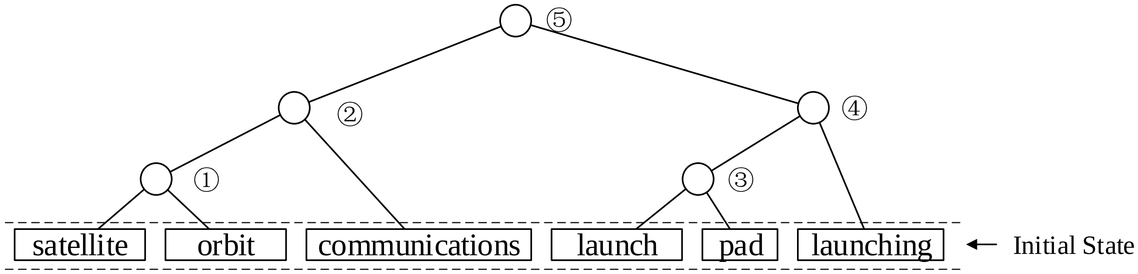


Figure 4.5: An example of building tree priors based on hierarchical agglomerative clustering (Lukasová, 1979, HAC). The construction starts from leaf nodes only, i.e., initial state. Then it repeatedly merges the clusters with the highest association score, as marked by the numbers, until there is only one left.

tLDA is unable to generate this word. Thus, n_i has an additional child node, which represents the word v_i itself, to ensure that v_i appears at the leaf level at least once so that it can be generated by tLDA.[5] Therefore, if the vocabulary size is V, there will be a total of $(N+1)V$ leaf nodes.

4.2.2 Hierarchical Agglomerative Clustering (HAC)

While a two-level tree is bushy (high branching factor) and flat, hierarchical agglomerative clustering (Lukasová, 1979, HAC) reduces the number of leaf nodes and encodes levels of word association information in its hierarchy (Figure 4.5). It conforms better to the intuition of tree priors that highly associated words should have the lowest common ancestor far from the root node.

The HAC process starts from V clusters representing the V words in the vocabulary (i.e., "Initial State" in Figure 4.5) and then builds the hierarchy. In each iteration, HAC selects the two clusters with the highest association score and creates a new internal node that connects to them. It repeats this process until

[5] All tree prior examples are real sub-trees of the priors built on Gigaword 5. See Section 4.3.

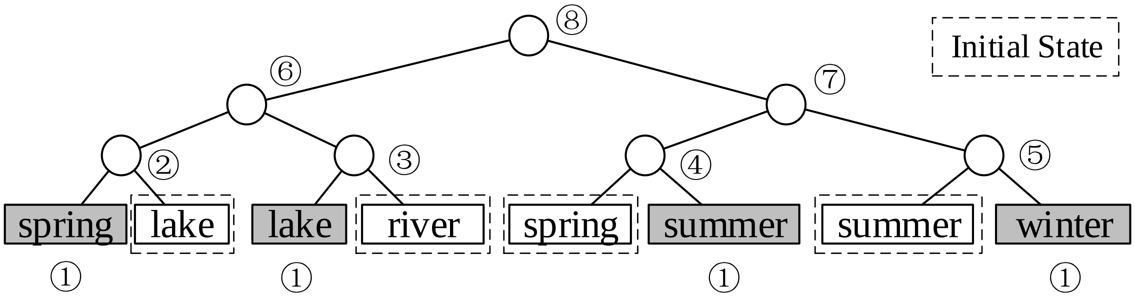


Figure 4.6: An example of constructing HAC with leaf duplication (HAC-LD) tree prior for the words “lake”, “river”, “spring”, and “summer” (“Initial State”), whose paired words are shaded in gray and marked with ①. HAC-LD alleviates the problem in HAC that a word with multiple senses can only be assigned to a single cluster close to one of its senses, e.g., the word “spring” which can be either a season or a body of water.

there is only one cluster left, as marked by the numbers beside the internal nodes in Figure 4.5.

In the clustering process, if two clusters both only have one word, their association score is just the two words' association score. If at least one of the two clusters, denoted by C_i and C_j, has multiple words, their association score is the average association score of the pairwise words from the two clusters:

$$S\left(C_i, C_j\right) = \frac{1}{|C_i||C_j|} \sum_{w_1 \in C_i} \sum_{w_2 \in C_j} S\left(w_1, w_2\right). \tag{4.2}$$

4.2.3 HAC with Leaf Duplication (HAC-LD)

In a tree prior constructed by HAC, a word appears in the leaf exactly once. This is fine for the words with a single sense, but may be problematic if a word has multiple senses. For example, the word “spring” could mean either a season (similar to “summer”) or a place with water (similar to “lake”). HAC can only assign it to a sub-tree close to one of its senses and will cause information loss on the other side.

To alleviate this problem, we create duplicate leaf nodes before running HAC. The leaf duplication pairs every word with its most semantically similar word according to word association scores and create a cluster with the pair. For instance, in Figure 4.6, “lake”, “river”, “spring”, and “summer” in white boxes (“Initial State”) are paired with “spring”, “lake”, “summer”, and “winter” in gray boxes respectively, as indicated by “①”. In this procedure, although “spring” is paired with “summer”, “lake”’s most similar word is “spring”, so that “spring” appears in both senses simultaneously, which reduces the information loss. Then we apply HAC as described in Section 4.2.2.

4.3 Experimental Results

We compute two versions of word association scores from Gigaword 5, using word2vec skip-gram model (Mikolov et al., 2013) and G2 likelihood ratio (Dunning, 1993).[6] Word2vec gives the vector representation of words rather than association scores, so for two words w_i and w_j, represented by vectors $\mathbf{v_i}$ and $\mathbf{v_j}$, their word2vec association score is their skip-gram probability:

$$S(w_i, w_j) = \frac{\exp(\mathbf{v_i} \cdot \mathbf{v_j})}{\sum_k \exp(\mathbf{v_i} \cdot \mathbf{v_k})}, \tag{4.3}$$

where $\cdot$ denotes dot product. Then we apply the three tree construction algorithms to construct a total of six tree priors. In the two-level trees, the value of N (i.e., the number of child nodes per internal node) is ten.

We evaluate the models on the corpora of Amazon reviews (Jindal and Liu,

[6] `https://catalog.ldc.upenn.edu/ldc2011t07`.

Corpus	#Vocabulary	#Docs	#Tokens	#Classes
20NewsGroups	9,194	18,769	1.75M	20
Amazon	9,410	39,392	1.51M	2

Table 4.1: Corpora Statistics.

2008) and 20NewsGroups (Lang, 1995). We apply the same tokenization and stopword removal methods. We then sort the words in the vocabularies by their document frequencies and return the top words, while also removing words that appear in more than 30% of the documents. The statistics of the corpora after preprocessing are given in Table 4.1.

Both corpora are split into five folds. For classification tasks, each fold is further equally divided into a development set and a test set when it is used for evaluation. All the results reported below are averages across five-fold cross-validation using twenty topics with hyper-parameters $\alpha = \beta = 0.01$. In 20NewsGroups, each post is assigned to one of twenty news groups, so we perform a twenty-class classification. For Amazon reviews, 4–5 star reviews are given positive labels, 1–2 stars are given negative, and reviews with 3 stars are discarded, which creates a binary classification task.

4.3.1 Perplexity

Before evaluating topic quality, we conduct a sanity check of the models' average perplexity (see Section 2.2) on the test sets (Table 4.2).

LDA achieves the lowest perplexity among all models on both corpora while tLDA models yield suboptimal perplexity results owing to the constraints given by

Model	Tree	20NewsGroups	Amazon
LDA	–	2158.74	999.98
tLDA	G2-2LV	2214.99	1018.72
	G2-HAC	2234.34	1017.17
	G2-HAC-LD	2251.65	1015.06
tLDA	W2V-2LV	2204.94	1016.31
	W2V-HAC	2222.53	1013.07
	W2V-HAC-LD	2234.08	1017.77

Table 4.2: The average perplexity results on the test sets by various models. Tree names indicate the word association score and tree prior construction algorithm. LDA gives the lowest perplexity, because tLDA models have constraint from the tree priors and sacrifice the perplexity.

tree priors.[7] As shown in the following sections, the sacrifice in perplexity brings improvement in topic coherence, while not hurting or slightly improving extrinsic performance using topics as features in supervised classification.

Tree priors built from word2vec skip-gram model generally outperform the ones created using the G2 likelihood ratio when using the same tree prior construction algorithm. Among the three tree prior construction algorithms, the two-level method is the best on the 20NewsGroups corpus. However, there is no such consistent pattern on Amazon reviews.

4.3.2 Topic Coherence

Instead of manually evaluating topic quality using word intrusion (Chang et al., 2009), we use an automatic alternative to calculate topic coherence (Lau et al., 2014, Section 2.2). For every topic, we extract its top ten words and compute average pairwise PMI (Church and Hanks, 1990) scores on a reference corpus of Wikipedia

[7] The constraints could be treated as additional implicit training data, as they are extracted or learned from an external dataset.

dump as of October 8, 2014.[8]

We include vanilla LDA and the latent concept topic model (Hu and Tsujii, 2016, LCTM) as baselines. LCTM also incorporates external knowledge from word embeddings. It assumes that latent concepts c's exist in the embedding space and generate nearby word w's embeddings via multivariate Gaussian distributions with means of their coordinates, i.e., $\Pr(w \mid c) = \mathcal{N}(\boldsymbol{\mu_c}, \sigma^2 I)$. And a topic k in LCTM is a multinomial distribution over these concepts c's, i.e., c is conditioned on k or $\Pr(c \mid k)$. To compare LCTM topics with LDA and tLDA, we marginalize over concepts and obtain the probability mass of every word in every topic as

$$\Pr(w \mid k) = \sum_{c=1}^{C} \Pr(w \mid c) \Pr(c \mid k), \tag{4.4}$$

where w and k denote the word and the topic respectively; C is the pre-defined number of latent concepts.

Most tLDA models yield more coherent topics than vanilla LDA (Figure 4.7). Among all tLDA models, the two-level tree built on word2vec skip-gram model improves the most. LCTM performs poorly: all its topics consist of words like “don”, “dodgers”, “au”, “alot”, “people”, “alicea”, “uw”, “arabia”, “sps”, and “entry” with slight differences in order.

To show how subjective topic quality improves over LDA, we extract the topics from 20NewsGroups given by vanilla LDA and the tLDA with two-level tree priors built on word2vec skip-gram model, pair them, and sort the pairs based on Kullback–Leibler divergence (Kullback and Leibler, 1951, KLD). In Table 4.3, we

[8] https://wiki.umiacs.umd.edu/clip/clipwiki/index.php?title=Data#Wikipedia

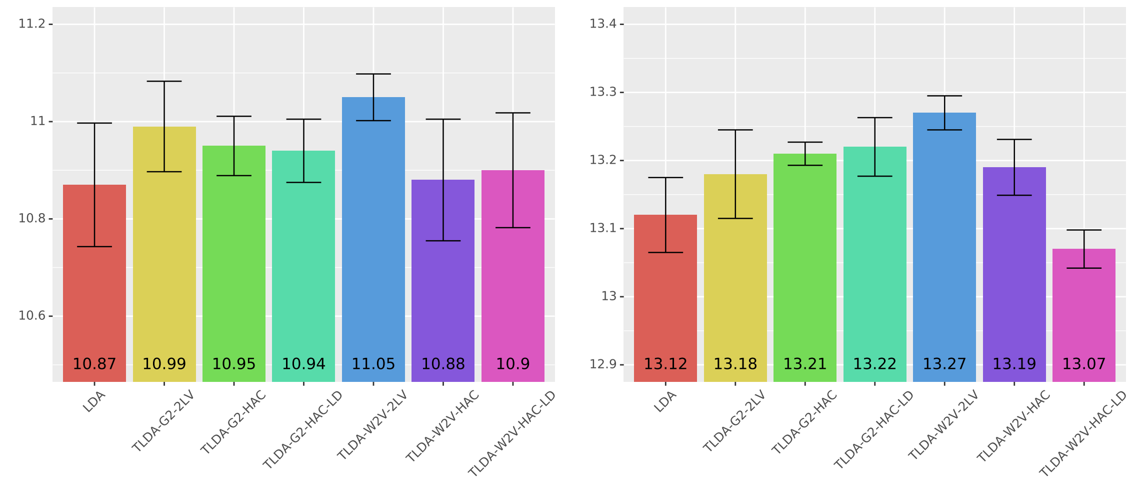


Figure 4.7: Average PMI scores of the top 10 words in topics given by LDA and tLDA on 20NewsGroups (left) and Amazon reviews (right). Most tLDA topics are more coherent than LDA topics, while the two-level tree priors created on word2vec improve the most. The PMI scores of LCTM are too low to be included: 8.86 ± 0.66 on 20NewsGroups and 6.34 ± 1.21 on Amazon reviews.

select and present three topics from each of the top, middle, and bottom third of the sorted topics.

The topics with low KLD, Christian, Security, and Middle East, are generally coherent and do not have significant differences. Although the topics of Sports have medium KLD and quite different words, both of them are still coherent. As KLD increases, tLDA topics gradually become more coherent than LDA and have more relevant words. In the University Research topics, tLDA includes more research-related words, e.g., “center”, “science”, and “institute”. In the Health topics, the tLDA topic has more coherent words like “patients”, “insurance”, “drugs”, “aids”, and “treatment”, while LDA includes less relevant words, e.g., “food”, “sex”, “cramer”, and “men”.

In the topics with high KLD, tLDA topics are also more coherent. For instance,

Topic	KLD	Model	Words
Christian	0.709	LDA	god, jesus, church, christ, christian, bible, man, christians, lord, sin
		tLDA	god, jesus, bible, christian, christ, church, christians, faith, people, lord
Security	0.720	LDA	key, encryption, chip, clipper, keys, government, public, security, system, law
		tLDA	key, encryption, chip, clipper, government, keys, privacy, security, system, public
Middle East	0.765	LDA	israel, jews, war, israeli, jewish, arab, people, world, peace, muslims
		tLDA	israel, jews, israeli, war, jewish, arab, muslims, people, peace, world
Sports	1.212	LDA	hockey, team, game, play, la, nhl, ca, period, pit, cup
		tLDA	game, team, year, games, play, players, hockey, season, win, baseball
University Research	1.647	LDA	university, information, national, april, states, year, research, number, united, american
		tLDA	university, research, information, april, national, **center**, **science**, year, number, **institute**
Health	1.914	LDA	medical, people, disease, health, cancer, *food*, *sex*, *cramer*, *men*, drug
		tLDA	health, medical, disease, drug, cancer, **patients**, **insurance**, **drugs**, **aids**, **treatment**
Images	1.995	LDA	image, ftp, software, graphics, *mail*, *data*, version, file, pub, images
		tLDA	file, image, **jpeg**, graphics, images, files, format, **bit**, **color**, program
Hardware	2.127	LDA	drive, card, mb, scsi, disk, *mac*, system, *pc*, *apple*, bit
		tLDA	drive, scsi, disk, mb, hard, **drives**, **dos**, **controller**, **ide**, system
People	2.512	LDA	armenian, people, turkish, armenians, armenia, turkey, turks, *didn*, soviet, *time*
		tLDA	armenian, turkish, armenians, armenia, turkey, turks, soviet, people, **russian**, genocide

Table 4.3: We sort topics into thirds by Kullback-Leibler divergence (Kullback and Leibler, 1951, KLD): low, medium, and high divergence between vanilla LDA and tLDA. Unique coherent words are in **blue and bold**. Unique incoherent words are in *red and italic*. tLDA brings in more topic-relevant words.

in the Images topics, the LDA topic contains less relevant words like “mail” and “data”, while the tLDA topic mostly consists of words related to images, and even includes words like “jpeg”, “color”, and “bit” that are not among the top words in the LDA topic.[9] In the topics for Hardware, there are more words closer to the hardware level of computers for tLDA, such as “drives”, “dos”, “controller”, and “ide”, in contrast to LDA, e.g., “mac”, “pc”, and “apple”. tLDA also ranks hardware-related words higher. For instance, “scsi” and “disk” come before “mb”. The words in the topics for People are generally coherent, although the tLDA topic has one more specific word of “russian” and the LDA topic includes “didn” and “time” that are less relevant to the topic.

4.3.3 Extrinsic Classification

To extrinsically evaluate topic quality, we use binary and multi-class classification on Amazon reviews and 20NewsGroups corpora using SVM-light (Joachims, 1998) and SVM-multiclass (Tsochantaridis et al., 2004) respectively.[10] We tune the parameter C, the trade-off between training error and margin, on the development set and apply the trained model with the best performance on the development set to the test set. The classification accuracies are given in Table 4.4.

We compare the accuracies with the features of bag-of-words (BoW) and topic

[9] The topic names are summarized manually, so some topics can be interpreted in another way, e.g., Image Transfer instead of Images. See Section 6.2.1.

[10] SVM-light: `http://svmlight.joachims.org/`. SVM-multiclass: `https://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html`.

Model	Tree	Path	20NewsGroups	Amazon
MFC	–	–	5.80	78.76
BoW	–	–	86.64	86.73
BoW+Vec	–	–	86.59	**87.30**
LDA	–	–	86.67	86.99
LCTM	–	–	86.52	86.83
tLDA	W2V-2LV	N	86.75	87.07
		Y	86.73	87.13
	W2V-HAC	–	86.79	87.19
	W2V-HAC-LD	N	86.73	87.02
		Y	86.94	86.88
tLDA	G2-2LV	N	86.82	87.15
		Y	**86.96**	87.05
	G2-HAC	–	86.63	87.11
	G2-HAC-LD	N	86.73	87.07
		Y	86.91	86.94

Table 4.4: Accuracies of topical classification on 20NewsGroups and sentiment analysis on Amazon reviews. Although not significantly improving the performance, tLDA topics at least do not hurt.

posteriors inferred by vanilla LDA, LCTM, and tLDA. For the tLDA models with two-level and HAC-LD tree priors, the path assignment is an additional feature, and we run experiments both with and without it. The tLDA models with HAC prior do not have this feature, because every word appears in the tree prior precisely once and the paths have a one-to-one mapping with the vocabulary. We also include the features of BoW and the average word vector for the document (BoW+Vec) a naïve baseline of most frequent class (MFC).

Features based on most tLDA topic posteriors perform at least as well as LDA-based topic features and often slightly better, although with no statistical significance. This proves that our tree priors do not sacrifice extrinsic performance for improving topic coherence. Also, the path assignment feature improves topical classification on 20NewsGroups but not sentiment classification on Amazon reviews.

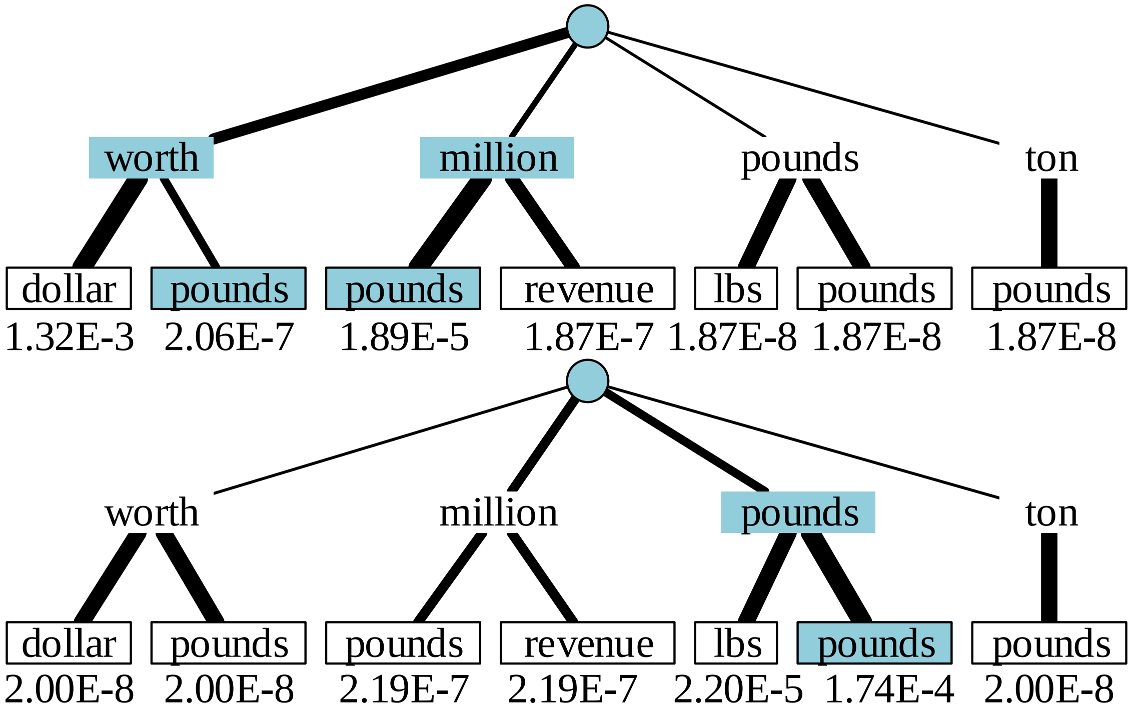


Figure 4.8: Sub-trees for the word "pounds" in two topics, from the 20NewsGroups corpus using a two-level tree prior built on word2vec. "Pounds" is more associated with the sense of British currency in the Politics topic (upper), while closer to the sense of weight unit in the Health topic (lower). High probability *paths* are shaded in blue and high probability *edges* have thicker lines.

LCTM-based features work worse than all topic model- and word2vec-based features and only beats the BoW baseline on Amazon reviews. Although the word2vec feature (BoW+Vec) performs the best on Amazon reviews, it lacks the interpretability of topic models.[11]

4.3.4 Learned Trees

In a tree prior, polysemous words may appear in several sub-trees. Its sense at a sub-tree could be identified by the words in the same or nearby sub-trees. For

[11] According to further analysis, the classification accuracies among the models do not have statistical significance.

instance, in Figure 4.6, the word “spring” appears in two sub-trees. Given the nearby words “lake” and “river”, we can tell that the “spring” in the left sub-tree denotes water. Similarly, the “spring” in the right sub-tree means a season, according to the sibling words “summer” and “winter”.

Tree-based topics distinguish polysemous words by assigning weights to their senses. Take the word “pounds” as an example. It can be either a British currency or a weight unit. In the topic of Politics with words “president”, “people”, “clinton”, “myers”, and “money”, the word “pounds” is more likely to be the unit of budgets. As we can see from the upper sub-tree in Figure 4.8, “pounds” is more likely to be reached in the sense of British currency via the paths of root → worth → pounds and root → million → pounds, with nearby words “worth”, “million”, “dollar”, and “revenue”. In the Health topic (“health”, “medical”, “disease”, “drug”, and “cancer”), “pounds” is likely to be the unit of people’s weights. Thus it is more likely to be reached from the weight unit sense via the path root → pounds → pounds, which is reflected from the lower sub-tree in Figure 4.8.

4.4 Summary

This chapter focuses on incorporating weighted word links, or word association scores, into topic modeling. We introduce three methods that find latent tree structures from dense and flat word associations, with the intent of extracting key information and reducing redundancy. We combine topic models and word association scores, based on either traditional statistical methods or more recent word embeddings, more simply and flexibly than in previous work (Hu and Tsujii, 2016). Moreover, they are not restricted to the ones we use above and can handle *any* word association scores.

With the tree priors built by our methods, tLDA yields more coherent topics than vanilla LDA and LCTM, both quantitatively and qualitatively, although it sacrifices some (less important) perplexity performance due to the constraint from tree priors. Meanwhile, it maintains extrinsic performance comparable to, if not better than, LDA and LCTM in binary and multi-class classification tasks with BoW and inferred topic posteriors. Also, it is less computationally costly than LCTM: tLDA Java implementation converges in twelve hours, while LCTM needs sixty hours on the same machine with 2.8GHz Intel Xeon CPUs and 110G of memory.

So far, we have incorporated the observable document links and easy-to-get word links into topic models and obtained more coherent topics and good extrinsic performance. In the next chapter, we will study weighted topic links which are unobservable due to the latent nature of topics. We will learn, instead of incorporate, weighted topic links to connect the topics across languages and develop a novel multilingual topic model.

Chapter 5: Topic Model for Learning Weighted Topic Links

In a topic model, topics are different from documents and words, because they are latent rather than observed. This makes it hard to obtain the ground-truth topic links, not to mention to incorporate them into topic models as external knowledge. However, we can instead learn the weighted topic links, and they can be useful in a multilingual case to connect similar topics *across* languages, hence a multilingual topic model (MTM).

Multilingual topic models uncover latent topics *across* languages. Latent topics—represented as distributions over words—summarize documents and help analysts discover trends (Lau et al., 2012), analyze emotions (Bao et al., 2009), or recommend content (Marlin, 2003). MTMs, in contrast to monolingual topic models, reveal commonalities and differences between documents in different languages and the cultures they represent (Ni et al., 2009; Shi et al., 2016; Gutiérrez et al., 2016). Like most multilingual algorithms, including multilingual word embeddings, there must be some source of knowledge to bridge the languages. For instance, document parallelism indicates the equivalence of documents in multiple languages (Søgaard et al., 2015; Hermann and Blunsom, 2014; Vulić and Moens, 2015; Mimno et al., 2009; Hao and Paul, 2018). Another source of cross-lingual knowledge is the word translations which map words in one language to those that have similar meanings in another language (Faruqui and Dyer, 2014; Lu et al., 2015; Ammar et al., 2016; Boyd-Graber and Blei, 2009; Jagarlamudi and Daumé III, 2010; Boyd-Graber and Resnik, 2010; Hu et al., 2014).

Existing MTMs extend latent Dirichlet allocation (Blei et al., 2003, LDA) and learn same numbers of topics or even *aligned* topics across languages. The polylingual topic model learns topics on parallel corpora and assumes the same topics across eleven European languages (Mimno et al., 2009). Hu et al. (2014) encode the word translations in a tree prior and pair each English topic with a Chinese topic. Code-Switched LDA learns language-specific topic distributions from multilingual documents, i.e., some documents contain words in multiple languages (Peng et al., 2014). It does not learn aligned topics, and can identify topics present in only one language, but this is done by a heuristic in postprocessing instead of jointly modeling it with topics. In addition, it requires the same numbers of topics across languages, which reduces its flexibility.

Most prior models have tended to work well because, even if it is not technically built into the model, their implicit assumption is that the data are comparable or even parallel and have been applied to datasets where this is true. However, this assumption does not always comport with reality, because documents from the same geographic region during the same period can discuss very different things across languages. Consider a day's worth of tweets, blogs, and newspapers in multicultural London: Hindi tweets might focus on a Bollywood actor's appearance on BBC, Chinese newspapers might discuss Lunar New Year, French blogs might fret about

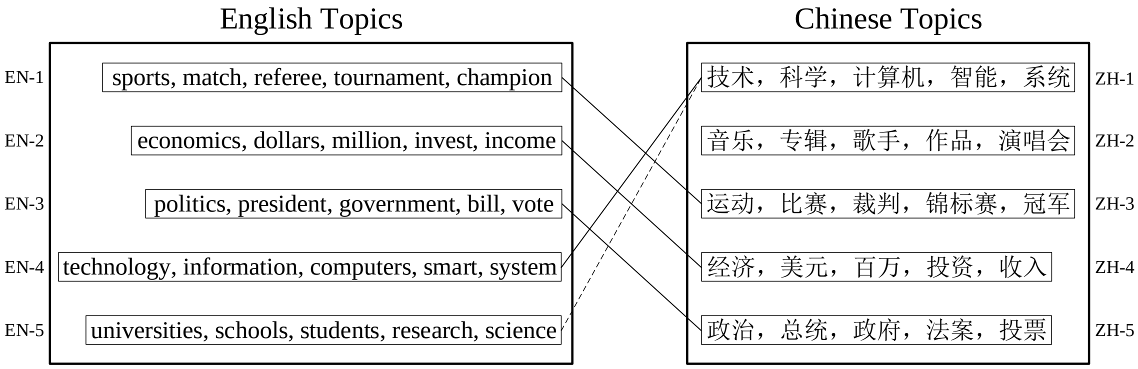

English Topics
Chinese Topics
EN-1
sports, match, referee, tournament, champion
EN-2
economics, dollars, million, invest, income
EN-3
politics, president, government, bill, vote
EN-4
technology, information, computers, smart, system
EN-5
universities, schools, students, research, science
技术，科学，计算机，智能，系统
ZH-1
音乐，专辑，歌手，作品，演唱会
ZH-2
运动，比赛，裁判，锦标赛，冠军
ZH-3
经济，美元，百万，投资，收入
ZH-4
政治，总统，政府，法案，投票
ZH-5

Figure 5.1: Past multilingual topic models (MTMs) learn aligned topics across languages, which is problematic on the corpora with low comparability. Our MTM overcomes it by learning weighted topic links without forcing topic alignment: topic pairs with many word translation pairs have high link weights, e.g., (EN-1, ZH-3), (EN-2, ZH-4), and (EN-3, ZH-5); topic pairs with partial overlap receive lower weights, e.g., (EN-5, ZH-1); a topic is unlinked if there is no corresponding topic in the other language (ZH-2).

Brexit, and English articles might dwell on changes in Tottenham's lineup. Even in a "comparable" setting, consideration of multiple languages brings to the forefront the fact that, while some topics are shared, the emphasis may differ across languages, and some topics may not have clear analogs across languages. For instance, in the news articles about Earthquake in the Chinese language pack released by the LORELEI program, English articles talk about earthquakes worldwide, while Chinese articles focus on the Wenchuan Earthquake, which occurred in 2008 in Sichuan Province.[1]

We, therefore, introduce a new multilingual topic model that assumes each language has its own topic sets which consist of the words in that language only. Our MTM jointly learns all topics but does not force the topics to be aligned across languages. Instead, it learns *real-valued weighted* links across languages and only

[1] LORELEI is short for LOw REsource Languages for Emergent Incidents: `https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents`.

assigns high topic link weight to a pair of topics when the two topics' top words have many direct translation pairs, e.g., (EN-1, ZH-3), (EN-2, ZH-4), and (EN-3, ZH-5) in Figure 5.1. If two topics have limited overlap, the link weight will be lower. For instance, topics EN-5 and ZH-1 have an overlap on "science" and "科学 (kē xué)", so they are weakly linked, while ZH-1 is strongly linked with EN-4. More importantly, the model allows a topic to remain unlinked if there is no corresponding topic in the other language (e.g., ZH-2 about Music), which makes the model robust in the (more common) case of partially comparable and even incomparable data with topic misalignment.

Via these weighted topic links, topic patterns can be conveyed from one language to the other as external knowledge for the latter. This helps improve topic quality for both languages and is particularly useful in scenarios that involve modeling topics on low-resource languages with very limited data, e.g., humanitarian assistance, peacekeeping, and/or infectious disease response. By learning the MTM on documents in a high-resource language, e.g., English, along with the documents in a low-resource language, e.g., Sinhalese, topic links will transfer relevant topic patterns from English to Singhalese, producing a better topic model on the low-resource language, while limiting the additional cost to other steps that will also need to be taken, such as finding or creating a word translation dictionary.

We describe our MTM in the bilingual case with languages S and T, and it is relatively easy to generalize it to multilingual situations. The MTM has two matrices—$\boldsymbol{\rho_{T \to S}}$ and $\boldsymbol{\rho_{S \to T}}$. They store topic link weight matrices and convert the topics from language T to S and S to T, respectively. Take $\boldsymbol{\rho_{S \to T}}$ for example, its values are learned by converting a word's topic distribution in language S to the topic space of language T and making it as close as possible to its translation word's topic distribution, as shown in Figure 5.2 and as will be discussed in more detail in Section 5.1. In this process, the shared topic pairs across languages will get higher weights, while a unique topic in a language will have a high-entropy weight distribution over the topics in the other language.

We validate the MTM in two classification tasks, one using inferred topic posteriors to predict Wikipedia document categories and the other looking for the need for rescue resources in disaster-related documents. Our MTM substantially outperforms other models as measured using F1 in both intra- and cross-lingual evaluations, while yielding coherent topics and meaningful topic links. We also demonstrate robust topic coherence even on low-comparability and small-size data.

5.1 Multilingual Topic Model for Connecting Cross-Lingual Topics

We introduce a formulation of posterior regularization (Section 2.3.2.2) that links languages in a topic model. For simplicity of exposition, we focus on the bilingual case, which has a language S with K_S topics and another language T with K_T topics and each topic is a distribution over the words in its language.

The MTM has two matrices, $\boldsymbol{\rho_{T \to S}}$ (size $K_S \times K_T$) and $\boldsymbol{\rho_{S \to T}}$ (size $K_T \times K_S$), that store topic link weights and convert the topics from language T to S and S to T respectively. Both matrices are critical for the MTM and neither can be derived from the other. Although we do not add constraints or regularizations on $\boldsymbol{\rho}$'s value

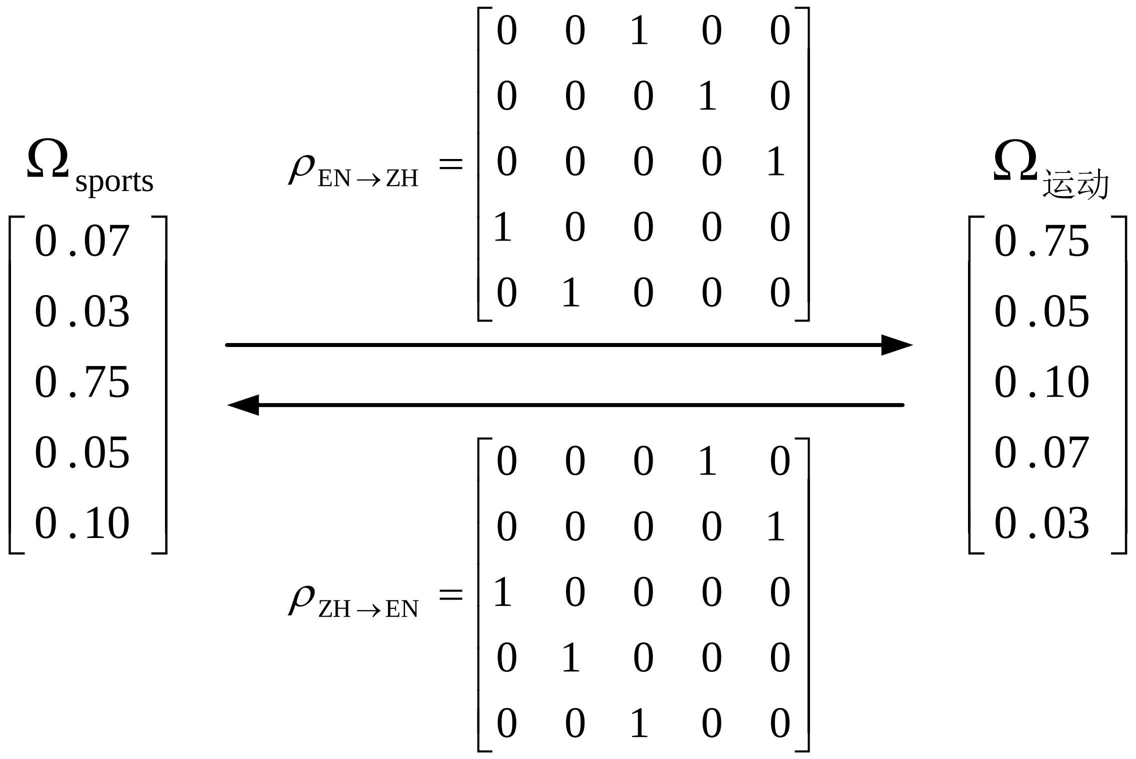


Figure 5.2: Our model uses topic link weight matrices $\boldsymbol{\rho}$'s to transform topics from one language to another. Unlike other models, it allows a topic to linked to another topic or multiple topics.

ranges, the values are between zero and one because the sum of each input/output training example vector is one. Each cell, $\rho_{T\rightarrow S,k_S,k_T}$, denotes the link weight of topics k_S and k_T while transforming from language T to S. The values of $\boldsymbol{\rho}$'s are learned from the translation pair's topic distributions—for a translation pair w_S and w_T, $\boldsymbol{\rho}$'s try to connect the topics (i.e., assign higher link weights) that have high probability mass in the two words' topic distributions. So to learn $\boldsymbol{\rho}$'s, we first define the topic distribution of a word w as the proportion of assignments to topic k to the assignments to all topics,

$$\Omega_{w,k} = \frac{N_{k,w}}{N_w}, \tag{5.1}$$

where N_w is w's total term frequency. The intuition is that if the two words in a translation pair are often assigned to the two topics in two languages, the two topics are likely to be corresponding topics in the two languages. Thus the values of $\boldsymbol{\rho}$'s are then optimized by converting the words' topic distributions in one language into the topic space of the other language using $\boldsymbol{\rho}$'s, and making them as close as possible to their translation words' topic distributions in the other language. For instance, given the translation pair of "sports" and "运动 (yùn dòng)", we want $\boldsymbol{\rho}_{\mathrm{EN}\to\mathrm{ZH}}\boldsymbol{\Omega}_{\mathrm{sports}}$ to be as close as possible to $\boldsymbol{\Omega}_{运动}$ and vice versa (Figure 5.2). The objective function for optimizing $\boldsymbol{\rho}$'s is the distance between a word's topic distribution, e.g., $\boldsymbol{\Omega}_{运动}$, and its translation's after transformation using $\boldsymbol{\rho}$'s, e.g., $\boldsymbol{\rho}_{\mathrm{EN}\to\mathrm{ZH}}\boldsymbol{\Omega}_{\mathrm{sports}}$, formulated as $\mathrm{Dis}(\boldsymbol{\Omega}_{运动}, \boldsymbol{\rho}_{\mathrm{EN}\to\mathrm{ZH}}\boldsymbol{\Omega}_{\mathrm{sports}})$ where $\mathrm{Dis}(\cdot,\cdot)$ denotes the Euclidean (or other) distance function of two topic distributions.

In addition, a translation pair is less reliable if one or both words have low term frequencies or high document frequencies. Thus, we add weights for translation pairs to reflect their importance and obtain the weighted distance product over all C translation pairs as

$$\prod_{c=1}^{C} \mathrm{Dis}\left(\boldsymbol{\Omega}_{\boldsymbol{S},\boldsymbol{c}}, \boldsymbol{\rho}_{\boldsymbol{T}\to\boldsymbol{S}}\boldsymbol{\Omega}_{\boldsymbol{T},\boldsymbol{c}}\right)^{\eta_c} \mathrm{Dis}\left(\boldsymbol{\rho}_{\boldsymbol{S}\to\boldsymbol{T}}\boldsymbol{\Omega}_{\boldsymbol{S},\boldsymbol{c}}, \boldsymbol{\Omega}_{\boldsymbol{T},\boldsymbol{c}}\right)^{\eta_c}, \tag{5.2}$$

where η_c is the weight of the c-th translation pair. We then compose the knowledge potential function $f(\mathbf{z},\mathbf{w})$ (Section 2.3.2.2) by taking the reciprocal and a logarithm of Equation 5.2 so that we can maximize it:

$$\begin{aligned} f(\mathbf{z},\mathbf{w}) = &-\sum_{c=1}^{C} \eta_c \log\left[\mathrm{Dis}\left(\boldsymbol{\Omega}_{\boldsymbol{S},\boldsymbol{c}}, \boldsymbol{\rho}_{\boldsymbol{T}\to\boldsymbol{S}}\boldsymbol{\Omega}_{\boldsymbol{T},\boldsymbol{c}}\right)\right] \\ &-\sum_{c=1}^{C} \eta_c \log\left[\mathrm{Dis}\left(\boldsymbol{\rho}_{\boldsymbol{S}\to\boldsymbol{T}}\boldsymbol{\Omega}_{\boldsymbol{S},\boldsymbol{c}}, \boldsymbol{\Omega}_{\boldsymbol{T},\boldsymbol{c}}\right)\right], \end{aligned} \tag{5.3}$$

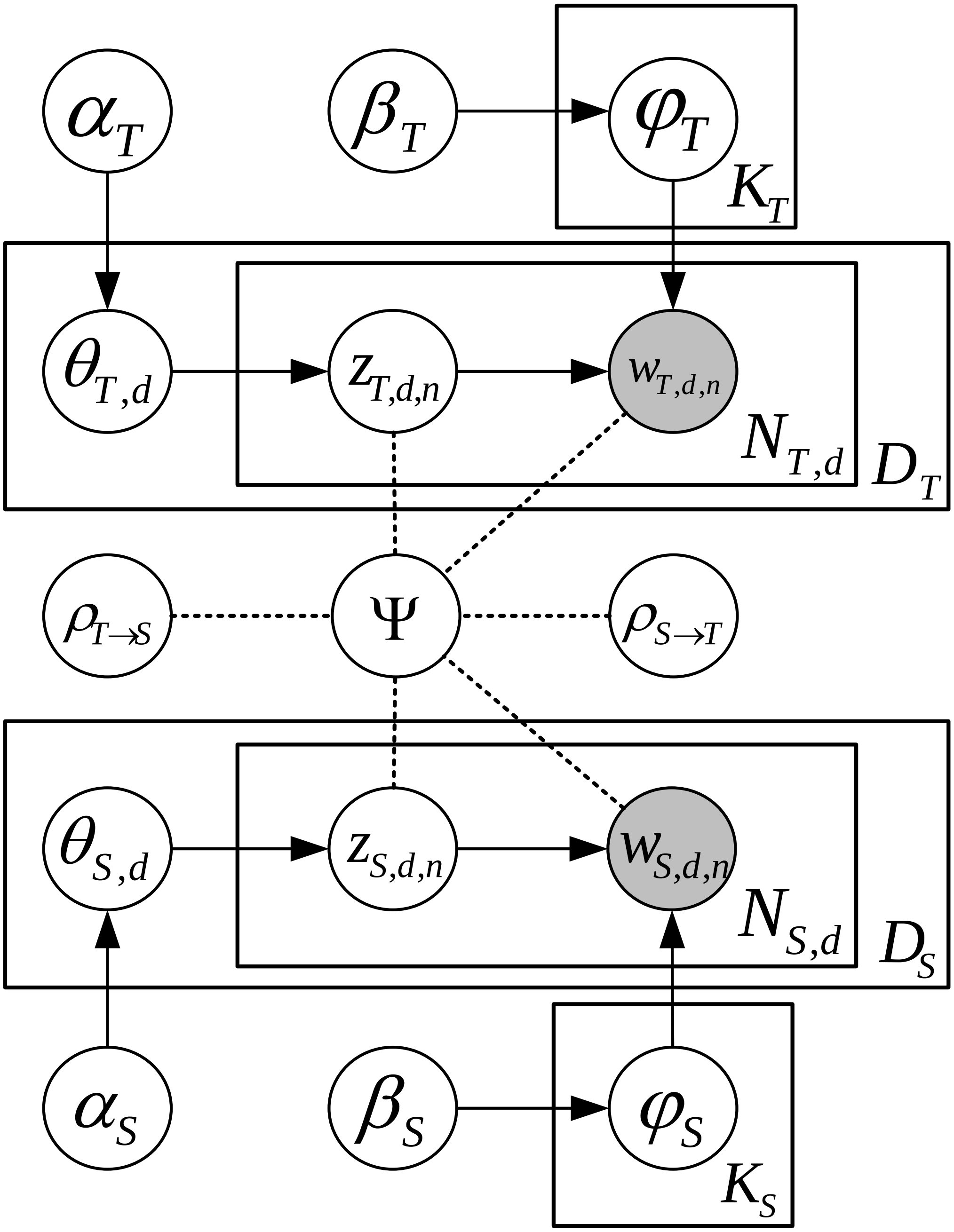


Figure 5.3: The graphical model of our multilingual topic model. The topic links $\boldsymbol{\rho}$'s, as instantiated by the function Ψ, encourage topics to encourage word translations to have consistent topic distributions.

This defines the posterior regularizer Ψ:

$$\Psi = \exp(f(\mathbf{z},\mathbf{w})) = \left(\prod_{c=1}^{C} \left[\mathrm{Dis}\left(\boldsymbol{\Omega_{S,c}}, \boldsymbol{\rho_{T\to S}}\boldsymbol{\Omega_{T,c}}\right)\right]^{\eta_c}\right)^{-1} \times \left(\prod_{c=1}^{C} \left[\mathrm{Dis}\left(\boldsymbol{\rho_{S\to T}}\boldsymbol{\Omega_{S,c}}, \boldsymbol{\Omega_{T,c}}\right)\right]^{\eta_c}\right)^{-1}, \tag{5.4}$$

where cross-lingual knowledge is encoded. We then use Ψ to connect the monolingual topic models to obtain a multilingual model (Figure 5.3):

1. For each topic $k \in \{1, \ldots, K_T\}$ in language T
   (a) Draw word distribution $\boldsymbol{\phi_{T,k}} \sim \text{Dirichlet}(\beta_T)$
2. For each document $d \in \{1, \ldots, D_T\}$ in language T
   (a) Draw topic distribution $\boldsymbol{\theta_{T,d}} \sim \text{Dirichlet}(\alpha_T)$
   (b) For each token $t_{T,d,n}$ in document d
      i. Draw a topic $z_{T,d,n} \sim \text{Multinomial}(\boldsymbol{\theta_{T,d}})$
      ii. Draw a word $w_{T,d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{T,z_{T,d,n}}})$
3. For each topic $k \in \{1, \ldots, K_S\}$ in language S
   (a) Draw word distribution $\boldsymbol{\phi_{S,k}} \sim \text{Dirichlet}(\beta_S)$
4. For each document $d \in \{1, \ldots, D_S\}$ in language S
   (a) Draw topic distribution $\boldsymbol{\theta_{S,d}} \sim \text{Dirichlet}(\alpha_S)$
   (b) For each token $t_{S,d,n}$ in document d
      i. Draw a topic $z_{S,d,n} \sim \text{Multinomial}(\boldsymbol{\theta_{S,d}})$
      ii. Draw a word $w_{S,d,n} \sim \text{Multinomial}(\boldsymbol{\phi_{S,z_{S,d,n}}})$
5. Draw the weighted topic distribution distance Ψ with Equation 5.4

As most downstream topic models, our MTM first generates the documents and tokens with two independent LDA components for languages S and T. Then it generates the posterior regularizer Ψ with weighted topic distribution distances. Although the two LDA components are separate in the generative process, they are correlated to incorporate the knowledge from the other language in posterior inference, as we will see in the next section.

5.2 Posterior Inference

The posterior inference is based on stochastic EM like other downstream topic models (Celeux, 1985). In each iteration, the E-step (Section 5.2.1) updates every token's topic assignment using Gibbs sampling, while fixing the values in the topic link weight matrices $\boldsymbol{\rho}$'s. The M-step (Section 5.2.2), on the other hand, optimizes $\boldsymbol{\rho}$'s while holding the topic assignments fixed.

5.2.1 E-step: Topic Assignment Sampling

Although topic generation looks independent from the posterior regularizer Ψ in the generative process, the topic assignment inference depends on Ψ (Equation 2.32). For our MTM, in addition to the usual word and topic dependencies, it encourages topic assignments that maximize the posterior regularizer Ψ, thus making the related translation pairs' (transformed) topic distributions close. This intuition is reflected in the Gibbs sampling equation to update $z_{T,d,n}$, the topic assignment of the n-th token of document d in language T:

$$\begin{aligned}
&\Pr\left(z_{T,d,n}=k \,|\, \mathbf{z}^{-\mathbf{T},\mathbf{d},\mathbf{n}}, w_{T,d,n}=v, \mathbf{w}^{-\mathbf{T},\mathbf{d},\mathbf{n}}, \boldsymbol{\rho}_{\boldsymbol{S}\to\boldsymbol{T}}, \boldsymbol{\rho}_{\boldsymbol{T}\to\boldsymbol{S}}, \alpha_T, \beta_T\right)\\
&\propto \underbrace{\left(N_{T,d,k}^{-T,d,n}+\alpha_T\right)\frac{N_{T,k,v}^{-T,d,n}+\beta_T}{N_{T,k,\cdot}^{-T,d,n}+V_T\beta_T}}_{\text{LDA Sampling}}\\
&\underbrace{\left(\prod_{v'\in\mathrm{Dic}(v)}\left[\mathrm{Dis}\left(\boldsymbol{\Omega}_{\boldsymbol{S},\boldsymbol{v'}}, \boldsymbol{\rho}_{\boldsymbol{T}\to\boldsymbol{S}}\boldsymbol{\Omega}_{\boldsymbol{T},\boldsymbol{v}}\right)\right]^{\eta_{v',v}}\right)^{-1}\left(\prod_{v'\in\mathrm{Dic}(v)}\left[\mathrm{Dis}\left(\boldsymbol{\rho}_{\boldsymbol{S}\to\boldsymbol{T}}\boldsymbol{\Omega}_{\boldsymbol{S},\boldsymbol{v'}}, \boldsymbol{\Omega}_{\boldsymbol{T},\boldsymbol{v}}\right)\right]^{\eta_{v',v}}\right)^{-1}}_{\text{Minimizing the Topic Distribution Distances}},
\end{aligned} \tag{5.5}$$

where the first two terms are the same as LDA: $N_{T,d,k}$ denotes the number of tokens in document d assigned to topic k; $N_{T,k,v}$ denotes the number of times word v is assigned to topic k; $\cdot$ denotes marginal counts; ${}^{-T,d,n}$ means the count excludes the token. The final term corresponds to the posterior regularizer: $\mathrm{Dic}(v)$ is word v's translation word set in language S; The values of $\boldsymbol{\Omega}_{\boldsymbol{T},\boldsymbol{v}}$, the topic distribution of word v, *assume* topic k is chosen as follows:

$$\Omega_{T,v,k'}=\frac{N_{T,k',v}^{-T,d,n}+\mathbb{1}\left(k'=k\right)}{N_{T,v}}, \tag{5.6}$$

where $\mathbb{1}\left(\cdot\right)$ is an indicator function. The Gibbs sampling Equation 5.5 prefers a topic, in addition to the usual constraints on co-occurrences, that can contribute more in minimizing the translation pairs' topic distribution distances after transformation by topic link weight matrices $\boldsymbol{\rho}$'s.

Similarly, the Gibbs sampling equation to update $z_{S,d,n}$, the topic assignment of the n-th token of document d in language S, is:

$$\begin{aligned}
&\Pr\left(z_{S,d,n} = k \,|\, \mathbf{z}^{-\mathbf{S,d,n}}, w_{S,d,n} = v, \mathbf{w}^{-\mathbf{S,d,n}}, \boldsymbol{\rho_{S\to T}}, \boldsymbol{\rho_{T\to S}}, \alpha_S, \beta_S\right) \\
&\propto \left(N_{S,d,k}^{-S,d,n} + \alpha_S\right) \frac{N_{S,k,v}^{-S,d,n} + \beta_S}{N_{S,k,\cdot}^{-S,d,n} + V_S\beta_S} \\
&\quad \left(\prod_{v'\in\mathrm{Dic}(v)} \left[\mathrm{Dis}\left(\boldsymbol{\Omega_{S,v}}, \boldsymbol{\rho_{T\to S}\Omega_{T,v'}}\right)\right]^{\eta_{v,v'}}\right)^{-1} \left(\prod_{v'\in\mathrm{Dic}(v)} \left[\mathrm{Dis}\left(\boldsymbol{\rho_{S\to T}\Omega_{S,v}}, \boldsymbol{\Omega_{T,v'}}\right)\right]^{\eta_{v,v'}}\right)^{-1}.
\end{aligned} \tag{5.7}$$

The values of $\boldsymbol{\Omega_{S,v}}$, assuming topic k is chosen, are

$$\Omega_{S,v,k'} = \frac{N_{S,k',v}^{-S,d,n} + \mathbb{1}\left(k' = k\right)}{N_{S,v}}. \tag{5.8}$$

The time complexity of inferring the topic assignment of a token v in language S is $O(K_S|\mathrm{Dic}(v)|K_SK_T)$. In the implementation, caching helps to reduce some repetitive computation: the values in $\boldsymbol{\rho_{T\to S}}$ and $\boldsymbol{\Omega_{T,v'}}$ (Equation 5.7) do not change when updating the tokens in language S, so we can pre-compute $\boldsymbol{\rho_{T\to S}\Omega_{T,v'}}$ and cache the values.

5.2.2 M-step: Parameter Optimization

In the M-step, we optimize the topic link weight matrices $\boldsymbol{\rho}$'s while fixing the topic assignments. As the posterior regularizer Ψ is the product over all translation pairs, we modify Ψ to obtain the objective functions $J(\boldsymbol{\rho_{T\to S}})$ and $J(\boldsymbol{\rho_{S\to T}})$ as the weighted logarithmic sums

$$\begin{aligned}
J(\boldsymbol{\rho_{T\to S}}) &= \sum_{c=1}^{C} \eta_c \log \sum_{i_S=1}^{K_S} \left(\Omega_{S,c,i_S} - \boldsymbol{\rho_{T\to S,i_S}\Omega_{T,c}}\right)^2 & (5.9) \\
J(\boldsymbol{\rho_{S\to T}}) &= \sum_{c=1}^{C} \eta_c \log \sum_{i_T=1}^{K_T} \left(\Omega_{T,c,i_T} - \boldsymbol{\rho_{S\to T,i_T}\Omega_{S,c}}\right)^2, & (5.10)
\end{aligned}$$

where the square root on the Euclidean distances is equivalent as a coefficient of 0.5 for the whole equation and thus dropped.[2]

The objective function is then minimized by using L-BFGS and the partial derivatives about $\rho_{T\to S,k_S,k_T}$ and $\rho_{S\to T,k_T,k_S}$ (Liu and Nocedal, 1989):

$$\frac{\partial J(\boldsymbol{\rho_{T\to S}})}{\rho_{T\to S,k_S,k_T}} = -\sum_{c=1}^{C} \frac{2\eta_c \Omega_{T,c,k_T}\left(\Omega_{S,c,k_S} - \boldsymbol{\rho_{T\to S,k_S}}\boldsymbol{\Omega_{T,c}}\right)}{\sum_{i_S=1}^{K_S}\left(\Omega_{S,c,i_S} - \boldsymbol{\rho_{T\to S,i_S}}\boldsymbol{\Omega_{T,c}}\right)^2} \tag{5.11}$$

$$\frac{\partial J(\boldsymbol{\rho_{S\to T}})}{\rho_{S\to T,k_T,k_S}} = -\sum_{c=1}^{C} \frac{2\eta_c \Omega_{S,c,k_S}\left(\Omega_{T,c,k_T} - \boldsymbol{\rho_{S\to T,k_T}}\boldsymbol{\Omega_{S,c}}\right)}{\sum_{i_T=1}^{K_T}\left(\Omega_{T,c,i_T} - \boldsymbol{\rho_{S\to T,i_T}}\boldsymbol{\Omega_{S,c}}\right)^2}. \tag{5.12}$$

5.3 Experimental Results

We first evaluate our model extrinsically by intra- and cross-lingual classification tasks with topic posteriors as features. Then we look into the model's intrinsic performance of topic coherence on five bilingual corpora when the corpora get less comparable and even incomparable. We also study how the topic coherence changes when the sizes of target language (non-English languages) corpora vary.

For the translation pair weighting, we explore equal weights and TF-IDF weights. A translation pair's TF-IDF weight is decided by the lower TF-IDF weight of the two words, based on the intuition that if a word is less important or reliable (i.e., of low TF-IDF weight), its information is less likely to be accurate and makes the whole pair less reliable.

[2] It makes sense to add regularization on $\boldsymbol{\rho}$'s to prevent overfitting, but the data already adds a strong constraint on $\boldsymbol{\rho}$'s—each word's $\boldsymbol{\Omega}$ values should add up to one.

Dataset	Lang. Pair	Lang.	#Docs	#Tokens	#Vocab.	#Trans.
Wikipedia	EN-ZH	EN	11,043	1,906,142	13,200	6,812
		ZH	10,135	1,169,056	13,972	
LORELEI	EN-SI	EN	1,100	32,714	6,920	6,330
		SI	4,790	168,082	31,629	

Table 5.1: Statistics of the bilingual corpora used in classification experiments with inferred topic posteriors. For Wikipedia, the task is to classify each document into one of six categories. For LORELEI, the goal is to distinguish the need of *evacuation* from other need types.

5.3.1 Classification with Topic Posteriors

We take two datasets for the classification experiments (Table 5.1). The first dataset contains Wikipedia documents in English (EN) and Chinese (ZH) (Yuan et al., 2018). Each document is labeled with one of six categories of *film*, *music*, *animals*, *politics*, *religion*, and *food*. The English-Chinese word translation dictionary is collected from MDBG, a website for learning Chinese.[3]

The second dataset is the Sinhalese (SI) language pack from the Low Resource Languages for Emergent Incidents (LORELEI) Program (Strassel and Tracey, 2016). The program aims to develop human language technology to identify emergent situations (e.g., earthquake, flood, war, etc.) and needs (e.g., shelters, medicine, food, etc.) at the regions where low-resource languages are frequently used in formal and/or informal media. It will support the government and other organizations in emergent missions such as humanitarian assistance, disaster relief, peacekeeping, or infectious disease response. The Sinhalese language pack contains documents related to disasters in both English and Sinhalese and a small subset of them are

[3] https://www.mdbg.net/chinese/dictionary?page=cc-cedict

annotated with one of the eight need types: *evacuation*, *food supply*, *search/rescue*, *utilities*, *infrastructure*, *medical assistance*, *shelter*, and *water supply* (Strassel et al., 2017). The dictionary comes along with the language pack.

We follow the same preprocessing mechanism with Yuan et al. (2018) and use SVM with a linear kernel for classification. For the Wikipedia dataset, we classify and report micro-F1 scores. For the LORELEI dataset, our goal is to distinguish the need for *evacuation* from other types.

We compare our model against several multilingual baselines, including tree LDA (Hu et al., 2014, tLDA) which encodes the dictionary as a tree prior (Boyd-Graber et al., 2007), Multilingual Topic Anchoring (Yuan et al., 2018, MTAnchor), and Multilingual Cultural-common Topic Analysis (Shi et al., 2016, MCTA). We also include LDA which runs monolingually on each language and a naïve baseline of most frequent class (MFC). For all the models, we set the number of topics at twenty and hyper-parameters $\alpha = 0.1$ and $\beta = 0.01$ (if applicable).

Our evaluations are both intra- and cross-lingual. The intra-lingual (IN) evaluation trains and tests the classifiers on the same language, while the cross-lingual (CR) evaluation trains the classifiers on English (Sinhalese/Chinese) and tests on Sinhalese/Chinese (English). In cross-lingual evaluations, MTAnchor, MCTA, and tLDA assume aligned topic spaces, so there is no need to convert the topic posteriors in different languages. LDA cannot transform topic spaces, so we do not apply any transformation and directly feed English topic posteriors to a Singhalese/Chinese classifier and vice versa.

For our MTM, we explore two methods to use the topic link weight matrices $\boldsymbol{\rho}$'s

Dataset	Method	EN-IN	SI/ZH-IN	EN-CR	SI/ZH-CR
LORELEI	MFC	14.46	15.09	14.46	15.09
	MCTA	12.99	26.53	4.08	15.58
	MTAnchor	20.78	32.65	24.49	24.68
	LDA	27.78	24.01	22.86	21.05
	tLDA	12.77	18.18	16.01	15.09
	MTM	**42.86**	23.08	22.22	26.67
	MTM + TOP	**42.86**	23.08	**35.29**	**33.33**
	MTM + TF-IDF	26.67	**38.10**	14.46	15.09
	MTM + TF-IDF + TOP	26.67	**38.10**	14.46	11.43
Wikipedia	MFC	16.52	20.82	16.93	17.32
	MCTA	51.56	33.35	23.24	39.79
	MTAnchor	80.71	75.33	57.62	54.54
	LDA	92.08	83.37	16.52	10.46
	tLDA	91.58	83.33	2.85	21.02
	MTM	92.98	**86.48**	74.69	64.48
	MTM + TOP	92.98	**86.48**	**78.13**	**83.08**
	MTM + TF-IDF	**94.07**	85.59	57.27	55.06
	MTM + TF-IDF + TOP	**94.07**	85.59	63.20	59.64

Table 5.2: Our MTM outperforms all the baseline models in both intra-lingual (IN) and cross-lingual (CR) evaluations in F1 scores. Connecting the top linked topics is a better way for topic transformation.

for topic space transformation. The first is to directly multiply $\boldsymbol{\rho}$ with a language's document topic distributions, i.e., $\boldsymbol{\rho}_{\text{ZH}\rightarrow\text{EN}}\boldsymbol{\theta}_{\text{ZH}}$ and vice versa. The other one, which we call top-linked topics (TOP), to take Chinese as an example, is to transfer each Chinese document's topic k's probability mass to the English topic which has the highest topic link weight with the Chinese topic:

$$\theta_{\text{EN},d,k_0} \leftarrow \theta_{\text{EN},d,k_0} + \theta_{\text{ZH},d,k}, \text{ where } k_0 = \arg\max_{k'} \rho_{\text{ZH}\rightarrow\text{EN},k',k}. \tag{5.13}$$

5.3.1.1 Classification Results

Our MTM performs better than baseline models both intra- and cross-lingually (Table 5.2).[4] TF-IDF weighting on translation pairs sometimes improves the intra-lingual performance, although it hurts the cross-lingual performance. In topic space transformation, connecting the top linked topics (TOP) is better than directly using the topic link weight matrices. This indicates that the values in $\boldsymbol{\rho}$'s have some noise and is worth further exploration.

5.3.2 Learned Topics

To show how the learned topics differ across models, we pick the English (EN) and Chinese (ZH) Movies topics from the Wikipedia dataset (Table 5.3). The English translations of Chinese words are given in brackets following the Chinese words. For each Chinese topic given by MTM, we attach the top three English topics with the highest topic link weights.[5]

The topics are about Movies, but the MCTA and MTAnchor topics do not rank “movie” or “电影 (diàn yǐng)” at the top. The tLDA topics, although aligned well, have some problems with the Chinese words. The word “胶片 (jiāo piàn)”, although its English translation is “film”, its actual meaning is not “movie” but “photographic film”. Another word, “释放 (shì fàng)”, corresponds to the sense of “let something go” for “release”, not “movie distribution”. tLDA links the words

[4] The performance of MTAnchor and MCTA are from Yuan et al. (2018).

[5] For each Chinese/English topic, its link weights to all English/Chinese topics sum up to one.

Model	Language	Words
MCTA	ZH	主演 (starring), 改编 (adapt), 本 (this), 小说 (novel), 拍摄 (shoot), 角色 (role), 战士 (fighter)
	EN	dog, san, movie, mexican, fighter, novel, california
MTAnchor	ZH	主演 (starring), 改编 (adapt), 饰演 (act), 本片 (this movie), 演员 (actor), 编剧 (playwright), 讲述 (narrate)
	EN	kong, hong, movie, official, martial, box, reception
LDA	ZH	电影 (movie), 部 (movie quantifier), 美国 (USA), 上映 (release), 英语 (English), 剧情 (plot), 片 (movie)
	EN	film, star, direct, release, action, plot, character
tLDA	ZH	电影 (movie), 胶片 (film), 星 (star), 动作 (action), 释放 (release), 影片 (movie), 剧情 (plot)
	EN	film, star, direct, action, release, plot, write
MTM	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 动画 (animation), 故事 (story), 作品 (works), 英语 (English)
	EN-1 (0.20)	film, direct, star, release, action, plot, production
	EN-2 (0.12)	kill, find, death, attack, escape, return, back
	EN-3 (0.11)	shrine, japanese, temple, japan, shinto, kami, god
MTM + TF-IDF	ZH	电影 (movie), 部 (movie quantifier), 上映 (release), 美国 (USA), 英语 (English), 导演 (director), 片 (movie)
	EN-1 (0.32)	film, direct, star, action, release, plot, movie
	EN-2 (0.24)	film, kill, find, escape, attack, return, back
	EN-3 (0.09)	character, series, star, game, trek, create, episode

Table 5.3: The topics of Movies given by models. For each Chinese topic given by our MTM, the top three English topics and their link weights are also given, while the link weights to all English topics sum up to one.

based on translations without looking at the context, which causes problems with multiple-sense words.

The LDA and MTM topics are generally coherent, despite slight differences in words and ordering. A unique output of our MTM is the weighted topic links. For the Chinese Movies topics given by our MTMs, the most relevant English topics are also about Movies, e.g., “film”, “direct”, “star”, and “release”. The second relevant topics have the words “kill”, “death”, “attack”, and “escape” which often appear in action movies. For the third relevant topics, the MTM model gives the words

Language	Weight	Words
ZH-0	–	学名 (scientific name), 它们 (they), 呈 (show), 白色 (white), 长 (long), 黑色 (black), 厘米 (centimeter)
EN-12	0.57	specie, bird, eagle, genus, white, owl, black
EN-19	0.13	breed, chicken, white, goose, bird, black, list
ZH-14	–	主义 (-ism)[6], 组织 (organization), 美国 (USA), 革命 (evolution), 运动 (campaign), 政府 (government), 人民 (people)
EN-16	0.32	sex, law, act, sexual, marriage, court, legal
EN-11	0.17	traffic, victim, government, trafficking, child, force, country
EN-1	–	abortion, government, report, muslim, death, arrest, iran
ZH-15	0.16	伊斯兰 (Islam), 穆斯林 (muslim), 伊斯兰教 (Islam), 阿拉伯语 (Arabic), 阿拉伯 (Arab), 世纪 (century), 帝国 (empire)
ZH-4	0.13	主义 (-ism), 社会 (society), 历史 (history), 文化 (culture), 发展 (develop), 研究 (research), 哲学 (philosophy)
EN-10	–	album, release, record, music, song, single, feature
ZH-9	0.30	专辑 (album), 张 (album quantifier), 发行 (release), 音乐 (music), 首 (song quantifier), 唱片 (record), 歌手 (singer)
ZH-17	0.20	音乐 (music), 乐团 (musical group), 艺术 (art), 创作 (create), 奖 (prize), 演出 (perform), 担任 (serve)

Table 5.4: Topics are linked because they have overlap in topical words. Although explicit word translations can help identify related topics, our MTM can also infer the topic relations beyond word translations, e.g., ZH-14 and EN-16 which have no overlap in words.

about Japanese cartoons, while the MTM with TF-IDF gives a Games topic which has some overlap with Movies, like “character”, “series”, and “episode”. Generally, the top three English topics all have overlap with the Chinese topic. One is a good match, while the other ones overlap with the Chinese topic in some perspective, as can be seen from the top words and reflected in the topic link weights. This shows that our MTM can link topics as long as they have some mutual perspective and represent it in the link weights.

5.3.3 Learned Topic Links

[6] -ism is a word suffix that denotes a system, principle, or ideological movement, e.g., terrorism,

We give more examples of cross-lingually linked topics and their weights in Table 5.4. The MTM assigns high weights to the topics which have more cross-lingual common words (as indicated by the dictionary). For instance, the words “white” and “black” in the Biology topics of ZH-0, EN-12, and EN-19. This is also the case for the Music topics of EN-10, ZH-9, and ZH-17.

Our MTM can also infer topic links beyond words. When the topical words have few direct translations but are related in senses, the MTM is still able to link them. ZH-14 is about the “campaigns” of “organizations” for “people” and against “government”, e.g., the Weather Underground Organization which ran campaigns against the US Government.[7] It has only one overlap word “government” with EN-16 and EN-11. However, MTM identifies the two English topics as the top linked topics for ZH-14: EN-16 is about the “campaign” in Sexual Rights, e.g., Campaign for Homosexual Law Reform in Ireland;[8] EN-11 talks about Crime with an emphasis on human trafficking, e.g., human trafficking in various countries.[9] This indicates that our MTM can incorporate the word translations and infer more cross-lingual word and topic relationships.

It also happens to the topics of EN-1 and ZH-4. EN-1 is about “abortion”

capitalism, and socialism.

[7] Chinese source page: https://zh.wikipedia.org/wiki/%E5%9C%B0%E4%B8%8B%E6%B0%A3%E8%B1%A1%E5%93%A1; English source page: https://en.wikipedia.org/wiki/Weather_Underground.

[8] https://en.wikipedia.org/wiki/Campaign_for_Homosexual_Law_Reform

[9] Human trafficking in Luxembourg: https://en.wikipedia.org/wiki/Human_trafficking_in_Luxembourg; Human trafficking in Slovenia: https://en.wikipedia.org/wiki/Human_trafficking_in_Slovenia.

in "muslim", e.g., abortion in Iran.[10] It is part of the "society", "history", and "culture" in ZH-4, e.g., Islamic Golden Age.[11] The two topics' top words do not have overlap either but are linked, although their overlap is limited, which is also reflected from the topic link weight.

5.3.4 Topic Coherence on Less Comparable Corpora

We intrinsically evaluate the models' intra-lingual topic coherence on two Wikipedia corpora with low comparability (Hao and Paul, 2018, Table 5.5). Each contains five bilingual corpora where one of the languages is always English, while the other ones are Arabic (AR), Chinese (ZH), Spanish (ES), Farsi (FA), and Russian (RU). Each bilingual corpus contains around 2,000 documents for both languages. The first Wikipedia corpora are partially comparable (PACO), where 30% of the documents have direct translations in the other language. The second corpora are incomparable (INCO)—no documents have direct translations. Dictionaries are extracted from Wiktionary.[12]

As the corpora are not highly comparable, their topic distributions differ substantially. Thus evaluation metrics for cross-lingual topic alignment are not good choices (Hao et al., 2018). We instead take an intra-lingual topic coherence metric (Lau et al., 2014, Section 2.2): for every topic, we extract the top N words and

[10] https://en.wikipedia.org/wiki/Abortion_in_Iran

[11] Chinese source page: https://zh.wikipedia.org/wiki/%E4%BC%8A%E6%96%AF%E8%98%AD%E9%BB%83%E9%87%91%E6%99%82%E4%BB%A3; English source page: https://en.wikipedia.org/wiki/Islamic_Golden_Age.

[12] https://dumps.wikimedia.org/enwiktionary/

Dataset	Lang. Pair	Lang.	#Docs	#Tokens	#Vocab.	#Trans.
PACO	EN-AR	EN	1,999	622,955	47,790	4,384
		AR	1,999	107,434	19,900	
	EN-ZH	EN	2,000	405,976	39,847	8,691
		ZH	1,997	86,585	30,481	
	EN-ES	EN	2,000	238,092	30,278	18,221
		ES	2,000	188,469	27,465	
	EN-FA	EN	2,000	513,855	41,685	4,419
		FA	1,814	37,158	9,987	
	EN-RU	EN	1,999	296,148	34,618	2,981
		RU	1,999	101,922	24,341	
INCO	EN-AR	EN	2,000	581,473	45,444	4,380
		AR	1,999	107,434	19,900	
	EN-ZH	EN	2,000	432,442	38,369	8,766
		ZH	1,997	86,585	30,481	
	EN-ES	EN	1,999	557,602	46,161	20,954
		ES	2,000	188,469	27,465	
	EN-FA	EN	2,000	324,858	34,278	4,280
		FA	1,814	37,158	9,987	
	EN-RU	EN	2,000	547,748	47,167	3,345
		RU	1,999	101,922	24,341	

Table 5.5: Statistics of the corpora for topic coherence evaluation. 30% of the documents in the partially comparable (PACO) corpora have direct translations in the other language, while no documents in the incomparable (INCO) corpora have direct translations.

compute the average pairwise PMI score on a reference corpus of a disjoint subset of Wikipedia documents (Hao and Paul, 2018).

We report the average coherence scores on five-fold cross-validation with values of N from 10 to 100 with a step size of 10. For the weighting on translation pairs, we take the same options as we do in classification tasks. For the baseline models, we choose monolingual LDA and tree LDA which encodes word translations in its tree prior (Boyd-Graber et al., 2007; Hu et al., 2014).

Our MTM mostly matches LDA in topic coherence and sometimes slightly better (Figures 5.4 and 5.5). TF-IDF weighting on translation pairs sometimes

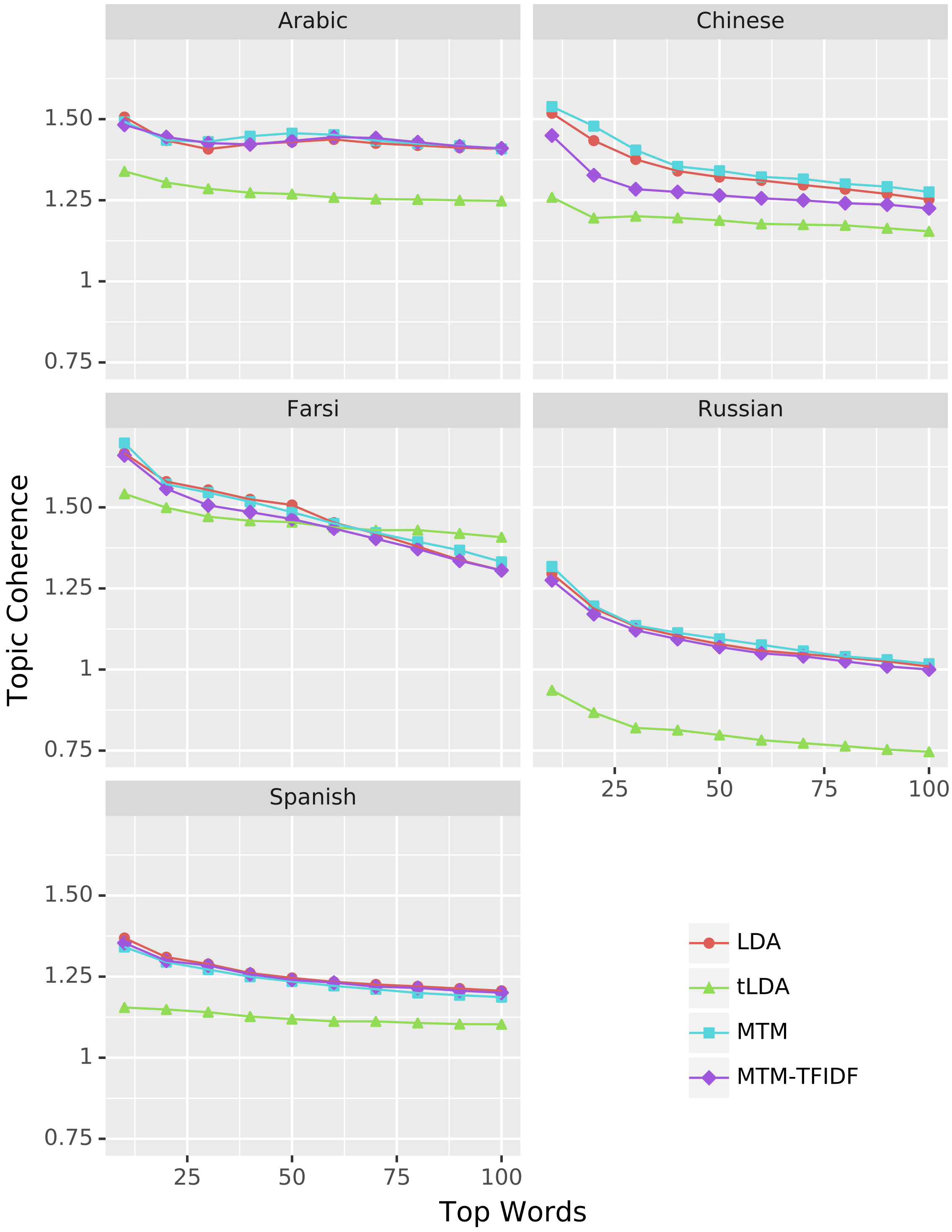


Figure 5.4: Topic coherence performance on PACO dataset with the number of top words in each topic.

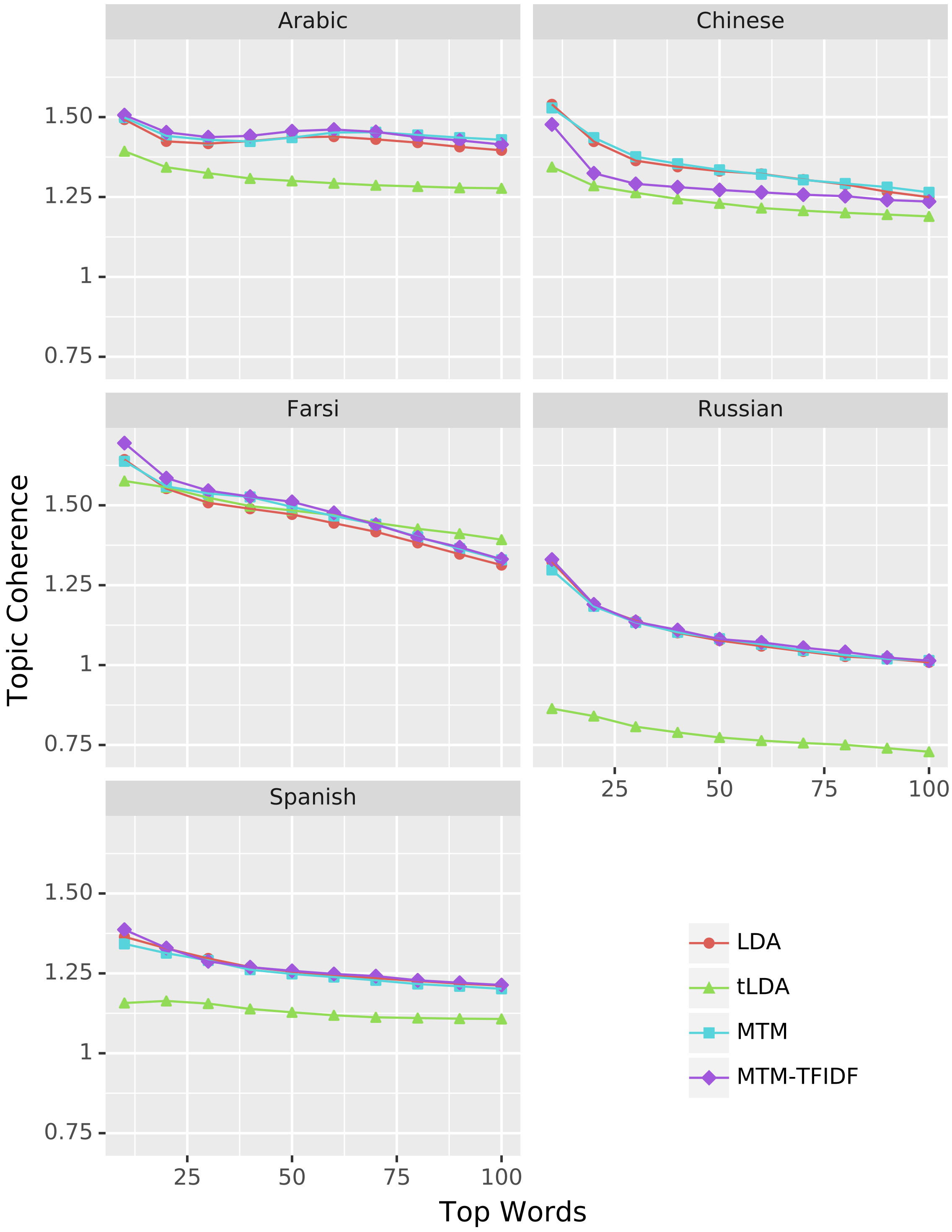


Figure 5.5: Topic coherence performance on INCO dataset with the number of top words in each topic.

further improves the topic coherence a little bit (Arabic, Farsi, Russian, and Spanish on INCO) but occasionally hurts (Chinese).

The scores on the PACO dataset are generally close to the ones on the INCO dataset according to the figures, but PACO scores are slightly higher numerically. In the PACO dataset, 30% of the documents have direct translations in the other language. It makes the topic space more aligned than the INCO dataset and provides more accurate topic information for each translation pair. Thus it is easier to achieve higher topic coherence scores on the PACO dataset.

Another baseline, tLDA, mostly works poorly, except on Farsi with a high number of top words. tLDA always tries to infer an aligned topic space for both languages, which is hard when the corpora are not comparable. To exchange for topic alignment, tLDA has to sacrifice the topic coherence on individual languages. Our MTM only connects topics when necessary, so it is more robust when the corpora get less comparable.

The results prove our MTM's robustness on low comparability data, on which it is likely to fail when forcing topic spaces to be aligned across languages. Our MTM only connects topics when necessary, thus can still give coherent topics like monolingual LDA.

5.3.5 Topic Coherence with Various Target Language Corpora Sizes

We study how the topic coherence changes when we vary the sizes of target language (non-English languages in PACO and INCO) corpora, to find out how much

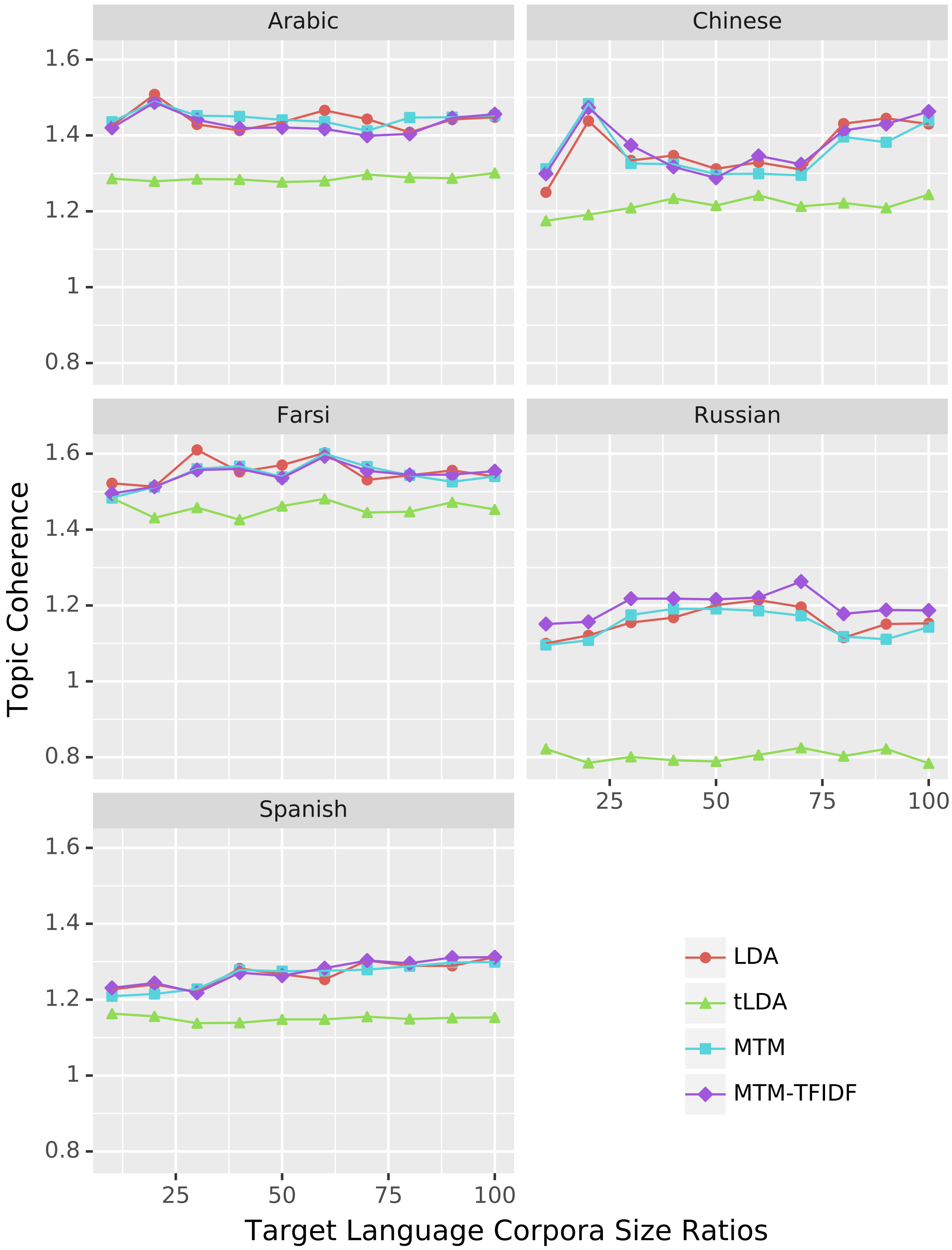


Figure 5.6: The models' performance of topic coherence on PACO dataset when the sizes of target language corpora grow from 10% to 100%, with a step size of 10%.

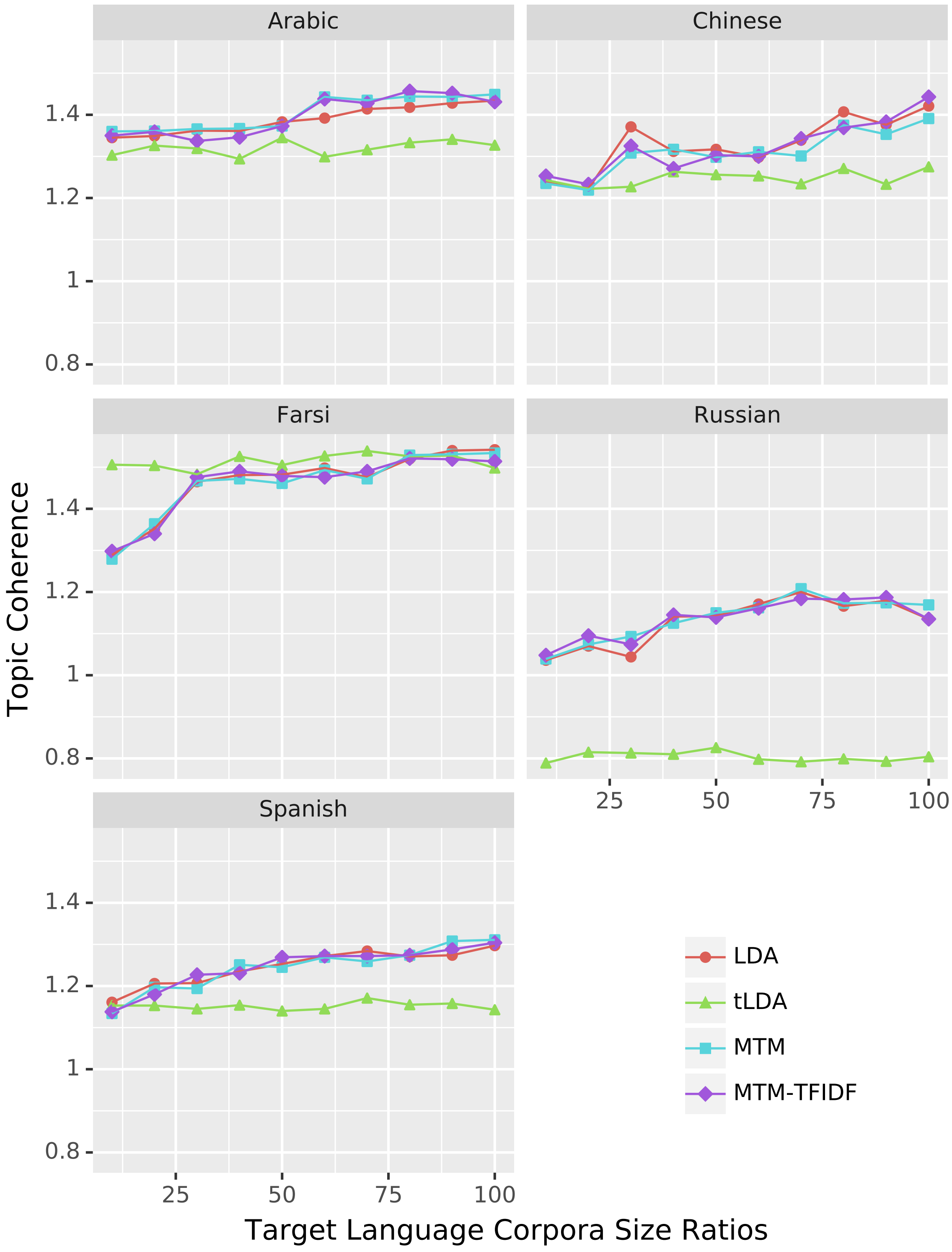


Figure 5.7: The models' performance of topic coherence on INCO dataset when the sizes of target language corpora grow from 10% to 100%, with a step size of 10%.

our MTM can help when the data is limited in the case of low-resource languages. Specifically, we start from 10% of the randomly-selected documents in target languages and incrementally add more target language documents at a step size of 10% until it reaches 100%. So take Arabic as an example, the data composition settings are (100% English, 10% Arabic), (100% English, 20% Arabic), until (100% English, 100% Arabic). We train monolingual LDA, tLDA, and MTMs with and without TF-IDF weighting on translation pairs on each setting, evaluate the topic coherence on the same reference corpora using the top thirty words of each topic and present them in Figures 5.6 and 5.7.

In most cases, the topic coherence gets better when the sizes of target language corpora enlarge, except a few cases like Arabic and Russian on PACO. This meets our intuition that with more available data, it is easier to train a better topic model. MTM is helpful in some cases when the target language corpora sizes are small, e.g., Chinese and Russian with 10% or 20% sizes of the corpora. In terms of TF-IDF weighting, there is no consistent result whether it is better than equal weights.

The tLDA with tree priors of dictionaries performs poorly in topic coherence, except Farsi in INCO. In most cases, its performance is way below other ones' and improves little when the target corpora sizes grow.

5.4 Summary

In this chapter, we focus on learning, instead of incorporating, topic links across languages, given the latent nature of topics. Thus, we introduce a novel multilingual topic model (MTM) which learns weighted topic links across languages. The MTM allows the topics in different languages to be connected *only* when necessary, based on the observations that topics often differ across languages and cultures, and even the same topic can have different emphases among languages. The topic link weights are learned by minimizing the Euclidean distances of translation pairs' (transformed) topic distributions, where each translation pair can be weighted, e.g., by TF-IDF.

Our MTM significantly outperforms baseline models in classification tasks both intra- and cross-lingually, while providing coherent topics and meaning topic links that can go beyond word translations. When the data get small and less comparable or even incomparable, our MTM still performs well or slightly better than monolingual LDA in topic coherence. This shows its robustness over past MTMs that force topic spaces to be aligned across languages.

Chapter 6: Conclusion and Future Work

Weighted links exist almost everywhere and connect objects with similar patterns. They contain rich information and could assist in various tasks. In the hierarchy of documents, topics, and words in topic modeling, weighted document links and word links are often observed and could provide external knowledge for topic modeling, while topic links are usually unobservable. This dissertation follows the induction and deduction insights to summarize the patterns from the weighted links and apply them in topic inference. Specifically, we explore the methods to uncover the latent structures in the observed weighted document and word links and suggests ways to incorporate them into topic modeling. For latent weighted topic links, we introduce a multilingual topic model to learn them across languages from word translations.

This dissertation develops methods based on topic models, contrast to the popular deep learning methods in the fields of natural language processing and machine learning. Although deep learning methods can also incorporate the external knowledge into its objective function and neural network structure pretty straightforwardly, topic models have their unique advantage of *interpretability*. Due to the non-linearity in neural networks, it is hard to pinpoint the bottleneck of deep learning algorithms. People can only try different network structures, tune the hyperparameters, and wish a better model (so some people call deep learning “a modern alchemy”). In topic models, on the contrary, every parameter is clearly interpretable, which is easy to diagnose and provides useful insight of the documents, so people like to use topic models to analyze documents, although it is more difficult to add a new distribution and derive the posterior inference than to add another layer in a neural network.

6.1 Summary of Contributions

In Chapter 3, we explore binary document links which indicate connected documents' topic similarities. Past methods either treat text and links separately or treat them jointly but focus exclusively on single links without delving into the latent structure of links. They have ignored interesting patterns in the document network, such as latent blocks, in which documents are densely connected and tend to be about similar topics. We use WSBM, a probabilistic block discovery algorithm, to find the latent blocks and extract informative topic priors from the blocks to guide documents' topic sampling. To make full use of the features we have, in addition to topical features in past methods, we also include lexical and block features for link prediction. Further, we employ hinge loss for classification, which better captures the side information than sigmoid loss. The model LBH-RTM, a relational topic model with lexical weights, block priors, and hinge loss, achieves better performance than RTM in both link prediction and topic coherence.

In Chapter 4, we incorporate real-valued word links, or lexical associations, that indicate word semantic relatedness to guide topic modeling. Because the lexical association table is too large and redundant to be incorporated, we introduce three straightforward but effective tree prior construction algorithms to remove the redundancy and build the word hierarchies. The hierarchies contain the most salient word association information and/or encode the magnitude. The tree priors are then fed to tree LDA and help improve topic coherence and enhance extrinsic classification performance. Although the tree priors in the experiments are constructed on two particular types of word associations, the flexibility of our construction algorithms accommodates *any* word associations.

In Chapter 5, we introduce weighted topic links that connect topics *across* languages. We also introduce a novel multilingual topic model (MTM). Given that topics often differ among languages and background cultures, unlike past MTMs that learn an aligned topic space across languages, our MTM only links topics (i.e., assigns a topic link value) when the two topics contain many word translation pairs. This substantially improves the performance of our MTM. Not only does it achieve higher F1 scores in intra- and cross-lingual classification tasks than monolingual LDA and past MTMs, but it also stays robust and gives coherent topics when the data get smaller and less comparable or even incomparable (when past MTMs have mostly failed altogether). In addition, our MTM uses a posterior regularizer to encode external knowledge and learn topic links. This flexibility allows us to try out *any* other formulations without changing the main model structure.

6.2 Directions for Future Work

In this subsection, we analyze some limitations of the work in this dissertation and propose corresponding solutions. In addition, we give some smaller-scope future directions which may improve the current work.

6.2.1 Primary Limitations and Solutions

We evaluate the topics both quantitatively and qualitatively. While quantitative evaluation is quite objective, qualitative evaluation relies a lot on the human evaluators who can give very different results. For instance, the Images topic in Table 4.3 may be identified as an Image Transfer topic and the current less relevant words fit well. To minimize the variance in human evaluation, we can have multiple human evaluators evaluate the topics and take the majority topic name. We can also refer to the documents with high posteriors in this topic and see what the topic should be according the documents' content.

When we identify documents' block assignments in Chapter 3, we assume that each document can be assigned to exactly one block. However, this sometimes does not conform to reality. To take scientific paper as an example, today there is more interdisciplinary work than ever before. For instance, for a paper that applies a topic model on images, it makes sense to assign it to either the topic model block or the computer vision block (Fei-Fei and Perona, 2005), but either option discards some useful information in the other one and may cause our model to perform less well. To overcome this problem, we should break the assumption of one-block assignment and instead assume mixed membership for each document (Kim and Leskovec, 2012). In this case, we can obtain more accurate priors for a document by taking a weighted sum or average of the block topic priors it belongs to.

Another limitation is to deal with multiple types of networks. On Twitter, users interact with each other by mentioning, retweeting, and following. In the real world, people have different facets to others, e.g., colleagues, family, and strangers (Goffman, 1978). Different types of links imply different relationships. For instance, mentioning often indicates a closer relationship than following; people talk with family members more than strangers in both depth and breadth. A straightforward solution for this situation would be to identify latent blocks separately for each network, and then take a weighted sum or average of the priors from all networks.

In Chapter 4, we build *static* tree priors that never change the structures during topic model training. The topic modeling thus may suffer or even fail from domain differences between the task corpus and the external corpus on which the lexical associations are learned. For instance, in the experiments, we build the tree priors on Gigaword 5, which consists of news articles, but we apply the tree priors to Amazon reviews, which have significantly different word distributions. Therefore, it is worthwhile to develop dynamic tree priors for tLDA, e.g., based on probabilistic hierarchical clustering with coalescent (Teh et al., 2007; Görür and Teh, 2009; Hu et al., 2013). In this case, the tree prior construction and topic modeling could be in a joint framework, each adjusting according to the patterns of the other, and thereby improving each other.

For our MTM in Chapter 5, we learn weighted topic links by converting words' topic distribution to their translations'. This is fine for the words with single sense, but the words with multiple senses may mislead the learning of topic links. For example, the word “spring” has high probabilities in the topics of Season and Water and its translations in Chinese are “春天 (chūn tiān)” and “泉水 (quán shuǐ)” (which correspond to Season and Water senses for “spring” respectively). Our topic link learning method then tries to connect the English Season and Water topics with the corresponding topics in Chinese. In this case, the English Season topic is connected to Chinese Season *and* Water topics, and so is the English Water topic, which produces wrong connections and will mislead the MTM. This problem also exists in other cross-lingual methods, e.g., in cross-lingual word embeddings, “spring” may not align well with either “春天 (chūn tiān)” or “泉水 (quán shuǐ)” (Fujinuma et al., 2019). To avoid this problem, we can add some heuristics when learning weighted topic links. A straightforward heuristic for a pair of topics is the number of translation pairs in their top words. The topic link weight learning algorithm is then penalized more if it does not connect the topics with more translation pairs but penalized less otherwise. In the example of “spring”, the heuristics can tell the learning algorithm that the English Season topic should be connected to the Chinese Season topic because they have more translation pairs in their top words, and so for the Water topics, but not connect English Season topic with Chinese Water topic or vice versa.

In addition, the datasets used in our experiments are relatively small, but their sizes are enough to validate our models. However, when applying our models on large datasets, it will probably take a long time for convergence. To improve the runtime performance, we can apply some approximation to the Gibbs sampling equation using the distribution sparsity, alias table, and Metropolis-Hastings sampling algorithm (Yao et al., 2009; Li et al., 2014; Yuan et al., 2015). Another approach is to use variational inference (Wainwright and Jordan, 2008) instead of Gibbs sampling for posterior inference and parallelize it (Zhai et al., 2012).

6.2.2 Other Future Directions

More Fine-grained Evaluation of Document Link Prediction In the experiments of document link prediction, we evaluate our model's performance on paper abstracts with citations and web pages with hyperlinks. The ground-truth links contain only the links that should and do exist, but there is a chance that some "good" links are missing, e.g., a missing citation.

In the information retrieval literature, retrieved documents are categorized into three classes: 1) documents relevant to the query, 2) irrelevant to the query, and 3) partially relevant to the query (Voorhees, 2001). Following this classification, we can categorize the links into three classes as well. The first two classes correspond to the notion of explicit positive and negative links in Chapter 3. These entail more effort in data annotation but allow us to conduct a more fine-grained evaluation on document link prediction, e.g., treating link prediction as a multi-class classification problem and evaluating the F1 scores on each class.

Document Link Suggestion The missing document links suggest a potential application for our LBH-RTM for document link suggestion. Because our model gives link probabilities instead of binary values for a link, we can use it to suggest links that do not exist but should exist. This can be useful to bootstrap the data—the model learns the topics with a few links and can then add the links with high probabilities (self-training) or selected by a human (human-in-the-loop).

Other Document Link Weights We deal with binary-valued document links which are quite pervasive. However, it is possible that the document links have integer or even real-valued weights. This requires us to generalize the combined weighted stochastic block model by replacing the probability distributions for generating links, e.g., using the distributions from the exponential family. This can make our model more widely applicable.

Weighting Methods for Topic Translation Pairs In Chapter 5, we use TF-IDF to weight the words, taking the smaller one of a translation pair as the pair's weight. This leaves many possibilities for evaluating the importance and reliability of translation pairs, including but not limited to the variations of TF-IDF (e.g., just term frequencies or inverse document frequencies), Okapi BM25 (Robertson et al., 1995, 1999), and other methods to combine the two words' scores into one for the pair. Hopefully, we can find a better metric to more accurately weight translation pairs and then improve the MTM's performance.

Topic Space Transformation with Topic Link Weight Matrices In topic space transformation, we explore two methods: directly multiply the topic link weight matrices with topic distributions and transfer the topic probability mass to the top-linked topic in the other language. This corresponds to two extremes—the first method takes all information including noise, while the second one throws away most information except the most confident part. It makes sense to find a balance point so that we can reduce the noise while keeping useful information. The new method could be entropy-based. For the topics which have high entropies in the weight distributions over the topics in the other language, they are likely to be unique topics, so their weights could be adjusted lower. By contrast, a topic's weight could be increased if it has a precise corresponding topic in the other language and its weight distribution has low entropy.

Appendix A: Derivation of the Posterior Inference for LBH-RTM

This appendix gives more details of deriving the Gibbs sampling equations and parameter optimization for LBH-RTM in Section 3.3.

A.1 Sampling Block Assignments

In a weighted stochastic block model (WSBM, Section 3.1.2), the joint probability of *all* link weights $\mathbf{A}$ and document block assignments $\mathbf{y}$ is

$$\Pr\left(\mathbf{A}, \mathbf{y} \,|\, a, b, \gamma\right) = \Pr\left(\mathbf{A} \,|\, \mathbf{y}, a, b\right) \Pr\left(\mathbf{y} \,|\, \gamma\right). \tag{A.1}$$

A.1.1 Undirected Links

We further expand $\Pr\left(\mathbf{A}, \mathbf{y} \,|\, a, b, \gamma\right)$ for undirected graph as

$$\Pr\left(\mathbf{A}, \mathbf{y} \,|\, a, b, \gamma\right) \tag{A.2}$$

$$= \iint \Pr\left(\mathbf{A} \,|\, \mathbf{y}, \boldsymbol{\Omega}\right) \Pr\left(\boldsymbol{\Omega} \,|\, a, b\right) \Pr\left(\mathbf{y} \,|\, \boldsymbol{\mu}\right) \Pr\left(\boldsymbol{\mu} \,|\, \gamma\right) \mathrm{d}\boldsymbol{\Omega}\mathrm{d}\boldsymbol{\mu} \tag{A.3}$$

$$= \iint \left(\prod_{l \le l'} \prod_{d \in l, d' \in l'} \frac{\Omega_{l,l'}^{A_{d,d'}}}{A_{d,d'}!} \exp\left(-\Omega_{l,l'}\right)\right) \left(\prod_{l \le l'} \frac{b^a}{\Gamma(a)} \Omega_{l,l'}^{a-1} \exp\left(-b\Omega_{l,l'}\right)\right)$$

$$\left(\prod_{l=1}^{L} \mu_l^{N_l}\right) \left(\frac{1}{\Delta(\gamma)} \prod_{l=1}^{L} \mu_l^{\gamma-1}\right) \mathrm{d}\boldsymbol{\Omega}\mathrm{d}\boldsymbol{\mu} \tag{A.4}$$

$$\propto \iint \left(\prod_{l \le l'} \Omega_{l,l'}^{S_w(l,l')+a-1} \exp\left(-(S_e(l,l')+b)\Omega_{l,l'}\right)\right) \left(\prod_{l=1}^{L} \mu_l^{N_l+\gamma-1}\right) \mathrm{d}\boldsymbol{\Omega}\mathrm{d}\boldsymbol{\mu} \tag{A.5}$$

$$\propto \Delta(\mathbf{N_l}+\gamma)\prod_{l\leq l'}\frac{\Gamma\left(S_w(l,l')+a\right)}{\left(S_e(l,l')+b\right)^{S_w(l,l')+a}}, \tag{A.6}$$

where $S_w(l,l')$ is the weight sum of *observed* links between blocks l and l'; $S_e(l,l')$ is the number of *all possible* links (i.e. assuming all links are observed) between blocks l and l'. Specifically, $S_e(l,l')$ is defined as

$$S_e(l,l') = \begin{cases} N_l N_{l'} & l \neq l' \\ \frac{1}{2}N_l(N_l-1) & l = l' \end{cases} \tag{A.7}$$

where N_l denotes the number of documents assigned the block l.

$\Delta\left(\mathbf{N_l}+\gamma\right)$ is defined as

$$\Delta\left(\mathbf{N_l}+\gamma\right) = \frac{\prod\limits_{l'=1}^{L}\Gamma\left(N_{l'}+\gamma\right)}{\Gamma\left(\sum\limits_{l'=1}^{L}N_{l'}+L\gamma\right)}, \tag{A.8}$$

where $\Gamma(\cdot)$ is a Gamma function:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t, \tag{A.9}$$

whose most important property, as introduced in Section 2.1.1, is $\Gamma(x+1) = x\Gamma(x)$.

We then derive the Gibbs sampling equation for document d, given the block assignments of other documents and link weights excluding d, as

$$\Pr\left(y_d = l \,|\, \mathbf{A}^{-\mathbf{d}}, \mathbf{y}^{-\mathbf{d}}, a, b, \gamma\right) \tag{A.10}$$

$$=\frac{\Pr\left(\mathbf{A},\mathbf{y}\,|\,a,b,\gamma\right)}{\Pr\left(\mathbf{A}^{-\mathbf{d}},\mathbf{y}^{-\mathbf{d}}\,|\,a,b,\gamma\right)} \tag{A.11}$$

$$\propto\frac{\Gamma(D-1+L\gamma)}{\Gamma(D+L\gamma)}\frac{\Gamma(N_l+\gamma)}{\Gamma(N_l^{-d}+\gamma)}$$

$$\prod_{l'=1}^{L}\frac{\Gamma\left(S_w(l,l')+a\right)}{\left(S_e(l,l')+b\right)^{S_w(l,l')+a}}\frac{\left(S_e^{-d}(l,l')+b\right)^{S_w^{-d}(l,l')+a}}{\Gamma\left(S_w^{-d}(l,l')+a\right)} \tag{A.12}$$

$$\propto \frac{N_l^{-d}+\gamma}{D-1+L\gamma} \prod_{l'=1}^{L} \frac{\left(S_e^{-d}(l,l')+b\right)^{S_w^{-d}(l,l')+a}}{\left(S_e(l,l')+b\right)^{S_w(l,l')+a}} \prod_{i=0}^{S_w(d,l')-1} \left(S_w^{-d}(l,l')+a+i\right) \tag{A.13}$$

$$\propto \left(N_l^{-d}+\gamma\right) \prod_{l'=1}^{L} \frac{\left(S_e^{-d}(l,l')+b\right)^{S_w^{-d}(l,l')+a}}{\left(S_e^{-d}(l,l')+b+S_e(d,l')\right)^{S_w^{-d}(l,l')+a+S_w(d,l')}}$$

$$\prod_{i=0}^{S_w(d,l')-1} \left(S_w^{-d}(l,l')+a+i\right), \tag{A.14}$$

where $S_w(d,l')$ denotes the weight sum of *observed* links between document d and block l'; $S_e(d,l')$ denotes the number of *all possible* links between document d and block l'. Namely, $S_e(d,l') = N_{l'}$.

A.1.2 Directed Links

The expansion of $\Pr\left(\mathbf{A},\mathbf{y}\,|\,a,b,\gamma\right)$ for directed graph is

$$\Pr\left(\mathbf{A},\mathbf{y}\,|\,a,b,\gamma\right) \tag{A.15}$$

$$\propto \iint \left(\prod_{l,l'} \prod_{d\in l,d'\in l'} \frac{\Omega_{l,l'}^{A_{d,d'}}}{A_{d,d'}!} \exp\left(-\Omega_{l,l'}\right)\right) \left(\prod_{l,l'} \frac{b^a}{\Gamma(a)} \Omega_{l,l'}^{a-1} \exp\left(-b\Omega_{l,l'}\right)\right)$$

$$\left(\prod_{l=1}^{L} \mu_l^{N_l}\right) \left(\frac{1}{\Delta(\gamma)} \prod_{l=1}^{L} \mu_l^{\gamma-1}\right) \mathrm{d}\boldsymbol{\Omega}\mathrm{d}\boldsymbol{\mu} \tag{A.16}$$

$$\propto \Delta\left(\mathbf{N_l}+\gamma\right) \prod_{l,l'} \frac{\Gamma\left(S_w(l,l')+a\right)}{\left(S_e(l,l')+b\right)^{S_w(l,l')+a}}, \tag{A.17}$$

where $S_e(l,l')$ is defined as

$$S_e(l,l') = \begin{cases} 2N_lN_{l'} & l \neq l' \\ N_l(N_l-1) & l = l' \end{cases} \tag{A.18}$$

The Gibbs sampling equation is derived as

$$\Pr\left(y_d = l\,|\,\mathbf{A^{-d}},\mathbf{y^{-d}},a,b,\gamma\right) \tag{A.19}$$

$$= \frac{\Pr\left(\mathbf{A},\mathbf{y}\,|\,a,b,\gamma\right)}{\Pr\left(\mathbf{A^{-d}},\mathbf{y^{-d}}\,|\,a,b,\gamma\right)} \tag{A.20}$$

$$\begin{aligned}
&\propto \frac{\Gamma\left(D-1+L\gamma\right)}{\Gamma\left(D+L\gamma\right)} \frac{\Gamma\left(N_l+\gamma\right)}{\Gamma\left(N_l^{-d}+\gamma\right)} \\
&\prod_{l'=1,l'\neq l}^{L} \frac{\Gamma\left(S_w(l,l')+a\right)}{\left(S_e(l,l')+b\right)^{S_w(l,l')+a}} \frac{\left(S_e^{-d}(l,l')+b\right)^{S_w^{-d}(l,l')+a}}{\Gamma\left(S_w^{-d}(l,l')+a\right)} \\
&\prod_{l'=1,l'\neq l}^{L} \frac{\Gamma\left(S_w(l',l)+a\right)}{\left(S_e(l',l)+b\right)^{S_w(l',l)+a}} \frac{\left(S_e^{-d}(l',l)+b\right)^{S_w^{-d}(l',l)+a}}{\Gamma\left(S_w^{-d}(l',l)+a\right)} \\
&\frac{\Gamma\left(S_w(l,l)+a\right)}{\left(S_e(l,l)+b\right)^{S_w(l,l)+a}} \frac{\left(S_e^{-d}(l,l)+b\right)^{S_w^{-d}(l,l)+a}}{\Gamma\left(S_w^{-d}(l,l)+a\right)}
\end{aligned} \tag{A.21}$$

$$\begin{aligned}
&\propto \frac{N_l^{-d}+\gamma}{D-1+L\gamma} \prod_{l'=1,l'\neq l}^{L} \frac{\left(S_e^{-d}(l,l')+b\right)^{S_w^{-d}(l,l')+a}}{\left(S_e(l,l')+b\right)^{S_w(l,l')+a}} \prod_{i=0}^{S_w(d,l')-1} \left(S_w^{-d}(l,l')+a+i\right) \\
&\prod_{l'=1,l'\neq l}^{L} \frac{\left(S_e^{-d}(l',l)+b\right)^{S_w^{-d}(l',l)+a}}{\left(S_e(l',l)+b\right)^{S_w(l',l)+a}} \prod_{i=0}^{S_w(l',d)-1} \left(S_w^{-d}(l',l)+a+i\right) \\
&\frac{\left(S_e^{-d}(l,l)+b\right)^{S_w^{-d}(l,l)+a}}{\left(S_e(l,l)+b\right)^{S_w(l,l)+a}} \prod_{i=0}^{S_w(d,l)+S_w(l,d)-1} \left(S_w^{-d}(l,l)+a+i\right)
\end{aligned} \tag{A.22}$$

$$\begin{aligned}
&\propto \left(N_l^{-d}+\gamma\right) \\
&\prod_{l'=1,l'\neq l}^{L} \frac{\left(S_e^{-d}(l,l')+b\right)^{S_w^{-d}(l,l')+a}}{\left(S_e^{-d}(l,l')+b+S_e(d,l')\right)^{S_w^{-d}(l,l')+a+S_w(d,l')}} \prod_{i=0}^{S_w(d,l')-1} \left(S_w^{-d}(l,l')+a+i\right) \\
&\prod_{l'=1,l'\neq l}^{L} \frac{\left(S_e^{-d}(l',l)+b\right)^{S_w^{-d}(l',l)+a}}{\left(S_e^{-d}(l',l)+b+S_e(l',d)\right)^{S_w^{-d}(l',l)+a+S_w(l',d)}} \prod_{i=0}^{S_w(l',d)-1} \left(S_w^{-d}(l',l)+a+i\right) \\
&\frac{\left(S_e^{-d}(l,l)+b\right)^{S_w^{-d}(l,l)+a}}{\left(S_e^{-d}(l,l)+b+S_e(l,d)+S_e(d,l)\right)^{S_w^{-d}(l,l)+a+S_w(d,l)+S_w(l,d)}} \\
&\prod_{i=0}^{S_w(d,l)+S_w(l,d)-1} \left(S_w^{-d}(l,l)+a+i\right).
\end{aligned} \tag{A.23}$$

A.2 Sampling Topic Assignments

The joint probability of topic assignments $\Pr\left(\mathbf{z}, \mathbf{w} \,|\, \alpha, \beta, \boldsymbol{\pi}, \mathbf{y}\right)$ for LBH-RTM (Section 3.2.3) is

$$\Pr\left(\mathbf{z}, \mathbf{w}, \mathbf{B} \,|\, \alpha, \beta, \boldsymbol{\pi}, \mathbf{y}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{A.24}$$

$$= \iint \Pr\left(\mathbf{z} \,|\, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{\theta} \,|\, \alpha, \boldsymbol{\pi}, \mathbf{y}\right) \Pr\left(\mathbf{w} \,|\, \mathbf{z}, \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\phi} \,|\, \beta\right) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\phi}$$

$$\Pr\left(\mathbf{B} \,|\, \mathbf{z}, \mathbf{w}, \mathbf{y}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{A.25}$$

$$= \iint \left(\prod_{d=1}^{D}\prod_{k=1}^{K} \theta_{d,k}^{N_{d,k}}\right) \left(\prod_{d=1}^{D} \frac{1}{\Delta\left(\alpha \boldsymbol{\pi}_{\boldsymbol{y_d}}\right)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha \pi_{y_d,k}-1}\right)$$

$$\left(\prod_{k=1}^{K}\prod_{v=1}^{V} \phi_{k,v}^{N_{k,v}}\right) \left(\prod_{k=1}^{K} \frac{1}{\Delta(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{\beta-1}\right) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\phi}$$

$$\prod_{d,d'} f\left(B_{d,d'} \,|\, \mathbf{z_d}, \mathbf{z_{d'}}, \mathbf{w_d}, \mathbf{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{A.26}$$

$$= \iint \left(\prod_{d=1}^{D} \frac{1}{\Delta\left(\alpha \boldsymbol{\pi}_{\boldsymbol{y_d}}\right)} \prod_{k=1}^{K} \theta_{d,k}^{N_{d,k}+\alpha \pi_{y_d,k}-1}\right) \left(\prod_{k=1}^{K} \frac{1}{\Delta(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{N_{k,v}+\beta-1}\right) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\phi}$$

$$\prod_{d,d'} f\left(B_{d,d'} \,|\, \mathbf{z_d}, \mathbf{z_{d'}}, \mathbf{w_d}, \mathbf{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{A.27}$$

$$= \left(\prod_{d=1}^{D} \frac{\Delta\left(\mathbf{N_d} + \alpha \boldsymbol{\pi}_{\boldsymbol{y_d}}\right)}{\Delta\left(\alpha \boldsymbol{\pi}_{\boldsymbol{y_d}}\right)}\right) \left(\prod_{k=1}^{K} \frac{\Delta\left(\mathbf{N_k} + \beta\right)}{\Delta(\beta)}\right)$$

$$\prod_{d,d'} f\left(B_{d,d'} \,|\, \mathbf{z_d}, \mathbf{z_{d'}}, \mathbf{w_d}, \mathbf{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right). \tag{A.28}$$

The Gibbs sampling equation is then derived as

$$\Pr\left(z_{d,n} = k \,|\, \mathbf{z}^{-\mathbf{d,n}}, w_{d,n} = v, \mathbf{w}^{-\mathbf{d,n}}, \mathbf{B}, \alpha, \beta, \boldsymbol{\pi}, \mathbf{y}^{-\mathbf{d}}, y_d = l, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{A.29}$$

$$= \frac{\Pr\left(z_{d,n} = k, \mathbf{z}^{-\mathbf{d,n}}, w_{d,n} = v, \mathbf{w}^{-\mathbf{d,n}}, \mathbf{B} \,|\, \alpha, \beta, \boldsymbol{\pi}, \mathbf{y}^{-\mathbf{d}}, y_d = l, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)}{\Pr\left(\mathbf{z}^{-\mathbf{d,n}}, \mathbf{w}^{-\mathbf{d,n}}, \mathbf{B}^{-\mathbf{d,n}} \,|\, \alpha, \beta, \boldsymbol{\pi}, \mathbf{y}^{-\mathbf{d}}, y_d = l, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)} \tag{A.30}$$

$$= \frac{\Delta\left(\mathbf{N_d} + \alpha \boldsymbol{\pi_l}\right)}{\Delta\left(\mathbf{N}_{\mathbf{d}}^{-\mathbf{d,n}} + \alpha \boldsymbol{\pi_l}\right)} \frac{\Delta\left(\mathbf{N_k} + \beta\right)}{\Delta\left(\mathbf{N}_{\mathbf{k}}^{-\mathbf{d,n}} + \beta\right)}$$

$$\prod_{d'} \frac{f\left(B_{d,d'} \,|\, z_{d,n}=k, \mathbf{z}^{-\mathbf{d},\mathbf{n}}, \mathbf{z}_{\mathbf{d}'}, \mathbf{w}_{\mathbf{d}}, \mathbf{w}_{\mathbf{d}'}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)}{f\left(B_{d,d'} \,|\, \mathbf{z}^{-\mathbf{d},\mathbf{n}}, \mathbf{z}_{\mathbf{d}'}, \mathbf{w}^{-\mathbf{d},\mathbf{n}}, \mathbf{w}_{\mathbf{d}'}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right)} \tag{A.31}$$

$$\propto \left(N_{d,k}^{-d,n} + \alpha\pi_{l,k}^{-d,n}\right) \frac{N_{k,v}^{-d,n}+\beta}{N_{k,\cdot}^{-d,n}+V\beta}$$

$$\prod_{d'} f\left(B_{d,d'} \,|\, z_{d,n}=k, \mathbf{z}^{-\mathbf{d},\mathbf{n}}, \mathbf{z}_{\mathbf{d}'}, \mathbf{w}_{\mathbf{d}}, \mathbf{w}_{\mathbf{d}'}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right), \tag{A.32}$$

where $\pi_{l,k}^{-d,n}$ is estimated based on maximal path assumption (Cowans, 2006; Wallach, 2008):

$$\pi_{l,k}^{-d,n} = \frac{\sum\limits_{d':y_{d'}=l} N_{d',k}^{-d,n} + \alpha'}{\sum\limits_{d':y_{d'}=l} N_{d',\cdot}^{-d,n} + K\alpha'}. \tag{A.33}$$

A.2.1 Sigmoid Loss

We split d' into two subsets: d^+ and d^-. d^+ denotes the documents that have positive links (observed links, with weight 1) with d. d^- denotes the documents that have negative links (sampled from unobserved links, with weight 0). When using sigmoid loss, the probability of a positive link between documents d and d^+ is

$$\Pr\left(B_{d,d^+}=1 \,|\, \mathbf{z}_{\mathbf{d}}, \mathbf{z}_{\mathbf{d}^+}, \mathbf{w}_{\mathbf{d}}, \mathbf{w}_{\mathbf{d}^+}, y_d, y_{d^+}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{A.34}$$

$$=\sigma\left(\boldsymbol{\eta}^\top\left(\overline{\mathbf{z}}_{\mathbf{d}} \circ \overline{\mathbf{z}}_{\mathbf{d}^+}\right) + \boldsymbol{\tau}^\top\left(\overline{\mathbf{w}}_{\mathbf{d}} \circ \overline{\mathbf{w}}_{\mathbf{d}^+}\right) + \rho_{y_d,y_{d^+}}\Omega_{y_d,y_{d^+}}\right) \tag{A.35}$$

$$=\sigma\left(\sum_{k=1}^{K}\eta_k \frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d^+,k}}{N_{d^+,\cdot}} + \sum_{v=1}^{V}\tau_v \frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d^+,v}}{N_{d^+,\cdot}} + \rho_{y_d,y_{d^+}}\Omega_{y_d,y_{d^+}}\right), \tag{A.36}$$

where $\sigma(x) = 1/(1+\exp(-x))$.

Contrarily, the probability of a negative link between documents d and d^- is

$$\Pr\left(B_{d,d^-}=0 \,|\, \mathbf{z}_{\mathbf{d}}, \mathbf{z}_{\mathbf{d}^-}, \mathbf{w}_{\mathbf{d}}, \mathbf{w}_{\mathbf{d}^-}, y_d, y_{d^-}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \tag{A.37}$$

$$=1-\sigma\left(\sum_{k=1}^{K}\eta_k \frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d^-,k}}{N_{d^-,\cdot}} + \sum_{v=1}^{V}\tau_v \frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d^-,v}}{N_{d^-,\cdot}} + \rho_{y_d,y_{d^-}}\Omega_{y_d,y_{d^-}}\right). \tag{A.38}$$

Therefore, the Gibbs sampling equation is

$$
\begin{aligned}
&\Pr\left(z_{d,n} = k \,|\, \text{rest}\right) \qquad \text{(A.39)}\\
&\propto \left(N_{d,k}^{-d,n} + \alpha \pi_{l,k}^{-d,n}\right) \frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta}\\
&\prod_{d^+} \sigma\left(\frac{\eta_k}{N_{d,\cdot}} \frac{N_{d^+,k}}{N_{d^+,\cdot}} + \sum_{k'=1}^{K} \eta_{k'} \frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}} \frac{N_{d^+,k'}}{N_{d^+,\cdot}} + \sum_{v=1}^{V} \tau_v \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d^+,v}}{N_{d^+,\cdot}} + \rho_{y_d,y_{d^+}} \Omega_{y_d,y_{d^+}}\right)\\
&\prod_{d^-} \left(1 - \sigma\left(\frac{\eta_k}{N_{d,\cdot}} \frac{N_{d^-,k}}{N_{d^-,\cdot}} + \sum_{k'=1}^{K} \eta_{k'} \frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}} \frac{N_{d^-,k'}}{N_{d^-,\cdot}} + \sum_{v=1}^{V} \tau_v \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d^-,v}}{N_{d^-,\cdot}} + \rho_{y_d,y_{d^-}} \Omega_{y_d,y_{d^-}}\right)\right).
\end{aligned}
\tag{A.40}
$$

A.2.2 Hinge Loss

When using hinge loss, the probability of a link (either positive or negative, but the weight of a negative link is -1) between documents d and d' is

$$\Pr\left(B_{d,d'} \,|\, \mathbf{z_d}, \mathbf{z_{d'}}, \mathbf{w_d}, \mathbf{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) = \exp\left(-2c \max\left(0, \zeta_{d,d'}\right)\right), \tag{A.41}$$

where c is the regularization parameter (it's set to 1 in our experiments, so it does not appear in Chapter 3); $\zeta_{d,d'}$ is defined as

$$\zeta_{d,d'} = 1 - B_{d,d'} R_{d,d'}, \tag{A.42}$$

$R_{d,d'}$ is defined in Equation A.53.

Equation A.41 can be rewritten by introducing a latent variable $\lambda_{d,d'}$ (Polson and Scott, 2011):

$$
\begin{aligned}
&\Pr(B_{d,d'} \,|\, \mathbf{z_d}, \mathbf{z_{d'}}, \mathbf{w_d}, \mathbf{w_{d'}}, y_d, y_{d'}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho})\\
&= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_{d,d'}}} \exp\left(-\frac{\left(c\zeta_{d,d'} + \lambda_{d,d'}\right)^2}{2\lambda_{d,d'}}\right) \mathrm{d}\lambda_{d,d'}.
\end{aligned}
\tag{A.43}
$$

Thus the Gibbs sampling equation is

$$\Pr\left(z_{d,n} = k \,|\, \text{rest}\right) \propto \left(N_{d,k}^{-d,n} + \alpha\pi_{l,k}\right) \frac{N_{k,v}^{-d,n} + \beta}{N_{k,\cdot}^{-d,n} + V\beta} \prod_{d'} \exp\left(-\frac{\left(c\zeta_{d,d'} + \lambda_{d,d'}\right)^2}{2\lambda_{d,d'}}\right). \tag{A.44}$$

The exponent of final term of the equation above can be expanded as

$$- \frac{\left(c\zeta_{d,d'} + \lambda_{d,d'}\right)^2}{2\lambda_{d,d'}} \tag{A.45}$$

$$\propto - \frac{c^2\zeta_{d,d'}^2 + 2c\lambda_{d,d'}\zeta_{d,d'}}{2\lambda_{d,d'}} \tag{A.46}$$

$$\propto - \frac{c^2\left(1 - B_{d,d'}R_{d,d'}\right)^2 + 2c\lambda_{d,d'}\left(1 - B_{d,d'}R_{d,d'}\right)}{2\lambda_{d,d'}} \tag{A.47}$$

$$\propto - \frac{c^2\left(-2B_{d,d'}R_{d,d'} + R_{d,d'}^2\right) - 2c\lambda_{d,d'}B_{d,d'}R_{d,d'}}{2\lambda_{d,d'}} \tag{A.48}$$

$$\propto - \frac{c^2 R_{d,d'}^2}{2\lambda_{d,d'}} + \frac{cB_{d,d'}\left(c + \lambda_{d,d'}\right)R_{d,d'}}{\lambda_{d,d'}} \tag{A.49}$$

$$\begin{aligned} &\propto - \frac{c^2\left(\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} + \sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d',k'}}{N_{d',\cdot}} + \sum_{v=1}^{V}\tau_v\frac{N_{d,v}}{N_{d',\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} + \rho_{y_d,y_{d'}}\Omega_{y_d,y_{d'}}\right)^2}{2\lambda_{d,d'}} \\ &\quad + \frac{cB_{d,d'}\left(c + \lambda_{d,d'}\right)\left(\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} + \sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d',k'}}{N_{d',\cdot}} + \sum_{v=1}^{V}\tau_v\frac{N_{d,v}}{N_{d',\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} + \rho_{y_d,y_{d'}}\Omega_{y_d,y_{d'}}\right)}{\lambda_{d,d'}} \end{aligned} \tag{A.50}$$

$$\begin{aligned} &\propto - \frac{c^2\left(\frac{\eta_k^2}{N_{d,\cdot}^2}\frac{N_{d',k}^2}{N_{d',\cdot}^2} + 2\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}}\left(\sum_{k'=1}^{K}\eta_{k'}\frac{N_{d,k'}^{-d,n}}{N_{d,\cdot}}\frac{N_{d',k'}}{N_{d',\cdot}} + \sum_{v=1}^{V}\tau_v\frac{N_{d,v}}{N_{d',\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} + \rho_{y_d,y_{d'}}\Omega_{y_d,y_{d'}}\right)\right)}{2\lambda_{d,d'}} \\ &\quad + \frac{cB_{d,d'}\left(c + \lambda_{d,d'}\right)\frac{\eta_k}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}}}{\lambda_{d,d'}} \end{aligned} \tag{A.51}$$

$$\begin{aligned} &\propto - \frac{c^2\left(\eta_k^2 N_{d',k}^2 + 2\eta_k N_{d',k}\left(\sum_{k'=1}^{K}\eta_{k'}N_{d,k'}^{-d,n}N_{d',k'} + \sum_{v=1}^{V}\tau_v N_{d,v}N_{d',v} + \rho_{y_d,y_{d'}}\Omega_{y_d,y_{d'}}N_{d,\cdot}N_{d',\cdot}\right)\right)}{2\lambda_{d,d'}N_{d,\cdot}^2 N_{d',\cdot}^2} \\ &\quad + \frac{cB_{d,d'}\left(c + \lambda_{d,d'}\right)\eta_k N_{d',k}}{\lambda_{d,d'}N_{d,\cdot}N_{d',\cdot}}. \end{aligned} \tag{A.52}$$

A.3 Optimizing Parameters

Let the regression value of documents d and d' be

$$R_{d,d'} = \boldsymbol{\eta}^\top \left(\overline{\mathbf{z}}_{\mathbf{d}} \circ \overline{\mathbf{z}}_{\mathbf{d'}}\right) + \boldsymbol{\tau}^\top \left(\overline{\mathbf{w}}_{\mathbf{d}} \circ \overline{\mathbf{w}}_{\mathbf{d'}}\right) + \rho_{y_d,y_{d'}} \Omega_{y_d,y_{d'}}. \tag{A.53}$$

Its partial derivatives are

$$\frac{\partial R_{d,d'}}{\partial \eta_k} = \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}} \tag{A.54}$$

$$\frac{\partial R_{d,d'}}{\partial \tau_v} = \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d',v}}{N_{d',\cdot}} \tag{A.55}$$

$$\frac{\partial R_{d,d'}}{\partial \rho_{y_d,y_{d'}}} = \Omega_{y_d,y_{d'}}. \tag{A.56}$$

A.3.1 Sigmoid Loss

To optimize regression parameters, we first compute the log likelihood of $\mathbf{B}$ as

$$\begin{aligned}
\mathcal{L}(\mathbf{B}) = & \log \Pr\left(\mathbf{B} \,|\, \mathbf{z}, \mathbf{w}, \mathbf{y}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) + \log \Pr\left(\boldsymbol{\eta} \,|\, \nu\right) \\
& + \log \Pr\left(\boldsymbol{\tau} \,|\, \nu\right) + \log \Pr\left(\boldsymbol{\rho} \,|\, \nu\right) & \text{(A.57)} \\
\propto & -\sum_{d,d^+} \log\left(1 + \exp\left(-R_{d,d^+}\right)\right) + \sum_{d,d^-} \left(\log\left(\exp\left(-R_{d,d^-}\right)\right) - \log\left(1 + \exp\left(-R_{d,d^-}\right)\right)\right) \\
& - \sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L} \sum_{l'=1}^{L} \frac{\rho_{l,l'}^2}{2\nu^2} & \text{(A.58)} \\
\propto & -\sum_{d,d'} \log\left(1 + \exp\left(-R_{d,d'}\right)\right) - \sum_{d,d^-} R_{d,d^-} \\
& - \sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L} \sum_{l'=1}^{L} \frac{\rho_{l,l'}^2}{2\nu^2}. & \text{(A.59)}
\end{aligned}$$

Its derivatives are

$$\frac{\partial \mathcal{L}(\mathbf{B})}{\partial \eta_k} \propto -\frac{\eta_k}{\nu^2} + \sum_{d,d'} \frac{\exp\left(-R_{d,d'}\right)}{1+\exp\left(-R_{d,d'}\right)} \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d',k}}{N_{d',\cdot}} - \sum_{d,d^-} \frac{N_{d,k}}{N_{d,\cdot}} \frac{N_{d^-,k}}{N_{d^-,\cdot}} \tag{A.60}$$

$$\frac{\partial \mathcal{L}(\mathbf{B})}{\partial \tau_v} \propto -\frac{\tau_v}{\nu^2} + \sum_{d,d'} \frac{\exp\left(-R_{d,d'}\right)}{1+\exp\left(-R_{d,d'}\right)} \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d',v}}{N_{d',\cdot}} - \sum_{d,d^-} \frac{N_{d,v}}{N_{d,\cdot}} \frac{N_{d^-,v}}{N_{d^-,\cdot}} \tag{A.61}$$

$$\frac{\partial \mathcal{L}(\mathbf{B})}{\partial \rho_{l,l'}} \propto -\frac{\rho_{l,l'}}{\nu^2} + \sum_{d\in l, d'\in l'} \frac{\exp\left(-R_{d,d'}\right)}{1+\exp\left(-R_{d,d'}\right)} \Omega_{l,l'} - \sum_{d\in l, d^-\in l'} \Omega_{l,l'}. \tag{A.62}$$

A.3.2 Hinge Loss

The log likelihood of $\mathbf{B}$ is

$$\begin{aligned}
\mathcal{L}(\mathbf{B}) = & \log \Pr\left(\mathbf{B} \,|\, \mathbf{z}, \mathbf{w}, \mathbf{y}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) + \log \Pr\left(\boldsymbol{\eta} \,|\, \nu\right) \\
& + \log \Pr\left(\boldsymbol{\tau} \,|\, \nu\right) + \log \Pr\left(\boldsymbol{\rho} \,|\, \nu\right) & \text{(A.63)} \\
\propto & -\sum_{d,d'} \frac{\left(c\zeta_{d,d'} + \lambda_{d,d'}\right)^2}{2\lambda_{d,d'}} - \sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L} \frac{\rho_{l,l'}^2}{2\nu^2} & \text{(A.64)} \\
\propto & -\sum_{d,d'} \frac{c^2\zeta_{d,d'}^2 + 2c\lambda_{d,d'}\zeta_{d,d'}}{2\lambda_{d,d'}} - \sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L} \frac{\rho_{l,l'}^2}{2\nu^2} & \text{(A.65)} \\
\propto & -\sum_{d,d'} \frac{c^2\left(1 - B_{d,d'}R_{d,d'}\right)^2 + 2c\lambda_{d,d'}\left(1 - B_{d,d'}R_{d,d'}\right)}{2\lambda_{d,d'}} \\
& - \sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L} \frac{\rho_{l,l'}^2}{2\nu^2} & \text{(A.66)} \\
\propto & -\sum_{d,d'} \frac{c^2R_{d,d'}^2 - 2c\left(c + \lambda_{d,d'}\right)B_{d,d'}R_{d,d'}}{2\lambda_{d,d'}} \\
& - \sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{v=1}^{V} \frac{\tau_v^2}{2\nu^2} - \sum_{l=1}^{L}\sum_{l'=1}^{L} \frac{\rho_{l,l'}^2}{2\nu^2}. & \text{(A.67)}
\end{aligned}$$

The partial derivatives of $R_{d,d'}^2$ are

$$\frac{\partial R_{d,d'}^2}{\partial \eta_k} = 2R_{d,d'}\frac{\partial R_{d,d'}}{\partial \eta_k} = 2R_{d,d'}\frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} \tag{A.68}$$

$$\frac{\partial R_{d,d'}^2}{\partial \tau_v} = 2R_{d,d'}\frac{\partial R_{d,d'}}{\partial \tau_v} = 2R_{d,d'}\frac{N_{d,v}}{N_{d,\cdot}}\frac{N_{d',v}}{N_{d',\cdot}} \tag{A.69}$$

$$\frac{\partial R_{d,d'}^2}{\partial \rho_{y_d,y_{d'}}} = 2R_{d,d'}\frac{\partial R_{d,d'}}{\partial \rho_{y_d,y_{d'}}} = 2R_{d,d'}\Omega_{y_d,y_{d'}}. \tag{A.70}$$

So the partial derivatives of $\mathcal{L}(\mathbf{B})$ are

$$\frac{\partial \mathcal{L}(\mathbf{B})}{\partial \eta_k} \propto -\sum_{d,d'} \frac{c\left(cR_{d,d'} - \left(c + \lambda_{d,d'}\right) B_{d,d'}\right)}{\lambda_{d,d'}} \frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} - \frac{\eta_k}{\nu^2} \tag{A.71}$$

$$\frac{\partial \mathcal{L}(\mathbf{B})}{\partial \tau_v} \propto -\sum_{d,d'} \frac{c\left(cR_{d,d'} - \left(c + \lambda_{d,d'}\right) B_{d,d'}\right)}{\lambda_{d,d'}} \frac{N_{d,k}}{N_{d,\cdot}}\frac{N_{d',k}}{N_{d',\cdot}} - \frac{\tau_v}{\nu^2} \tag{A.72}$$

$$\frac{\partial \mathcal{L}(\mathbf{B})}{\partial \rho_{l,l'}} \propto -\sum_{d\in l, d'\in l'} \frac{c\left(cR_{d,d'} - \left(c + \lambda_{d,d'}\right) B_{d,d'}\right)}{\lambda_{d,d'}} \Omega_{l,l'} - \frac{\rho_{l,l'}}{\nu^2}. \tag{A.73}$$

The likelihood of latent variable $\lambda_{d,d'}$ is

$$\begin{aligned} &\Pr\left(\lambda_{d,d'} \,|\, \mathbf{z}, \mathbf{w}, \mathbf{y}, \boldsymbol{\Omega}, \mathbf{B}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \\ \propto& \frac{1}{\sqrt{2\pi\lambda_{d,d'}}} \exp\left(-\frac{\left(\lambda_{d,d'} + c\zeta_{d,d'}\right)^2}{2\lambda_{d,d'}}\right) && \text{(A.74)} \\ \propto& \frac{1}{\sqrt{2\pi\lambda_{d,d'}}} \exp\left(-\frac{c^2\zeta_{d,d'}^2}{2\lambda_{d,d'}} - \frac{\lambda_{d,d'}}{2}\right) && \text{(A.75)} \\ \propto& \mathcal{GIG}\left(\lambda_{d,d'}\,;\, \frac{1}{2}, 1, c^2\zeta_{d,d'}^2\right), && \text{(A.76)} \end{aligned}$$

where $\mathcal{GIG}$ denotes generalized inverse Gaussian distribution which is defined as

$$\mathcal{GIG}\left(x\,;\, p, a, b\right) = C\left(p, a, b\right) x^{p-1} \exp\left(-\frac{1}{2}\left(\frac{b}{x} + ax\right)\right), \tag{A.77}$$

where $C(p, a, b)$ is a normalizer.

We can sample $\lambda_{d,d'}^{-1}$ (then $\lambda_{d,d'}$) from an inverse Gaussian distribution

$$\Pr\left(\lambda_{d,d'}^{-1} \,|\, \mathbf{z}, \mathbf{w}, \mathbf{y}, \boldsymbol{\Omega}, \mathbf{B}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\rho}\right) \propto \mathcal{IG}\left(\lambda_{d,d'}^{-1}\,;\, \frac{1}{c|\zeta_{d,d'}|}, 1\right), \tag{A.78}$$

where

$$\mathcal{IG}\left(x\,;\, a, b\right) = \sqrt{\frac{b}{2\pi x^3}} \exp\left(-\frac{b\left(x-a\right)^2}{2a^2x}\right), \tag{A.79}$$

for $a > 0$ and $b > 0$.

A.4 Sampling Process

The following is the sampling process of LBS-RTM. It is less complex than that of LBH-RTM (Algorithm 1).

Algorithm 2 Sampling Process of LBS-RTM

```
Sample implicit negative links as explicit ones from a uniform distribution
Initialize every topic assignment $z_{d,n}$ from a uniform distribution
for $m = 1$ to $M$ do
    Optimize $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ using L-BFGS (Equations A.59, A.60, A.61, and A.62)
    for each document $d = 1$ to $D$ do
        Draw block assignment $y_d$ from the multinomial distribution (Equation A.14)
        for each token $n$ in document $d$ do
            Draw a topic assignment $z_{d,n}$ from the multinomial distribution (Equations A.40)
        end for
    end for
end for
```

Appendix B: Derivation of the Posterior Inference for the Multilingual Topic Model

This appendix includes the detailed derivation of the Gibbs sampling equations for the multilingual topic model in Section 5.

The joint likelihood of generating the corpora in languages S and T is

$$
\begin{aligned}
&\Pr\left(\mathbf{z_S}, \mathbf{w_S}, \mathbf{z_T}, \mathbf{w_T} \,|\, \alpha_S, \beta_S, \alpha_T, \beta_T, \boldsymbol{\rho}, \boldsymbol{\eta}\right) && \text{(B.1)}\\
=&\Psi\left(\boldsymbol{\rho}, \boldsymbol{\eta}, \mathbf{z_S}, \mathbf{w_S}, \mathbf{z_T}, \mathbf{w_T}\right) \Pr\left(\mathbf{z_S}, \mathbf{w_S} \,|\, \alpha_S, \beta_S\right) \Pr\left(\mathbf{z_T}, \mathbf{w_T} \,|\, \alpha_T, \beta_T\right) && \text{(B.2)}\\
=&\exp\left(f\left(\boldsymbol{\rho}, \mathbf{z_S}, \mathbf{z_T}, \mathbf{w_S}, \mathbf{w_T}\right)\right)\\
&\iiiint \Pr\left(\mathbf{w_S} \,|\, \mathbf{z_S}, \boldsymbol{\phi_S}\right) \Pr\left(\boldsymbol{\phi_S} \,|\, \beta_S\right) \Pr\left(\mathbf{z_S} \,|\, \boldsymbol{\theta_S}\right) \Pr\left(\boldsymbol{\theta_S} \,|\, \alpha_S\right)\\
&\Pr\left(\mathbf{w_T} \,|\, \mathbf{z_T}, \boldsymbol{\phi_T}\right) \Pr\left(\boldsymbol{\phi_T} \,|\, \beta_T\right) \Pr\left(\mathbf{z_T} \,|\, \boldsymbol{\theta_T}\right) \Pr\left(\boldsymbol{\theta_T} \,|\, \alpha_T\right) \mathrm{d}\boldsymbol{\phi_S} \mathrm{d}\boldsymbol{\theta_S} \mathrm{d}\boldsymbol{\phi_T} \mathrm{d}\boldsymbol{\theta_T} && \text{(B.3)}\\
=&\exp\left(f\left(\boldsymbol{\rho}, \boldsymbol{\eta}, \mathbf{z_S}, \mathbf{z_T}, \mathbf{w_S}, \mathbf{w_T}\right)\right) \iiiint \left(\prod_{k_S=1}^{K_S} \prod_{v_S=1}^{V_S} \phi_{S,k_S,v_S}^{N_{S,k_S,v_S}}\right)\\
&\left(\prod_{k_S=1}^{K_S} \frac{1}{\Delta(\beta_S)} \prod_{v_S=1}^{V_S} \phi_{S,k_S,v_S}^{\beta_S-1}\right) \left(\prod_{d_S=1}^{D_S} \prod_{k_S=1}^{K_S} \theta_{S,d_S,k_S}^{N_{S,d_S,k_S}}\right) \left(\prod_{d_S=1}^{D_S} \frac{1}{\Delta(\alpha_S)} \prod_{k_S=1}^{K_S} \theta_{S,d_S,k_S}^{\alpha_S-1}\right)\\
&\left(\prod_{k_T=1}^{K_T} \prod_{v_T=1}^{V_T} \phi_{T,k_T,v_T}^{N_{T,k_T,v_T}}\right) \left(\prod_{k_T=1}^{K_T} \frac{1}{\Delta(\beta_T)} \prod_{v_T=1}^{V_T} \phi_{T,k_T,v_T}^{\beta_T-1}\right) \left(\prod_{d_T=1}^{D_T} \prod_{k_T=1}^{K_T} \theta_{T,d_T,k_T}^{N_{T,d_T,k_T}}\right)\\
&\left(\prod_{d_T=1}^{D_T} \frac{1}{\Delta(\alpha_T)} \prod_{k_T=1}^{K_T} \theta_{T,d_T,k_T}^{\alpha_T-1}\right) \mathrm{d}\boldsymbol{\phi_S} \mathrm{d}\boldsymbol{\theta_S} \mathrm{d}\boldsymbol{\phi_T} \mathrm{d}\boldsymbol{\theta_T} && \text{(B.4)}\\
\propto&\exp\left(f\left(\boldsymbol{\rho}, \boldsymbol{\eta}, \mathbf{z_S}, \mathbf{z_T}, \mathbf{w_S}, \mathbf{w_T}\right)\right) \iiiint \left(\prod_{k_S=1}^{K_S} \prod_{v_S=1}^{V_S} \phi_{S,k_S,v_S}^{N_{S,k_S,v_S}+\beta_S-1}\right)
\end{aligned}
$$

$$
\left(\prod_{d_S=1}^{D_S}\prod_{k_S=1}^{K_S}\theta_{S,d_S,k_S}^{N_{S,d_S,k_S}+\alpha_S-1}\right)\left(\prod_{k_T=1}^{K_T}\prod_{v_T=1}^{V_T}\phi_{T,k_T,v_T}^{N_{T,k_T,v_T}+\beta_T-1}\right)
\left(\prod_{d_T=1}^{D_T}\prod_{k_T=1}^{K_T}\theta_{T,d_T,k_T}^{N_{T,d_T,k_T}+\alpha_T-1}\right)\mathrm{d}\boldsymbol{\phi_S}\mathrm{d}\boldsymbol{\theta_S}\mathrm{d}\boldsymbol{\phi_T}\mathrm{d}\boldsymbol{\theta_T} \tag{B.5}
$$

$$
= \exp\left(f\left(\boldsymbol{\rho},\boldsymbol{\eta},\mathbf{z_S},\mathbf{z_T},\mathbf{w_S},\mathbf{w_T}\right)\right)\left(\prod_{k_S=1}^{K_S}\Delta(\mathbf{N_{S,k_S}}+\beta_S)\right)
\left(\prod_{d_S=1}^{D_S}\Delta(\mathbf{N_{S,d_S}}+\alpha_S)\right)\left(\prod_{k_T=1}^{K_T}\Delta(\mathbf{N_{T,k_T}}+\beta_T)\right)\left(\prod_{d_T=1}^{D_T}\Delta(\mathbf{N_{T,d_T}}+\alpha_T)\right), \tag{B.6}
$$

where

$$
\Delta\left(\mathbf{N_{S,k_S}}+\beta_S\right)=\frac{\prod\limits_{v_S=1}^{V_S}\Gamma\left(N_{S,k_S,v_S}+\beta_S\right)}{\Gamma\left(\sum\limits_{v_S=1}^{V_S}N_{S,k_S,v_S}+V_S\beta_S\right)}, \tag{B.7}
$$

where $\Gamma(\cdot)$ is a Gamma function and $\Gamma(x+1)=x\Gamma(x)$.

Thus, for a token in language S, the Gibbs sampling equation for updating its topic assignment is

$$
\Pr\left(z_{S,d,n}=k\,|\,\mathbf{z}^{-\mathbf{S,d,n}},\mathbf{w}^{-\mathbf{S,d,n}},w_{S,d,n}=v,\mathbf{z_T},\mathbf{w_T},\alpha_S,\beta_S,\boldsymbol{\rho},\boldsymbol{\eta}\right)
$$

$$
\propto\frac{\Delta(\mathbf{N_{S,d_S}}+\alpha_S)}{\Delta(\mathbf{N}_{\mathbf{S,d_S}}^{-\mathbf{S,d,n}}+\alpha_S)}\frac{\Delta(\mathbf{N_{S,k}}+\beta_S)}{\Delta(\mathbf{N}_{\mathbf{S,k}}^{-\mathbf{S,d,n}}+\beta_S)}
\frac{\exp\left(f\left(\boldsymbol{\rho},\boldsymbol{\eta},\mathbf{z_T},\mathbf{w_T},\mathbf{z}_{\mathbf{S}}^{-\mathbf{S,d,n}},\mathbf{w}_{\mathbf{S}}^{-\mathbf{S,d,n}},z_{S,d,n}=k,w_{S,d,n}=v\right)\right)}{\exp\left(f\left(\boldsymbol{\rho},\boldsymbol{\eta},\mathbf{z_T},\mathbf{w_T},\mathbf{z}_{\mathbf{S}}^{-\mathbf{S,d,n}},\mathbf{w}_{\mathbf{S}}^{-\mathbf{S,d,n}}\right)\right)} \tag{B.8}
$$

$$
\propto\left(N_{S,d,k}^{-S,d,n}+\alpha_S\right)\frac{N_{S,k,v}^{-S,d,n}+\beta_S}{N_{S,k,\cdot}^{-S,d,n}+V_S\beta_S}\left(\prod_{v'\in\mathrm{Dic}(v)}\left[\mathrm{Dis}\left(\boldsymbol{\Omega_{S,v}},\boldsymbol{\rho_{T\to S}}\boldsymbol{\Omega_{T,v'}}\right)\right]^{\eta_{v,v'}}\right)^{-1}
\left(\prod_{v'\in\mathrm{Dic}(v)}\left[\mathrm{Dis}\left(\boldsymbol{\rho_{S\to T}}\boldsymbol{\Omega_{S,v}},\boldsymbol{\Omega_{T,v'}}\right)\right]^{\eta_{v,v'}}\right)^{-1}. \tag{B.9}
$$

The derivation for sampling a token's topic assignment in language T is similar—just swap subscriptions S and T:

$$
\Pr\left(z_{T,d,n}=k\,|\,\mathbf{z}^{-\mathbf{T,d,n}},\mathbf{w}^{-\mathbf{T,d,n}},w_{T,d,n}=v,\mathbf{z_S},\mathbf{w_S},\alpha_T,\beta_T,\boldsymbol{\rho}\right)
$$

$$
\begin{aligned}
&\propto \frac{\Delta(\mathbf{N}_{\mathbf{T},\mathbf{d_T}}+\alpha_T)}{\Delta(\mathbf{N}^{-\mathbf{T},\mathbf{d},\mathbf{n}}_{\mathbf{T},\mathbf{d_T}}+\alpha_T)} \frac{\Delta(\mathbf{N}_{\mathbf{T},\mathbf{k}}+\beta_T)}{\Delta(\mathbf{N}^{-\mathbf{T},\mathbf{d},\mathbf{n}}_{\mathbf{T},\mathbf{k}}+\beta_T)} \\
&\quad \frac{\exp\left(f\left(\boldsymbol{\rho},\boldsymbol{\eta},\mathbf{z_S},\mathbf{w_S},\mathbf{z_T}^{-\mathbf{T},\mathbf{d},\mathbf{n}},\mathbf{w_T}^{-\mathbf{T},\mathbf{d},\mathbf{n}},z_{T,d,n}=k,w_{T,d,n}=v\right)\right)}{\exp\left(f\left(\boldsymbol{\rho},\boldsymbol{\eta},\mathbf{z_S},\mathbf{w_S},\mathbf{z_T}^{-\mathbf{T},\mathbf{d},\mathbf{n}},\mathbf{w_T}^{-\mathbf{T},\mathbf{d},\mathbf{n}}\right)\right)}
\end{aligned}
\tag{B.10}
$$

$$
\begin{aligned}
&\propto \left(N^{-T,d,n}_{T,d,k}+\alpha_T\right)\frac{N^{-T,d,n}_{T,k,v}+\beta_T}{N^{-T,d,n}_{T,k,\cdot}+V_T\beta_T}\left(\prod_{v'\in\mathrm{Dic}(v)}\left[\mathrm{Dis}\left(\boldsymbol{\Omega_{S,v'}},\boldsymbol{\rho_{T\to S}}\boldsymbol{\Omega_{T,v}}\right)\right]^{\eta_{v',v}}\right)^{-1} \\
&\quad \left(\prod_{v'\in\mathrm{Dic}(v)}\left[\mathrm{Dis}\left(\boldsymbol{\rho_{S\to T}}\boldsymbol{\Omega_{S,v'}},\boldsymbol{\Omega_{T,v}}\right)\right]^{\eta_{v',v}}\right)^{-1}.
\end{aligned}
\tag{B.11}
$$

Bibliography

Christopher Aicher, Abigail Z. Jacobs, and Aaron Clauset. 2014. Learning latent block structure in weighted networks. *Journal of Complex Networks*, pages 221–248.

Francesca Albertini and Eduardo D. Sontag. 1992. For neural networks, function determines form. In *Proceedings of IEEE Conference on Decision and Control*.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *arXiv:1602.01925*.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the International Conference of Machine Learning*.

Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. 2009. Joint emotion-topic modeling for social affective text mining. In *Proceedings of the International Conference on Data Mining*.

Ole Barndorff-Nielsen and Christian Halgreen. 1977. Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. *Probability Theory and Related Fields*, pages 309–311.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2007. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the Association for Computing Machinery*, pages 1–30.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the Conference Uncertainty in Artificial Intelligence*.

Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. *Applications of Topic Models*, volume 11 of *Foundations and Trends in Information Retrieval*. NOW Publishers.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Gilles Celeux. 1985. The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, pages 73–82.

Jonathan Chang and David M. Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, pages 124–150.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Snigdha Chaturvedi, Hal Daumé III, Taesun Moon, and Shashank Srivastava. 2012. A topical graph kernel for link prediction in labeled graphs. In *Proceedings of the International Conference of Machine Learning*.

Raj Chhikara. 1988. *The Inverse Gaussian Distribution: Theory: Methodology, and Applications*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, pages 22–29.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the International Conference of Machine Learning*.

Philip J. Cowans. 2006. *Probabilistic Document Modelling*. Ph.D. thesis, University of Cambridge.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of the Association for Computational Linguistics*.

Hal Daumé III. 2009. Markov random topic fields. In *Proceedings of the Association for Computational Linguistics*.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, pages 391–407.

Susan T. Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology*, pages 188–230.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, pages 61–74.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Li Fei-Fei and Pietro Perona. 2005. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

Thomas S. Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Michael J. Paul. 2019. A better way to diagnose poorly-mixed cross-lingual embeddings: A resource-free graph-based evaluation metric. In *Proceedings of the Association for Computational Linguistics*.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 564–584.

Erving Goffman. 1978. *The Presentation of Self in Everyday Life*. Harmondsworth London.

Joshua Goodman. 1996. Parsing algorithms and metrics. In *Proceedings of the Association for Computational Linguistics*.

Dilan Görür and Yee Whye Teh. 2009. An efficient sequential Monte Carlo algorithm for coalescent clustering. In *Proceedings of Advances in Neural Information Processing Systems*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, pages 5228–5235.

E. Dario Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, pages 47–60.

Michael U. Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, pages 307–361.

Shudong Hao, Jordan L. Boyd-Graber, and Michael J. Paul. 2018. Lessons from the Bible on modern topics: Adapting topic model evaluation to multilingual and low-resource settings. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Shudong Hao and Michael J. Paul. 2018. Learning multilingual topics from incomparable corpora. In *Proceedings of International Conference on Computational Linguistics*.

Junxian He, Zhiting Hu, Taylor Berg-Kirkpatrick, Ying Huang, and Eric P. Xing. 2017. Efficient correlated topic modeling with topic embedding. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*.

Gregor Heinrich. 2008. Parameter estimation for text analysis. *Technical Report*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributional semantics. In *Proceedings of the Association for Computational Linguistics*.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social Networks*, pages 109–137.

Weihua Hu and Jun'ichi Tsujii. 2016. A latent concept topic model for robust topic inference using word embeddings. In *Proceedings of the Association for Computational Linguistics*.

Yuening Hu, Jordan Boyd-Graber, Hal Daumé III, and Z. Irene Ying. 2013. Binary to bushy: Bayesian hierarchical clustering with the Beta coalescent. In *Proceedings of Advances in Neural Information Processing Systems*.

Yuening Hu, Ke Zhai, Vlad Eidelman, and Jordan Boyd-Graber. 2014. Polylingual tree-based topic models for translation domain adaptation. In *Proceedings of the Association for Computational Linguistics*.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the European Conference on Information Retrieval*.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of ACM International Conference on Web Search and Data Mining*.

Thorsten Joachims. 1998. Making large-scale SVM learning practical. Technical report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Myunghwan Kim and Jure Leskovec. 2012. Latent multi-group membership graph model. In *Proceedings of the International Conference of Machine Learning*.

Renée Koplon and Eduardo D. Sontag. 1997. Using Fourier-neural recurrent networks to fit sequential input/output data. *Neurocomputing*, pages 225–248.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the International Conference of Machine Learning*.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line trend analysis with topic models: #twitter trends detection topic model online. In *Proceedings of International Conference on Computational Linguistics*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the Association for Computational Linguistics*.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, and Myron Gutmann. 2009. Computational social science. *Science*, pages 721–723.

Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J. Smola. 2014. Reducing the sampling complexity of topic models. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, pages 1019–1031.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, pages 503–528.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Alena Lukasová. 1979. Hierarchical agglomerative clustering procedure. *Pattern Recognition*, pages 365–381.

Benjamin Marlin. 2003. Modeling user rating profiles for collaborative filtering. In *Proceedings of Advances in Neural Information Processing Systems*.

Ben Marwick. 2013. Discovery of emergent issues and controversies in anthropology using text mining, topic modeling, and social network analysis of microblog content. *Data Mining Applications with R*, page 514.

Jon D. McAuliffe and David M. Blei. 2008. Supervised topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of Internet portals with machine learning. *Information Retrieval*, pages 127–163.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, pages 39–41.

David Mimno and Andrew McCallum. 2012. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the Conference Uncertainty in Artificial Intelligence*.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of Empirical Methods in Natural Language Processing*.

Andriy Mnih and Yee Whye Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. *ArXiv*.

Robert Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of Empirical Methods in Natural Language Processing*.

David Newman, Edwin V. Bonilla, and Wray Buntine. 2011. Improving topic coherence with regularized topic models. In *Proceedings of Advances in Neural Information Processing Systems*.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of Advances in Neural Information Processing Systems*.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of the World Wide Web Conference*.

Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. 1998. Latent semantic indexing: A probabilistic analysis. In *Proceedings of ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*.

Nanyun Peng, Yiming Wang, and Mark Dredze. 2014. Learning polylingual topic models from code-switched social media documents. In *Proceedings of the Association for Computational Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Nicholas G. Polson and Steven L. Scott. 2011. Data augmentation for support vector machines. *Bayesian Analysis*, pages 1–23.

Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the Association for Computational Linguistics*.

Philip Resnik and Eric Hardisty. 2010. Gibbs sampling for the uninitiated. Technical report, UMIACS-TR-2010-04, University of Maryland, College Park.

Stephen E. Robertson, Steve Walker, Micheline Beaulieu, and Peter Willett. 1999. Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track. *NIST Special Publication Sp*, pages 253–264.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. *NIST Special Publication Sp*, page 109.

V. Seshadri. 1997. Halphen's laws. *Encyclopedia of Statistical Sciences*.

Micha Sharir. 1981. A strong-connectivity algorithm and its applications in data flow analysis. *Computers & Mathematics with Applications*, pages 67–72.

Bei Shi, Wai Lam, Lidong Bing, and Yinqing Xu. 2016. Detecting common discussion topics across culture from news reader comments. In *Proceedings of the Association for Computational Linguistics*.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. In *Proceedings of the Association for Computational Linguistics*.

Stephanie M. Strassel, Ann Bies, and Jennifer Tracey. 2017. Situational awareness for low resource languages: The LORELEI situation frame annotation task. In *Proceedings of the European Conference on Information Retrieval*.

Stephanie M. Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Language Resources and Evaluation Conference*.

Yee Whye Teh, Hal Daumé III, and Daniel M. Roy. 2007. Bayesian agglomerative clustering with coalescents. In *Proceedings of Advances in Neural Information Processing Systems*.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, pages 1566–1581.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the International Conference of Machine Learning*.

Graham JG Upton. 1992. Fisher's exact test. *Journal of the Royal Statistical Society*, pages 395–402.

Ellen M. Voorhees. 2001. Evaluation by highly relevant documents. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the Association for Computational Linguistics*.

Martin J. Wainwright and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, pages 1–305.

Hanna M. Wallach. 2008. *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.

Yuchung J. Wang and George Y. Wong. 1987. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, pages 8–19.

Xing Wei and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tze-I Yang, Andrew J. Torget, and Rada Mihalcea. 2011. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2015a. Birds of a feather linked together: A discriminative topic model using link-based priors. In *Proceedings of Empirical Methods in Natural Language Processing*.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2016. A discriminative topic model using document network structure. In *Proceedings of the Association for Computational Linguistics*.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2017. Adapting topic models using lexical associations with tree priors. In *Proceedings of Empirical Methods in Natural Language Processing*.

Yi Yang, Doug Downey, and Jordan Boyd-Graber. 2015b. Efficient methods for incorporating knowledge into topic models. In *Proceedings of Empirical Methods in Natural Language Processing*.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. 2015. LightLDA: Big topic models on modest computer clusters. In *Proceedings of the World Wide Web Conference*.

Michelle Yuan, Benjamin Van Durme, and Jordan Boyd-Graber. 2018. Multilingual anchoring: Interactive topic modeling and alignment across languages. In *Proceedings of Advances in Neural Information Processing Systems*.

Ke Zhai and Jordan Boyd-Graber. 2013. Online latent Dirichlet allocation with infinite vocabulary. In *Proceedings of the International Conference of Machine Learning*.

Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad L. Alkhouja. 2012. Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. In *Proceedings of the World Wide Web Conference*.

Jun Zhu, Amr Ahmed, and Eric P. Xing. 2012. MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research*, pages 2237–2278.

Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. 2014. Gibbs max-margin topic models with data augmentation. *Journal of Machine Learning Research*, pages 1073–1110.