Document Topic Modeling Methods Based on Context Information

李, 文博

出版情報：九州大学, 2021, 博士（情報科学）, 課程博士
バージョン：
権利関係：

Document Topic Modeling Methods Based on Context Information

Wenbo Li

Graduate School of Information Sciences and Electrical Engineering

Kyushu University

A thesis submitted for the degree of

Doctor of Computer Science

July 2021

Abstract

Topic modeling is an unsupervised method which takes a set of documents as input and automatically clusters word groups and similar expressions that best characterize a set of documents, which is widely used for dealing with sparse high-dimensional features of text data and finding latent semantic relationships between documents. Despite its significant progress in various tasks and applications, for some more complicated text analysis tasks such as sequential documents processing and word sense analysis, the crucial information is finer-grained semantic relationships between words (e.g., polysemy and synonyms) or documents (e.g., semantic relevance of sequential documents). These fine-grained semantic relations are normally hidden in the context information of the dataset, such as the surrounding documents for each sequential document of a document sequence (document-level context) and the surrounding words for each word in a document (word-level context). Most existing methods either ignore the context information or simply introduce them without considering the intrinsic relationship with the target task.

In this thesis, we mainly consider two types of contextual information: document-level context information and word-level context information. For the type of the document-level context information, its usage in topic modeling is intuitive, i.e., using the topics of the proper surrounding documents to clarify the topic distribution of each target document. This usage is suitable for the task of topic modeling of sequential documents. Specifically, we first consider the document-level context information and propose a new topic modeling methods for sequential documents with topic evolution based on hybrid inter-document topic dependency. For the sequential documents, we focus on modeling topic evolution, i.e., various and frequent emergences, growths and fades of topics, which is commonly applied to emerging topic detection tasks, such as emerging topic clustering and novel topic detection.

For the word-level context, there are at least two aspects of usages in the topic modeling process: (1) using the topics of the context words to assist the topic assignment for each word in a document; (2) generate both the word and its context words based on an assigned topic. For (1), it is suitable for handling the topic segmentation task for each document, i.e., dividing a document into a sequence of topically coherent segments. For (2), it is more suitable for dealing with the word sense disambiguation task in topic modeling, since the unit contained in the document is no longer a word, but a combination of each word and its corresponding context words. Based on the

two aspects of word-level contexts, we propose two topic models for two tasks: topic segmentation task and word sense disambiguation (WSD) task. The former task is to divide a document into a sequence of topically coherent segments, while preserving long topic change-points and keeping short topic segments from getting merged. The latter aim is to discover finer-grained word semantic differences in the topic modeling process, such as different entities or standpoints, and handle the disambiguation problem.

The experiments for our models on the three tasks are conducted on their corresponding standard datasets, respectively. For the sequential topic modeling task, our experiments conducted on six standard datasets on topic modeling show that our proposals outperform the state-of-the-art models, in terms of the accuracy of topic modeling, the quality of topic clustering, and th effectiveness of outlier detection. For the topic segmentation task, experimental results show that our proposal also produces significant improvements in both topic coherence and topic segmentation on three standard datasets. For the WSD task in topic modeling, our experiments on three standard datasets show that our proposal outperforms other state-of-the-art methods in terms of word sense estimation, topic modeling, and document classification.

Acknowledgments

First of all, I would like to thank Professor Einoshin Suzuki for his strict academic training during my doctoral course. The training helps me to improve the capability of logical thinking, critical thinking and academic writing. Without these training, I would never have finished this thesis.

I am deeply grateful to Professor Einoshin Suzuki, Professor Hiroto Saigo and Professor Tetsu Matsukawa for their supervision and extensive discussions on my published conference and journal papers, which are the foundation of this thesis. Many thanks to Professor Tetsu Matsukawa for helping me with daily affairs. Special thanks also go to Dr. Bin Tong and Dr. Kaikai Zhao, who gave me a lot of suggestions during these years.

Also, I would like to thank present and past members in Suzuki Lab, Saigo Lab and Ikeda Lab. My thanks go to Yuanyuan Li, Muhammad Fikko Fadjrimiratno, Qingpu Yang, Ruihan Wang, Kang Zhang, Qiming Zhou, Ning Dong, Chang Liu and Yi Zhou, for their assistances in various ways during my Ph.D. studies. Particular acknowledgement is also due to Yi Zhou, Yuanyuan Li and Qingpu Yang, who offered me great help at my beginning life in Japan.

Last but most importantly, I am forever grateful to my parents for their understanding, endless patience and encouragement when it was most required.

Contents

Abstract i

Acknowledgments iii

Contents iv

List of Figures viii

List of Tables xii

1 Introduction 1

1.1 Background and Motivation 1

1.1.1 Sequential Document Topic Modeling 3

1.1.2 Document Topic Segmentation 6

1.1.3 Word Sense-Aware Topic Modeling 8

1.2 Contribution of This Thesis 11

1.3 Thesis Organization 14

2 Topic Modeling for Sequential Documents Based on Hybrid Inter-Document Topic Dependency 15

2.1 Overview 15

2.2 Sequential Topic Modeling 18
2.2.1 Sequential Documents Dependencies 18
2.2.2 Sequential Outlier Topic Model 26
2.2.3 Parameter Estimation 27
2.2.4 Fine-grained SOT 30
2.3 Experiments 33
2.3.1 Datasets 33
2.3.2 Baseline Models 34
2.3.3 Perplexities 35
2.3.4 Outlier Detection 38
2.3.5 Emerging Topic Detection 42
2.3.6 Running Time 46
2.3.7 Parameter Sensitivities 47
2.4 Summary 49

3 Context-Aware Latent Dirichlet Allocation for Topic Segmentation 50
3.1 Overview 50
3.2 Context-Aware Topic Modeling 52
3.2.1 Context Word Pairs-Topic Distribution 52
3.2.2 Context-Aware Latent Dirichlet Allocation 55
3.2.3 Parameter Estimation 56
3.2.4 Topic Coherency Ratio 57
3.3 Experiments 59
3.3.1 Topic Coherence 61
3.3.2 Topic Segmentation 62
3.4 Summary 65

4 Adaptive and Hybrid Context-Aware Fine-Grained Word Sense Disambiguation in Topic Modeling Based Document Representation 66

4.1 Overview 66

4.2 Sense Aware Topic Modeling 70

4.2.1 Bag-of-Senses 70

4.2.2 Hybrid-Context based Sense Estimation 71

4.2.3 Word Sense Generation 74

4.2.4 Adaptive Context Estimation 76

4.2.5 Model Description 77

4.2.6 Parameter Estimation 79

4.3 Experiments 82

4.3.1 Experimental Setup 83

4.3.2 Quantitative Analysis 83

4.3.2.1 Sense Estimation Quality 84

4.3.2.2 Document Classification 85

4.3.2.3 Topic Modeling Accuracy 88

4.3.2.4 Effectiveness of the Adaptive Context Window Length 89

4.3.2.5 Efficiency Comparison 90

4.3.3 Qualitative Analysis 92

4.3.3.1 Dataset-Specific Word Sense Discovery 92

4.3.3.2 Word Sense Evolution Detection 95

4.3.3.3 Effectiveness of "Bag-of-Senses" 98

4.4 Summary 100

5 Conclusions and Future Work 101

5.1 Conclusions 101

5.2 Future Work 103

Bibliography **108**

Published Papers **121**

List of Figures

1.1 Illustrations for (a) plate notation of Latent Dirichlet Allocation (LDA [7]), (b) diagram of topic model which uses document-level context, (c) diagram of topic model which uses word-level context for topic assignment, and (d) diagram of topic model which uses word-level contexts in document generation, where w' refers to a context word, S is the size of context window, θ_d refers to the topic distribution of document d, ϕ is the word distribution for each K topics, α and β are their hyper parameters. 2

1.2 Example of a part of topic evolution in sequential documents. Blocks in different colors represent documents belonging to various topics. We assume that D_{i-1} and D_i belong to topics 2 and 1, respectively and D_{i+1} is an outlier. Note that, in this example, we assume each document belongs to only one topic. 4

2.1 Schematic illustration of the estimated topic distributions of documents without considering dependency, with single-dependency and with hybrid dependency. Gray points refer to outliers and the other points represent documents belonging to different topics. Single-dependency based methods are likely to decrease the dissimilarity of the two topic clusters. By considering the hybrid topic dependencies, the intra-topic similarity of each topic becomes higher and the dissimilarity of the two topic clusters are preserved. 20
2.2 Schematic illustration of hybrid inter-document topic dependencies. The arrows represent the Consecutive Dependency, the Trend dependency, and the Independency, where the thickness of each arrow corresponds to its probability. 25
2.3 Graphical Models for Sequential Outlier Topic Model (SOT). 26
2.4 ROC curves in four datasets of MOP = 0.2 (a), 0.6 (b) and 0.8 (c). . . 37
2.5 AUCs over datasets of different MOP rates. 38
2.6 Values of generated subinterval length λ ($\lambda \in [1,9]$) for each document and visualization for outlier detection results in two cases of (a) MOP = 0.2 and (b) MOP = 0.8. The second line of each subfigure is the test sequences and different topics are distinguished by color. The following lines represent the positions of true positives and are labeled in red. 39
2.7 Intra-class and inter-topic distances on sequences with MOPs of 0.2 and 0.8. 41
2.8 Novelties on different datasets (100 topics). 42
2.9 AUC under different values of L in both datasets (MOP=0.2 (a) and 0.8 (b)). 47

2.10 F1-Score under different outlier sensitivity factor ε in both datasets. . 48

3.1 Schematic illustration of a topic assignment for word "Liverpool" with and without considering its context words (respectively labeled by red and blue). We see that if "Liverpool" co-occurs with word "football" in the same context, it is more likely to be assigned to the topic of "sports", while "geography" if co-occurs with "population". 53
3.2 Graphical model for Context-Aware LDA. 54
3.3 NPMIs of different L values (a-b) and different topic numbers k (c-h). 61
3.4 PK and WindowDiff scores in terms of the number K of topics in (a) Wikicities$^+$, (b) Wikielements$^+$ and (c) CellphoneReviews$^+$. The stacked part above each bar is the improvement from RTM algorithm. 64

4.1 Comparison examples of the sense clusters in a dictionary (Figure 4.1 (a)) and the dataset-specific sense clusters (Figure 4.1 (b)). 71
4.2 Plate notation of HCT. 77
4.3 Comparison of average similarities between vectors of each word and its Top-n ($n \in [1,2000]$) nearest words based on the cosine similarity on 20NG dataset. 84
4.4 F1-Scores of HCT and HCT-L with different values of h on T-COM, 20NG and Tweet. 90
4.5 Running time per iteration (s) on 20NG of all the baselines and HCT-L with $h = 5,10,20$ (denoted by HCT-L5, HCT-L10 and HCT-L20, respectively). 91

4.6 Visualization for each example word w by their Word Sense Vectors ($\boldsymbol{v}_w$) and the corresponding Local Sense Weights (μ_w) in the 20NG. Each point in (a-f) or bar in (g-l) refers to a word item in the dataset, where each color corresponds to a sense cluster. 93

4.7 Silhouette Coefficients for each example words with different clustering numbers of kMeans. 95

4.8 Visualization for topic vectors of word pairs "*Fish-Oil*" and "*Raynaud*" (a) and "*Indomethacin*" and "*Alzheimer*", with only the topic dimensions whose top-10 high-frequency topic words contain word families of "prevent" and "treat". 98

4.9 Comparison of document vectors on 20NG of the traditional"Bag-of-Words" based topic model (LDA [7]) and the "Bag-of-Senses" based HCT. Each point corresponds to a document, where the red, blue, and yellow ones refer to the documents containing the word "*key*" with its top-3 high frequent word sense clusters, respectively. Note that, documents containing multiple senses are labeled by the color of the sense with the largest number. 99

List of Tables

2.1 Comparison of perplexities with different latent topic numbers K. Bold fonts highlight the best results. 36

2.2 Comparison of Purities and NMIs (the mean and the standard deviation). Bold fonts highlight the best results ($K = 100$). 44

2.3 Time Cost (Seconds) per Iteration on TDT Collection 46

3.1 Topic segmentation results. PK and WD scores are in %. Bold fonts indicate best scores yielded by the models except for C-LDA-R and * indicates the best scores among all the models. 62

4.1 Comparison of the average similarities between the sense clusters of each word ($\overline{C_{sc}}$). 84

4.2 Comparison of the classification performance and NPMI on "Tweet", "T-COM" and "20NG". 86

4.3 Context words for each sense cluster of the example words. Bold fonts indicate the high-frequency context words which help clarify the semantic difference. The color for each cluster symbol $\boldsymbol{c}$ corresponds to that of each cluster in Figure 4.6. 96

4.4 Interpretations in the Longman Dictionary for the generated sense clusters. c_i(s) in each row represents the possibly related cluster(s). . . . 97

Chapter 1

Introduction

1.1 Background and Motivation

Topic modeling is an unsupervised method which takes a set of documents as input and automatically clusters word groups and similar expressions that best characterize the documents. It is widely used for dealing with sparse high-dimensional features of text data and finding latent semantic relationships between documents, such as Latent Dirichlet Allocation (LDA) [7]. It takes a global view of the word distributions across the corpus to assign a topic to each word occurrence and generate a topic distribution for each document. With the progressive prevalence of mobile devices in recent decades, massive text data are continuously generated in various forms, e.g., news and tweets. For these different kinds of text data, we consider more about fine-grained semantic relationships between words (e.g., polysemy and synonyms) or documents (e.g., semantic relevance of sequential documents). These fine-grained semantic relations are normally hidden in the context information of the dataset. The context information can consist of two aspects: (1) document-level context, and (2) word-level context. The former refers to the surrounding documents for each sequential docu-

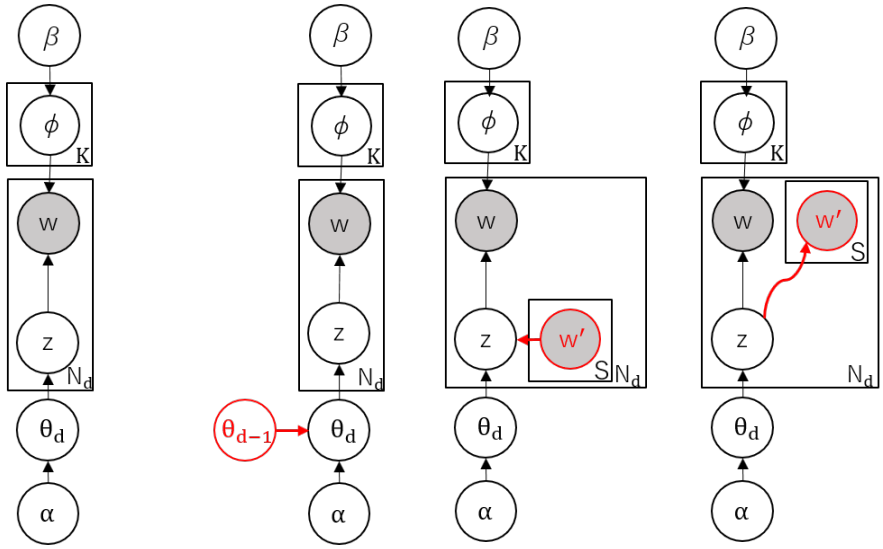


Figure 1.1: Illustrations for (a) plate notation of Latent Dirichlet Allocation (LDA [7]), (b) diagram of topic model which uses document-level context, (c) diagram of topic model which uses word-level context for topic assignment, and (d) diagram of topic model which uses word-level contexts in document generation, where w' refers to a context word, S is the size of context window, θ_d refers to the topic distribution of document d, ϕ is the word distribution for each K topics, α and β are their hyper parameters.

ment of a document sequence, while the latter represents the surrounding words for each word in a document. The usage of context information in topic modeling varies according to different target tasks.

For the type of the document-level context information, its usage in topic modeling is intuitive, i.e., using the topics of the proper surrounding documents to clarify the topic distribution of each target document. This usage is suitable for the task of topic modeling of sequential documents. However, for the type of the word-level context, there are typically two variables inside the document in a topic model: the word and

its corresponding topic, such as the variables w and z in LDA (Figure 1.1 (a)), which is a popular form of statistical topic modeling. LDA mainly consists of four parts, the word w, the topic assignment z for w, the topic distribution for a document θ_d and the probability of words belonging to each topic ϕ, which are what needs to be calculated. The algorithm tries to determine, for a given document, how many words belong to a specific topic. Documents are represented as a mixture of topics and a topic is a set of words. Therefore, there are at least two aspects of usages in the topic modeling process: (1) using the topics of the context words to assist the topic assignment for each word in a document (Figure 1.1 (c)); (2) generate both a word and its context words based on an assigned topic (Figure 1.1 (d)). For (1), it is suitable for handling the topic segmentation task for each document, i.e., dividing a document into a sequence of topically coherent segments. For (2), it is more suitable for dealing with the word sense disambiguation task in topic modeling, since the unit contained in the document is no longer a word, but a combination of each word and its corresponding context words. Different word senses of identical words can be learned by different contexts in which they are combined in the modeling process. In this study, we focus on three tasks, sequential document topic modeling, document topic segmentation and word sense-aware topic modeling, to study how to effectively use different context information to improve fine-grained semantic discovery in topic modeling.

1.1.1 Sequential Document Topic Modeling

For these sequential documents, we focus on modeling topic evolution, i.e., various and frequent emergences, growths and fades of topics, which is commonly applied to emerging topic detection tasks, such as emerging topic clustering[1] and novel topic

[1]The emerging topic clustering task is to group the sequential documents belonging to the same emerging topic into a set known as cluster without knowing their category [25].

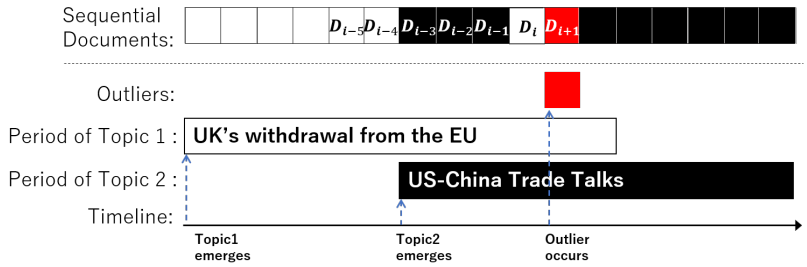


Figure 1.2: Example of a part of topic evolution in sequential documents. Blocks in different colors represent documents belonging to various topics. We assume that D_{i-1} and D_i belong to topics 2 and 1, respectively and D_{i+1} is an outlier. Note that, in this example, we assume each document belongs to only one topic.

detection[1] [25, 37]. For these tasks, the critical issue in modeling is to capture fades and emergences of each topic and discriminate outliers from each emerging topic.

Unlike batch datasets, sequential documents are collected in chronological order and usually exhibit a phenomenon called "bursty feature", which refers to a tendency that a large amount of sequential documents about a particular topic is generated over a relatively short time period [34]. Therefore, the topic of each document may be strongly related to those of its previous ones, and thus to preserve the sequential information among inter-documents in modeling is particularly essential. Recently, many topic models have been proposed for sequential documents by considering topic dependencies for each document, where the topic dependency refers to the characteristics that the topic distribution of a document depends on that of its previous one or the mean of previous document set [1, 38, 56, 99, 103]. Some of these models consider the topic dependencies between consecutive documents [1, 99, 103], while others focus on those between documents and their corresponding previous document sets [38, 56]. The former improves the sensitivity of new topic detection while the latter ensures that

[1]The Novel Topic Detection (also called First Story Detection) refers to detecting the first document to discuss a topic, which allows to know when to start a new cluster [25].

each sequential document follows the topic trend, where the topic trend refers to the dominant topic distribution of the previous document set. They all made significant progresses in various specific applications, such as DCT [56] in clustering streaming short texts and s-LDA [1] in novel topic detection. However, these single-dependency based methods are mainly based on an idealized assumption that the occurrences of documents belonging to one topic are always continuous. In real sequential documents such as those on the Internet, topics are often interwoven and outliers which belong to independent topics are frequently scattered in the sequence. For example, in texts from "Citizen Journalism"[1] such as Twitter, an outlier refers to a non-influential record, e.g., personal diaries or local anecdotes, which are frequent but not widely disseminated. Hence, these single-dependency based models have difficulties in discriminating outliers between emerging topics, as well as assigning an appropriate topic dependency for each document [37, 60].

As shown in Figure 1.2, blocks in black, white, and gray represent documents belonging to Topic 1, Topic 2 and outlier, respectively. For the case of considering topic dependencies between consecutive documents, we see that the topic distribution of D_i depends on that of D_{i-1} and the topic distribution of D_{i+1} depends on that of D_i. However, in this case, the single-dependency assumption disturbs the estimation of the topic distributions of D_i and D_{i+1}, since D_{i+1} represents an outlier whose topic should be independent, while the topic of D_i belongs to a fading topic which should depend on those of its long-term previous documents such as D_j ($j = i-4, i-5, ...$) but not D_{i-1}. On the other hand, if we consider the topic dependency of a document on its previous document set, we see that both D_{i-1} and D_i depend on their respective previous document sets, which mostly belong to the fading topic of Topic 1. However,

[1]Citizen Journalism (also known as "We Media") is a media based on public citizens who collect, report, analyze and disseminate news or information [8].

the topic estimations for D_{i-1} needs special care in this case, since D_{i-1} is a document belonging to a new emerging topic that should depend on its recent previous document D_{i-2}. This interweaving situation of fading and emerging topics often happens in sequential datasets, e.g., news and tweets [33,34,37]. In these cases, to handle outliers under such frequent topic changes, the topic dependencies should be adaptable and flexible.

In this task, we study the problem of topic modeling for sequential documents with topic evolution. The aim is to assign an appropriate dependency to each sequential document belonging to emerging topics and reserve independence for the topic distributions of outliers. For the above motivation, we give an assumption which is closer to the real data, i.e., a document has three probabilities of belonging to a fading topic, a new emerging topic and an outlier. According to this assumption, we propose two sequential topic models based on hybrid inter-document topic dependency: Sequential Outlier Topic Model (SOT) and Fine-grained Sequential Outlier Topic Model (f-SOT). Specifically, in SOT, we use the three probabilities to assign an appropriate dependency for each document in modeling. For f-SOT, to deal with more complex topic evolution, we extend SOT by considering fine-grained dependency relations.

1.1.2 Document Topic Segmentation

Topic segmentation is the task of dividing a document into a sequence of topically coherent segments [78]. Specifically, besides the topic distribution, the order of topic segments is also an essential part of document semantic information [75]. Even with the same topic distribution, different orders might represent different or even opposite standpoints. For example, a commentary at the end often determines the guidance of the public opinion, such as the coverage of politics, in particular, election campaigns

[10, 40]. The challenge of this task is to ensure both the coherency and the saliency of the topic segments, where the coherency refers to keeping long topic segments without being split, while the saliency reserving short topic segments without being absorbed with longer ones.

Conventional topic modeling, such as Latent Dirichlet Allocation (LDA) [7], has made significant progress in various specific applications by handling sparse high dimensional features and finding latent semantic relationships [54, 108]. Nevertheless, the "bag of words" based models are unable to capture the order of topics within each document. A simple solution is to consider the physical structure [3] (e.g., sentences and paragraphs) of each document and use a Hidden Markov Model (HMM) structure [6, 21, 82, 97, 100] or predefine a common canonical topic ordering to model the order of topics [20]. However, in recent decades, massive document data are continuously generated in various forms (e.g., news and postings) and from multiple modes (e.g., voice and video). The above models cannot handle these documents with no physical structure information.

Another way is to use high-frequency words as keywords of topics [83]. Detecting and utilizing keywords on the topic assignments improve the coherency of topic segments, especially in documents with well-proportioned topic distribution and sufficient keywords. However, relying heavily on extracted keywords limits the saliency of topic segments. For example, for a document with an uneven topic distribution, extracting enough keywords for all the segments is difficult. As a result, less proportionate topic segments are likely to be absorbed by topic segments with higher proportions, due to insufficient keywords.

The fundamental reason for the limited saliency and coherency is that the topic assignment of each word is highly uncertain. Most words can represent multiple topics, due to their polysemy. The distributional hypothesis [81], which states that words in

similar contexts have similar meanings, is one of the primary theories used to quantify the meaning of words according to their context (e.g., Word2vec [30]). Inspired by it, we assume that the topic of each word in a document is related to its context, that is, similar contexts correspond to similar topics. Intuitively, even if a word can be assigned to multiple topics, given its context, we can assign a corresponding topic more certainly. For example, the word "Liverpool" can belong to a topic of sports, geography or art, etc. However, if we combine it to the words in its context (e.g., "Liverpool" & "football" or "Liverpool" & "Beatles"), the assignment is much clearer.

In this task, we study the problem of how to balance saliency and coherency in topic modeling and propose a new generative model, Contextual Latent Dirichlet Allocation (Contextual-LDA). Instead of relying on keywords or the HMM structures, in the topic assignment of each word, we consider both the topic distributions and the co-occurrence distributions of context words under each topic. Besides, we also design a post-processing algorithm to optimize the generated topic segments. The proposed model enjoys two substantial merits over the state-of-the-art methods: (1) the topic of each word is generated by both topic distribution and a set of word pairs in its context, which ensures both satisfactory saliency and coherency in topic segmentation; (2) it is independent of the physical structure, such as sentences or paragraphs, and the predefined canonical topic ordering, which enhances the applicability to more datasets.

1.1.3 Word Sense-Aware Topic Modeling

Word sense-aware topic modeling is to identify the senses of polysemic words in the topic modeling process, i.e., Word Sense Disambiguation (WSD) [102] in topic modeling. Conventional solutions typically introduce an external standard knowledge library (e.g., Wikipedia, WordNet [64]) as machine-readable sense inventories for data

enrichment[1] [9, 14, 18, 23, 35]. However, in most lexicographic practice, word senses are abstractions from clusters of corpus citations, i.e., the category and the explanation for each sense strongly depend on the semantic coverage[2] of the related dataset. A more extensive semantic coverage corresponds to a coarser granularity of the word semantic division [47]. Moreover, words may also have new senses over time [9, 47]. Therefore, in these knowledge libraries, word senses which are rare, emerging, or confined to a specific domain are typically ignored [47]. For instance, for the word "*religion*" in a politics-related dataset, we may be more concerned about a finer-grained semantic division of different religious groups, e.g., "*the Islam*" and "*the Christian*", rather than just handling them abstractly as "*a belief in one or more gods*". Besides, the semantics understanding of words in a dataset may also exhibit a unique perspective. For example, for the word "*homosexual*" in a social network corpus, the position it stands for might be more valuable than its original meaning. This kind of unique perspective on semantic understanding is always dateset specific and implicitly contained in the co-occurrence pattern of each word and its context [46, 47]. Therefore, handling these fine-grained WSD problems without data enrichment is a critical issue in the document representation task. The challenge for this disambiguation problem is to divide various senses of each polysemous word while preserving the differences between different words, especially synonyms.

Several researchers model multiple word senses without data enrichment by separate context clusters [36,66,77]. Specifically, they group the contexts of all occurrences for each word into discriminated sense clusters, use these clusters to re-label the words based on the contexts of each occurrence, and then learn word or document representations based on these re-labeled words. These context clustering-based methods

[1]Data enrichment is defined as merging third-party data from an external authoritative source with an existing database of first-party customer data [62].

[2]Semantic coverage is the coverage of themes relative to a dataset.

can capture different usages of word senses in a dataset without external knowledge libraries. However, relying solely on each clustered contexts is likely to decrease the differences between synonyms, since they often occur in highly similar contexts when representing similar or identical senses, such as the sense "*belief*" for words "*faith*" and "*religion*". Besides, the context in which a word occurs is not necessarily sufficient to specify its sense. For example, "*kick*" contributes more to clarifying the sense of the word "*ball*" than "*play*" because "*play*" has a broader sense than "*kick*".

Another kind of solution is to introduce an auxiliary module which is linked through an intermediate variable $\boldsymbol{t}$, e.g., the topic assignment for each word [9, 57, 86]. Identical words combined with different values of $\boldsymbol{t}$ correspond to different senses. This approach can take advantage of the complementarity of different models and improve document representation performance. However, there are two risks for the applicability of the word sense division: (1) the differences in senses of identical words with the same value of $\boldsymbol{t}$ could be ignored, and (2) identical words with different $\boldsymbol{t}$ values could be misinterpreted as representing different senses. For example, the word "*key*" in the topic of "*electronic*", might has at least two senses of "*buttons on a keyboard*" and "*string of bits for scrambling and unscrambling*", and the sense "*buttons on a keyboard*" may correspond to at least two topics of "*electronic*" and "*music*". Therefore, it is not always appropriate to impose such a semantic division for each word.

Either of these two kinds of solutions seems unable to construct a common word sense disambiguation standard in document representation. The fundamental reason is that the different senses of a word are mainly assumed to be independent and their intrinsic relationships are ignored. These relationships are an essential basis to clarify the usage differences in other words. For example, the difference between the senses of "*belief*" for "*religion*" and "*faith*" lies in that "*faith*" in something does not necessarily pre-suppose that the belief could not be proven wrong, while "*religion*" is not [67].

Such internal differences of synonyms are challenging to be captured only according to the sense related to their contexts in which they occur, and should also depend on their other senses, e.g., another sense "*ceremonies and duties related to a belief*" of "*religion*" may help clarify its difference to "*faith*" [95]. Therefore, in estimating a word sense, the context of the word where it occurs, and the contexts of its other senses should both be considered in a weighted integrating manner. The former gives a semantic division for identical words, while the latter provides discrimination for different words with the same (or similar) senses. Moreover, the context window length of each word also needs a special care since only the context related to the word sense should be taken into account in sense estimation.

In this task, we focus on the problem of fine-grained word sense disambiguation in topic modeling. We propose a hybrid context based word sense aware topic model (named HCT), where each sense of a word is estimated by integrating their topic distributions of both the context words in which it occurs and those of its other occurrences. Besides, we introduce the "Bag-of-Senses" (BoS) assumption that a document is a multiset of word senses, based on which HCT generates a word sense instead of the words themselves. The proposed model enjoys two substantial merits over the state-of-the-art methods: (1) no data enrichment or auxiliary module is needed, (2) it is an end-to-end model in which the topic vectors for hybrid contexts as well as their weights for each word are all considered as variables and learned jointly.

1.2 Contribution of This Thesis

In this doctoral thesis, we make three main contributions on the problems of document topic modeling based on context information. The contributions are briefly summarized as follows.

- In the task of sequential documents topic modeling, we consider the document-level context information and propose a new topic model for sequential documents with topic evolution based on hybrid inter-document topic dependency. Topic modeling for sequential documents is the basis of many attractive applications such as emerging topic clustering and novel topic detection. Sequential documents such as news and social media information streams prevail on the Internet and gain increasing importance. Most of the existing topic models introduce their inter-document dependencies between topic distributions for these tasks. Considering such dependencies enables preserving sequential relationships between documents and effectively capturing fades or emergences of each topic. However, they basically consider only one kind of dependency, e.g., the topic distribution of a document solely depends on that of its previous one or the mean of the fixed number of the previous documents. In a real situation, adjacent emerging topics are often intertwined, and outliers which belong to independent topics are scattered in the sequence. These single-dependency based models have difficulties in handling the topic evolution in such multi-topic and outlier mixed sequential documents. To solve this problem, our method considers three kinds of topic dependencies for each document to handle documents belonging to a fading topic, an emerging topic, or an independent topic. Moreover, to deal with more complex topic evolution, we extend SOT by considering fine-grained dependency relations.

- In the task of topic segmentation, we use the word-level context information to assist topic assignment. Specifically, we propose a new generative model for topic segmentation based on Latent Dirichlet Allocation. The task is to divide a document into a sequence of topically coherent segments, while preserving long

topic change-points (coherency) and keeping short topic segments from getting merged (saliency). Most of the existing models either fuse topic segments by keywords or focus on modeling word co-occurrence patterns without merging. They can hardly achieve both coherency and saliency since many words have high uncertainties in topic assignments due to their polysemous nature. To solve this problem, we introduce topic-specific co-occurrence of word pairs within contexts in modeling, to generate more coherent segments and alleviate the influence of irrelevant words on topic assignment. We also design an optimization algorithm to eliminate redundant items in the generated topic segments.

- In the task of handling WSD problem in topic modeling, we use the word-level context information to document generation. Specifically, we propose a hybrid context based topic model for word sense disambiguation in document representation. Traditional methods mainly rely on knowledge libraries for data enrichment; however, semantics division for a word may vary from different domain-specific datasets. We aim to discover more particular word semantic differences for each input dataset and handle the disambiguation problem without data enrichment. The challenge for this disambiguation is to (1) divide various senses for each polysemous word while (2) preserve the differences between synonyms. Most of the existing models are either based on separate context clusters or integrating an auxiliary module to specify word senses. They can hardly achieve both (1) and (2) since different senses of a word are assumed to be independent and their intrinsic relationships are ignored. To solve this problem, we estimate a word sense by both the context in which it occurs and the contexts of its other occurrences. Besides, we introduce the "Bag-of-Senses" (BoS) assumption: a document is a multiset of word senses, and the senses are generated instead of

the words.

1.3 Thesis Organization

The rest of this thesis is organized as follows. In Chapter 2, we first introduce a sequential topic model based on hybrid inter-document topic dependency: Sequential Outlier Topic Model (SOT) and Fine-grained Sequential Outlier Topic Model (f-SOT). In Chapter 3, we introduce a new generative model, Context-Aware Latent Dirichlet Allocation (C-LDA), for topic segmentation. In the topic assignment, we consider both the topic distributions and the topic-specific occurrence of word pairs in contexts. In Chapter 4, we introduce a hybrid context based word sense aware topic model (named HCT), where each sense of a word is estimated by integrating their topic distributions of both the context words in which it occurs and those of its other occurrences. In Chapter 5, we conclude the thesis and discuss future work.

Chapter 2

Topic Modeling for Sequential Documents Based on Hybrid Inter-Document Topic Dependency

2.1 Overview

In this chapter, we restrict our attention to the document-level context based topic modeling for sequential documents. For sequential topic modeling, Dynamic Topic Model (DTM) [5] is one of the first proposals which handle topic drifts by modeling the topic dependencies between consecutive documents via Gaussian process. Based on DTM, many related studies have been proposed, such as Topic Tracking Model (TTM) [38], Temporal Latent Dirichlet Allocation (TM-LDA) [101] and Topic over Time Model (ToT) [99]. These dependencies are however obtained through Gaussian distributions, the expectation of which corresponds to the previous parameters. To more sensitively capture the topic evolutions and transitions, Streaming Latent Dirichlet Allocation (s-

LDA) [1] uses Dirichlet distributions to model the document-specific topic distributions and offers flexibilities over the consecutive topic dependencies. Nevertheless, relying solely on the consecutive documents might not supply rich enough content to infer a per-document multinomial distribution of topics [56]. To solve this problem, several studies extend the scope of the topic dependency on long-term history [55, 56]. Dynamic Clustering Topic Model (DCT) [56] considers the dependency between topic distributions of documents and their corresponding previous document sets. Emerging Topic Tracking Model (ETT) [37] models topic evolution by generating latent topics from word co-occurrences and estimating novelty for each word from the sequence with a given time decay. They all have made significant progress in several directions, e.g., short-text data clustering and semantic correlation detection. However, they are more suitable for datasets with clear transitions of emerging topics and with outliers filtered out beforehand [1, 56].

Several neural network based methods can also be utilized in sequential documents modeling, such as GT-Sem [109], Transformed-W2V [49], Aligned-W2V [32] and Dynamic Word2vec [107]. These methods capture the topic evolution of sequential documents by learning the semantic changes of words in the document sequence. The word semantic changes are basically learned by a similar two-step pattern: (1) learning static word embeddings in each time slice separately, and (2) constructing a transformation function of words between any two time slices to associate each dimension of word embeddings across time slices. All of them have provided powerful performance in many sequential document related NLP tasks. However, these time-slicing based embedding methods are difficult to detect outlier documents, since outliers are mixed into each time slice and used to learn word vectors together with other documents. Dividing a corpus into separate time slices may also result in a too-small training set to train an accurate word embedding. Besides, the word embedding models are based

on a view of the local word collocation patterns that are observed in a text corpus while the latent topic models take a more global view of the word distributions across the corpus to assign a topic to each word occurrence. Many studies have shown that these two paradigms are complementary in how they represent the semantics of documents [58, 87].

In this chapter, we study the problem of topic modeling for sequential documents with topic evolution. The aim is to assign an appropriate dependency to each sequential document belonging to emerging topics and reserve independence for the topic distributions of outliers. For the above motivation, we give an assumption which is closer to the real data, i.e., a document has three probabilities of belonging to a fading topic, a new emerging topic and an outlier. According to this assumption, we propose two sequential topic models based on hybrid inter-document topic dependency: Sequential Outlier Topic Model (SOT) and Fine-grained Sequential Outlier Topic Model (f-SOT). Specifically, in SOT, we use the three probabilities to assign an appropriate dependency for each document in modeling. For f-SOT, to deal with more complex topic evolution, we extend SOT by considering fine-grained dependency relations. We compare them to state-of-the-art methods [1, 37, 56] on six standard datasets and show that our proposals are superior in terms of the accuracy of topic modeling, the quality of topic clustering, and the effectiveness of outlier detection. The rest of this section is organized as follows. Section 2.2 briefly summarizes the related work of topic models used for sequential documents. Section 2.3 presents the details of the proposed two methods and their model inference. We evaluate the method in Section 2.4 and draw a conclusion in Section 2.5.

2.2 Sequential Topic Modeling

This section describes our methodology in detail. In Section 2.2.1, we describe the target and related problems as well as details of how the dependencies affect the topic modeling for sequential documents. Section 2.2.2 and Section 2.2.3 describe our Sequential Outlier Topic Model, SOT, and its corresponding parameter estimation. Section 2.2.4 extends SOT to fine-grained SOT (f-SOT) to deal with more complex cases, such as datasets with a large number of topics occurring in each period.

2.2.1 Sequential Documents Dependencies

As a topic model, the basic task is, for a sequence of n documents $\mathbb{D} = (D_0, D_1, ..., D_n)$, to obtain the topic distribution $\boldsymbol{\theta}_{D_i}$ for each document D_i and word distribution $\boldsymbol{\phi}_{D_i}$ of D_i over K topics, where $\boldsymbol{\theta}_{D_i}$ and $\boldsymbol{\phi}_{D_i,k}$ are both assumed to obey Dirichlet distribution with hyper parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. The number K of topics is assumed fixed and one word occurrence in a document corresponds to one topic [7]. For sequential documents, the target problem is to improve the sensitivity of novel topic detection and the quality of emerging topic clustering. The sensitivity to a novel topic detection is measured by Novelty [105] score, which quantifies the freshness of a word or a topic in the sequence and we will explain it in detail in the experiment part.

To distinguish among documents belonging to each emerging topic and outlier in the modeling, there are three related problems: (1) to increase the intra-topic similarity for document topic distributions belonging to each topic, (2) to preserve the inter-topic dissimilarity between each pair of topic clusters, and (3) to discriminate outliers automatically from topic evolution. Referring to the definitions in studies [22, 88], the intra-topic simialrity S_k for a topic k and inter-topic dissimilarity I_{k_1,k_2} for two topics

k_1 and k_2 are defined as[1]:

$$S_k = \frac{1}{|\mathbb{D}_k|} \sum_{D_i \in \mathbb{D}_k} f(\boldsymbol{\theta}_{D_i}, \overline{\boldsymbol{\theta}}_{\mathbb{D}_k}), \tag{2.1}$$

$$I_{k_1,k_2} = \frac{1}{f(\overline{\boldsymbol{\theta}}_{\mathbb{D}_{k_1}}, \overline{\boldsymbol{\theta}}_{\mathbb{D}_{k_2}})}, \tag{2.2}$$

where $\mathbb{D}_k$ refers to the set of documents belonging to topic k[2]. $\overline{\boldsymbol{\theta}}_{\mathbb{D}_k}$ is the mean topic distribution of all documents in $\mathbb{D}_k$. $f(\cdot)$ refers to a similarity degree between two distributions such as the inverse of Euclidean distance, Kullback-Leibler divergence, or cross entropy. We use these evaluation measures when the ground truth is available. An outlier in sequential data is more regarded as a contextual outlier, which is anomalous in a specific context but not otherwise [13, 27]. Therefore, we define an outlier as a document belonging to a rare-occurring topic in a given time period.

Traditionally, the total probability of a topic model, e.g., LDA, is:

$$\begin{aligned} &P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \prod_{k=1}^{K} P(\boldsymbol{\phi}_{d,k} | \boldsymbol{\beta}) \prod_{d=1}^{|\mathbb{D}|} P(\boldsymbol{\theta}_{\boldsymbol{d}} | \boldsymbol{\alpha}) \prod_{m=1}^{V_d} P(Z_{d,m} | \boldsymbol{\theta}_{\boldsymbol{d}}) P(W_{d,m} | \boldsymbol{\phi}_{d,Z_{d,m}}), \end{aligned} \tag{2.3}$$

where $\boldsymbol{W}$ refers to the word set in $\mathbb{D}$ and $\boldsymbol{Z}$ refers to their corresponding topics. $\boldsymbol{\Theta}$ represents the distributions of the documents in $\mathbb{D}$. $\boldsymbol{\Phi}$ is the word distributions over K topics. $W_{d,m}$ refers to the mth word in document d. $Z_{d,m}$ represents the generated topic of $W_{d,m}$. $\boldsymbol{\phi}_{d,k}$ refers to the word distribution for document d of topic k. For

[1] These two equations are our definitions for the target problems based on the similar definitions in those two studies [22, 88].

[2] The topic k in these equations is the ground-truth topic of a document, which is determined by the topic label or topic keyword of the dataset.

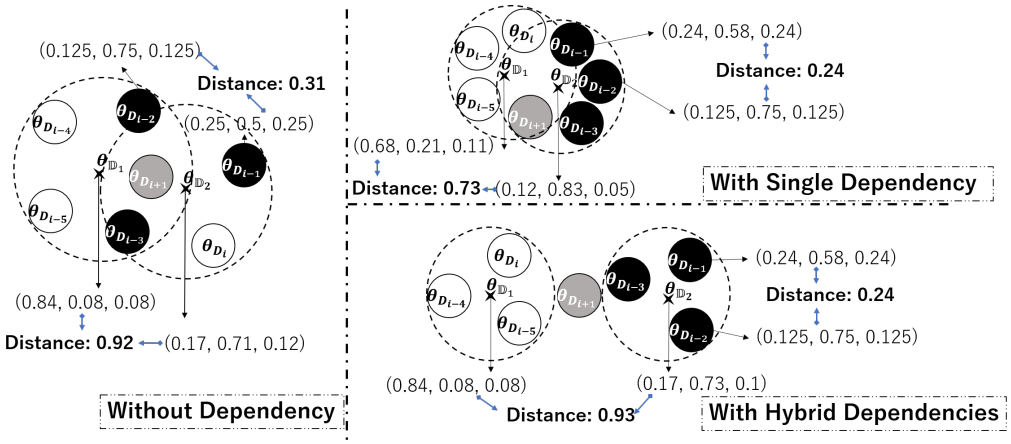


Figure 2.1: Schematic illustration of the estimated topic distributions of documents without considering dependency, with single-dependency and with hybrid dependency. Gray points refer to outliers and the other points represent documents belonging to different topics. Single-dependency based methods are likely to decrease the dissimilarity of the two topic clusters. By considering the hybrid topic dependencies, the intra-topic similarity of each topic becomes higher and the dissimilarity of the two topic clusters are preserved.

sequential topic modeling, topic distributions of the previous documents are typically used as prior information for topic estimation. Specifically, $P(\boldsymbol{\theta_d}|\boldsymbol{\alpha})$ is replaced by $P(\boldsymbol{\theta_d}|\boldsymbol{\theta}^*_d, \boldsymbol{\alpha})$ in total probability, where $\boldsymbol{\theta}^*_d$ is the topic distribution calculated based on the previous documents.

Assigning topic dependencies for documents can increase the intra-topic similarity of topic distributions for documents belonging to the same topic. Suppose that, as Figure 2.1 shows, the documents D_{i-1} and D_{i-2} both consist of 8 words and 3 latent topics (Topic 1, Topic 2 and Topic 3). Moreover, in D_{i-1}, we assume that there are 2 words belonging to Topic 1, 4 words belonging to Topic 2, and 2 words belonging to Topic 3. For D_{i-2}, we assume there is 1 word belonging to Topic 1, 6 words belonging to Topic 2, and 1 word belonging to Topic 3. According to the "Bag-of-Words"[1] assumption, the topic distributions $\boldsymbol{\theta}_{D_{i-1}}$ and $\boldsymbol{\theta}_{D_{i-2}}$ of D_{i-1} and D_{i-2} are estimated by the topic histograms of all their contained words, respectively, where $\boldsymbol{\theta}_{D_{i-1}}$=(2/8, 4/8, 2/8)=$(0.25, 0.5, 0.25)$ and $\boldsymbol{\theta}_{D_{i-2}}$=(1/8, 6/8, 1/8)=$(0.125, 0.75, 0.125)$. For simplicity, we use Euclidean distance to evaluate the similarity between the two topic distributions. The distance $dis(\boldsymbol{\theta}_{D_{i-1}}, \boldsymbol{\theta}_{D_{i-2}})$ between $\boldsymbol{\theta}_{D_{i-1}}$ and $\boldsymbol{\theta}_{D_{i-2}}$ is 0.31. On the other hand, if we consider their topic dependency and assume the new topic distribution of D_{i-1} (denoted by $\boldsymbol{\theta}'_{D_{i-1}}$) depends on $\boldsymbol{\theta}_{D_{i-2}}$, based on the the conjugate nature of the Dirichlet and the Multinomial distribution [52], $\boldsymbol{\theta}'_{D_{i-1}}$ becomes $(\frac{2+0.125}{9}, \frac{4+0.75}{9}, \frac{2+0.125}{9}) \approx (0.23, 0.54, 0.23)$ and the new distance $dis'(\boldsymbol{\theta}'_{D_{i-1}}, \boldsymbol{\theta}_{D_{i-2}}) \approx 0.26$ is shorter than $dis(\boldsymbol{\theta}_{D_{i-1}}, \boldsymbol{\theta}_{D_{i-2}})$.

However, the problem is how to assign appropriate topic dependency to preserve the inter-topic dissimilarity of each topic cluster. For methods based on single-dependencies, this problem is always unavoidable, since a document may depend on either the topic distribution of the previous document or the overall mean of the previous ones. Sim-

[1]The "Bag-of-Words" (BoW) assumes that a document is a multiset of words, disregarding grammar and even word order but keeping multiplicity [89].

ilarly, we take the six documents in Figure 1.2 and Figure 2.1 as examples. Suppose that, the rest of topic distributions of D_i, D_{i-3}, D_{i-4} and D_{i-1} are $\boldsymbol{\theta}_{D_i} = (0.75, 0.125, 0.125)$ (6 words of Topic 1, 1 word of Topic 2 and 1 word of Topic 3), $\boldsymbol{\theta}_{D_{i-3}} = (0.125, 0.875, 0)$ (1 words of Topic 1, 7 words of Topic 2 and 0 word of Topic 3), $\boldsymbol{\theta}_{D_{i-4}} = (0.875, 0.125, 0)$ (7 words of Topic 1, 1 word of Topic 2 and 0 word of Topic 3), $\boldsymbol{\theta}_{D_{i-5}} = (0.875, 0, 0.125)$ (7 words of Topic 1, 0 word of Topic 2 and 1 word of Topic 3), respectively. Therefore, the center of the two topic clusters are $\overline{\boldsymbol{\theta}}_{\mathbb{D}_1} \approx (0.84, 0.08, 0.08)$ and $\overline{\boldsymbol{\theta}}_{\mathbb{D}_2} \approx (0.17, 0.71, 0.12)$. Besides, we assume the mean topic distribution of D_{i+1} is $\overline{\boldsymbol{\theta}}_{<D_{i+1}} = (0.9, 0.05, 0.05)$ since Topic 1 dominates the topics of the previous document set. We see the original distance $dis(\overline{\boldsymbol{\theta}}_{\mathbb{D}_1}, \overline{\boldsymbol{\theta}}_{\mathbb{D}_2})$ is 0.92. If we let $\boldsymbol{\theta}_{D_i}$ be dependent on $\boldsymbol{\theta}_{D_{i-1}}$, then the topic distribution for the D_i becomes $\boldsymbol{\theta}_{D_i} = (\frac{6+0.25}{9}, \frac{1+0.5}{9}, \frac{1+0.25}{9}) \approx (0.69, 0.17, 0.14)$, and the new centers $\overline{\boldsymbol{\theta}}'_{\mathbb{D}_1} \approx (0.68, 0.21, 0.11)$ as well as $\overline{\boldsymbol{\theta}}'_{\mathbb{D}_2} \approx (0.12, 0.83, 0.05)$, and thus the new distance $dis'(\overline{\boldsymbol{\theta}}'_{\mathbb{D}_1}, \overline{\boldsymbol{\theta}}'_{\mathbb{D}_2})$ decrease to 0.72. On the other hand, if we let $\boldsymbol{\theta}_{D_{i-1}}$ depend on $\overline{\boldsymbol{\theta}}_{<i-1}$ (where $\overline{\boldsymbol{\theta}}_{<i-1} \approx \overline{\boldsymbol{\theta}}_{<i+1}$), we obtain the same result since $\boldsymbol{\theta}_{D_{i-1}}$ will be assigned to the topic of $\overline{\boldsymbol{\theta}}_{\mathbb{D}_1}$ that it does not belong to.

Based on these results, we see that if the topic of a document is different from that of the one it depends, its topic distribution is probably assigned to a wrong topic and thus the inter-topic dissimilarities for each pair of topics are decreased. Therefore, the topic distribution of a document which belongs to a new emerging topic should be dependent on its previous one while that of a document which belongs to a fading topic should depend on the previous document set. Moreover, outliers should be independent of any previous topic distributions. According to this motivation, we introduce a variable named dependency type e_d which takes one of 0, 1, 2 as its value to specify the appropriate kind of dependency for a document d, and a variable $\boldsymbol{\eta}_d = (\eta_{d,0}, \eta_{d,1}, \eta_{d,2})$ to handle the probabilities for three kinds of dependencies: (1) Consecutive Dependency ($\eta_{d,0}$), (2) Trend Dependency ($\eta_{d,1}$), and (3) Independency ($\eta_{d,2}$). Consecutive

Dependency refers to the dependency of a document topic distribution on that of its previous one while the Trend Dependency refers to the dependency of a topic distribution on the overall mean of its corresponding previous L documents, where L is named Trend Dependency factor. Independencies represent the degree to what extent those topic distributions are independent of the ones for their consecutive documents and trends.

Let $\boldsymbol{\theta}_d$ be the topic distribution of document d, $\boldsymbol{\theta}_{d-1}$ be the topic distribution of its previous document and $\overline{\boldsymbol{\theta}}_{<d-1}$ be the mean of the distributions of the previous document set $\mathbb{D}_{<d-1} = D_{d-L}, ..., D_{d-1}$. The probabilities for the three dependencies are obtained by comparing $\boldsymbol{\theta}_d$ with $\boldsymbol{\theta}_{d-1}$ and $\overline{\boldsymbol{\theta}}_{<d-1}$. Specifically, e_d obeys a Categorical distribution $Cat(\boldsymbol{\eta}_d)$ to determine whether d depends on its consecutive document, its trend or neither of them. The detailed definitions are as follows:

$$e_d \sim Cat(\boldsymbol{\eta}_d), \tag{2.4}$$

$$\boldsymbol{\eta}_d = (\eta_{d,0}, \eta_{d,1}, \eta_{d,2}), \tag{2.5}$$

where

$$\begin{aligned}
\eta_{d,0} &= \frac{f(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d-1})}{f(\boldsymbol{\theta}_d, \overline{\boldsymbol{\theta}}_{<d-1}) + f(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d-1}) + \varepsilon}, \\
\eta_{d,1} &= \frac{f(\boldsymbol{\theta}_d, \overline{\boldsymbol{\theta}}_{<d-1})}{f(\boldsymbol{\theta}_d, \overline{\boldsymbol{\theta}}_{<d-1}) + f(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d-1}) + \varepsilon}, \\
\eta_{d,2} &= \frac{\varepsilon}{f(\boldsymbol{\theta}_d, \overline{\boldsymbol{\theta}}_{<d-1}) + f(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d-1}) + \varepsilon}.
\end{aligned} \tag{2.6}$$

In our experiments, we choose the inverse of symmetric Kullback-Leibler divergence as $f(\cdot)$ for Eq. (2.6). The parameter ε refers to an outlier sensitivity factor, which

enables e_d to generate $\eta_{d,2}$ at a certain probability. We formalize the Consecutive Dependency as

$$P(\boldsymbol{\theta}_d|\boldsymbol{\alpha},\boldsymbol{\theta}_{d-1}) \sim Dir(\boldsymbol{\alpha}+\boldsymbol{\theta}_{d-1}), \tag{2.7}$$

and the Trend Dependency as

$$P(\boldsymbol{\theta}_d|\boldsymbol{\alpha},\overline{\boldsymbol{\theta}}_{<d-1}) \sim Dir(\boldsymbol{\alpha}+\overline{\boldsymbol{\theta}}_{<d-1}), \tag{2.8}$$

and the Independency as

$$P(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \sim Dir(\boldsymbol{\alpha}), \tag{2.9}$$

where *Dir* refers to Dirichlet distribution and $\boldsymbol{\alpha}$ is its fixed prior parameter.

Therefore, the topic distributions can be more accurate in terms of their ground-truth topics in estimation. When there is no topic evolution, both $f(\boldsymbol{\theta}_d,\overline{\boldsymbol{\theta}}_{<d-1})$ and $f(\boldsymbol{\theta}_d,\boldsymbol{\theta}_{d-1})$ of current document d are high and thus d has high probabilities of both the Trend Dependency and Consecutive Dependency. In this case, there is no difference between the two kinds of dependencies, i.e., $\boldsymbol{\theta}_{d-1}$ and $\overline{\boldsymbol{\theta}}_{<d-1}$ are highly similar, since they share the same topic at present. On the other hand, when topic evolution occurs, there are three cases: (1) higher $f(\boldsymbol{\theta}_d,\overline{\boldsymbol{\theta}}_{<d-1})$ and lower $f(\boldsymbol{\theta}_d,\boldsymbol{\theta}_{d-1})$; (2) lower $f(\boldsymbol{\theta}_d,\overline{\boldsymbol{\theta}}_{<d-1})$ and higher $f(\boldsymbol{\theta}_d,\boldsymbol{\theta}_{d-1})$; (3) lower $f(\boldsymbol{\theta}_d,\overline{\boldsymbol{\theta}}_{<d-1})$ and lower $f(\boldsymbol{\theta}_d,\boldsymbol{\theta}_{d-1})$. The first case indicates that the current document is more likely to belong to a fading topic, the second one indicates it has a high probability to belong to an emerging topic, and the third case indicates that it is more likely to be an outlier. Based on the above definitions, only if both of $f(\boldsymbol{\theta}_d,\overline{\boldsymbol{\theta}}_{<d-1})$ and $f(\boldsymbol{\theta}_d,\boldsymbol{\theta}_{d-1})$ are of lower values, document d can be regarded as an outlier.

Also, take the documents D_{i-1}, D_i, D_{i+1} as an example. Suppose that the topic distribution of outlier D_{i+1} is $\boldsymbol{\theta}_{D_{i+1}} = (0.1, 0.2, 0.7)$, and the mean topic distributions of

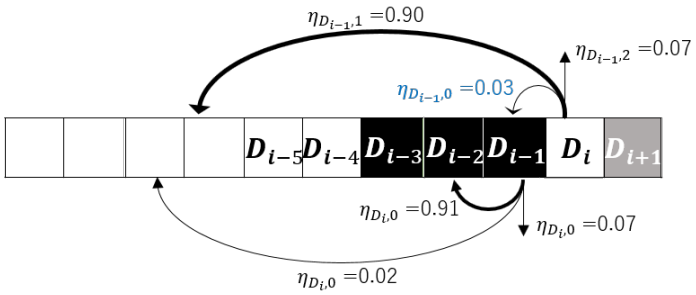


Figure 2.2: Schematic illustration of hybrid inter-document topic dependencies. The arrows represent the Consecutive Dependency, the Trend dependency, and the Independency, where the thickness of each arrow corresponds to its probability.

their previous document sets are the same $\overline{\boldsymbol{\theta}}_{<D_{i+1}} \approx \overline{\boldsymbol{\theta}}_{<D_i} \approx \overline{\boldsymbol{\theta}}_{<D_{i-1}} \approx (0.9, 0.05, 0.05)$ since Topic 1 dominates the topics of the previous document set. Based on Eq. (2.6), we obtain the $\boldsymbol{\eta}_{D_{i-1}} \approx (0.91, 0.02, 0.07)$, $\boldsymbol{\eta}_{D_i} \approx (0.03, 0.90, 0.07)$ and $\boldsymbol{\eta}_{D_{i+1}} = (0.21, 0.16, 0.63)$ for D_{i-1}, D_i, and D_{i+1}, respectively, with the parameter $\varepsilon = 2$. As shown in Figure 2.2, we see that the topic distribution of D_{i-1} has the highest probability $\eta_{D_{i-1},0}$, depending on that of its previous one D_{i-2}, the topic distribution of D_i has the highest probability $\eta_{D_i,1}$, depending on the mean topic distribution of their previous ones $\overline{\boldsymbol{\theta}}_{<D_i}$, and D_{i-2} has the highest probability $\eta_{D_{i+1},2}$, being independent. Coming back to the example in Figure 2.1, if $\boldsymbol{\theta}_{D_{i-1}}$ depends on $\boldsymbol{\theta}_{D_{i-2}}$, then the center of the two topic clusters are $\overline{\boldsymbol{\theta}}''_{\mathbb{D}_1} \approx (0.84, 0.08, 0.08)$ and $\overline{\boldsymbol{\theta}}''_{\mathbb{D}_2} \approx (0.17, 0.73, 0.1)$. The new distance $dis''(\overline{\boldsymbol{\theta}}''_{\mathbb{D}_1}, \overline{\boldsymbol{\theta}}''_{\mathbb{D}_2}) = 0.93$, which keep the original distance between the two cluster centers(0.92). These results all fit our intuition that the topic distribution for each document, e.g., D_{i-1}, D_{i-2}, is likely to be close to those of the documents with the same topic and keep outlier documents (e.g., D_{i+1}) being independent during the estimation steps.

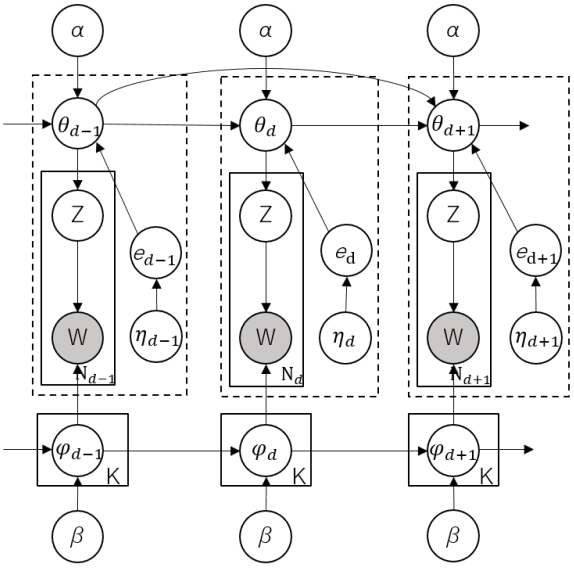


Figure 2.3: Graphical Models for Sequential Outlier Topic Model (SOT).

2.2.2 Sequential Outlier Topic Model

As Figure 2.3 shows, our Sequential Outlier Topic model (SOT) contains two new variables e_d and $\boldsymbol{\eta}_d$, where e_d represents the dependency type, which is either Consecutive Dependency, Trend Dependency or Independency, and $\boldsymbol{\eta}_d$ refers to their corresponding probabilities. For a sequence of documents $\mathbb{D}$ with a vocabulary of size V and latent topics indexed in $\{1, ..., K\}$, SOT is associated to the following generative model:

1. Generate documents according to a standard LDA model.

2. For each document d:

   (a) Calculate $\boldsymbol{\eta}_d$ based on (1) and (2).

   (b) Draw dependency type e_d: $e_d \sim Cat(\boldsymbol{\eta_d})$.

   (c) Draw topic distribution $\boldsymbol{\theta}_d$ of document d:

$$\boldsymbol{\theta}_d \sim \begin{cases} Dir(\boldsymbol{\alpha}+\overline{\boldsymbol{\theta}}_{<d-1}), & e_d = 0 \\ Dir(\boldsymbol{\alpha}+\boldsymbol{\theta}_{d-1}), & e_d = 1 \\ Dir(\boldsymbol{\alpha}), & e_d = 2 \end{cases}$$

(d) Draw word distribution $\boldsymbol{\phi}_{d,k}$ from $\boldsymbol{\phi}_d$ for each topic k of d:

$\boldsymbol{\phi}_{d,k} \sim Dir(\boldsymbol{\beta}+\boldsymbol{\phi}_{d-1,k})$

(e) For each word W in d (index by m):

i. Choose a topic $Z_{d,m}$ assignment: $Z_{d,m} \sim Cat(\boldsymbol{\theta}_d)$.

ii. Draw $W_{d,m}$: $W_{d,m} \sim Cat(\boldsymbol{\phi}_{d,z_{d,m}})$,

where $\boldsymbol{\phi}_{d,k}$ is the word distribution of d over topic k. In the process of topic generation, $\boldsymbol{\eta}_d$ is adjusted by the similarities between $\boldsymbol{\theta}_d$ with $\boldsymbol{\theta}_{d-1}$ and $\overline{\boldsymbol{\theta}}_{<d-1}$.

2.2.3 Parameter Estimation

For complex probability models, to obtain the optimal parameters directly by point estimation is difficult, therefore, except for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the parameters of our model are approximately estimated by Gibbs sampling [29], which is one of the widely used sampling methods based on Markov Chain Monte Carlo (MCMC) [11]. In the procedure, we need to calculate the conditional distribution $P_{d,m,k} = P(Z_{d,m} = k|\boldsymbol{Z}_{d,-(d,m)}, \boldsymbol{W}_d, \boldsymbol{\theta}_{d-1}, \boldsymbol{\phi}_{d-1}, \overline{\boldsymbol{\theta}}_{<d-1}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}_d)$, where $\boldsymbol{Z}_{d,-(d,m)}$ refers to the topic assignments for all words in d except word $W_{d,m}$. $P_{d,m,k}$ is computed as follows:

$$P_{d,m,k} = \int P(W_{d,m} = t|\boldsymbol{\phi}_d) P(\boldsymbol{\phi}_d|\boldsymbol{W}_{-(d,m)}, \boldsymbol{Z}_{d,-(d,m)}, \boldsymbol{\phi}_{d-1}, \boldsymbol{\beta}) d\boldsymbol{\phi}_d$$
$$\int P(Z_{d,m} = k|\boldsymbol{\theta}_d) \int P(\boldsymbol{\theta}_d|\boldsymbol{W}_{-(d,m)}, \boldsymbol{Z}_{d,-(d,m)}, e_d, \boldsymbol{\alpha}) P(e_d|\boldsymbol{\eta}_d) d\boldsymbol{\theta}_d de_d.$$

Since e_d is a discrete variable and generated from a multinomial distribution with parameter $\boldsymbol{\eta}_d$, the integral of e_d is the summation of its different values.

$$P_{d,m,k} \propto \left(\eta_{d,0} \cdot \Upsilon_{d,0} + \eta_{d,1} \cdot \Upsilon_{d,1} + \eta_{d,2} \cdot \Upsilon_{d,2}\right) \int P(W_{d,m} = t|\boldsymbol{\phi}_d) P(\boldsymbol{\phi}_d|\boldsymbol{W}_{-(d,m)}, \boldsymbol{Z}_{d,-(d,m)}, \boldsymbol{\phi}_{d-1}, \boldsymbol{\beta}) d\boldsymbol{\phi}_d,$$

where

$$\begin{aligned}
\Upsilon_{d,0} &= \int P(Z_{d,m} = k|\boldsymbol{\theta}_d) P(\boldsymbol{\theta}_d|\boldsymbol{W}_{-(d,m)}, \boldsymbol{Z}_{d,-(d,m)}, \overline{\boldsymbol{\theta}}_{<d-1}, \boldsymbol{\alpha}) d\boldsymbol{\theta}_d, \\
\Upsilon_{d,1} &= \int P(Z_{d,m} = k|\boldsymbol{\theta}_d) P(\boldsymbol{\theta}_d|\boldsymbol{W}_{-(d,m)}, \boldsymbol{Z}_{d,-(d,m)}, \boldsymbol{\theta}_{d-1}, \boldsymbol{\alpha}) d\boldsymbol{\theta}_d, \\
\Upsilon_{d,2} &= \int P(Z_{d,m} = k|\boldsymbol{\theta}_d) P(\boldsymbol{\theta}_d|\boldsymbol{W}_{-(d,m)}, \boldsymbol{Z}_{d,-(d,m)}, \boldsymbol{\alpha}) d\boldsymbol{\theta}_d.
\end{aligned}$$

By the definition of Dirichlet distribution, conditional distribution $P_{d,m,k}$ can be further simplified as a sum of expectations of Dirichlet distributions as (10).

$$\begin{aligned}
P_{d,m,k} =& E_{Dir(\boldsymbol{\phi}_{d-1}+\boldsymbol{\beta})}(\phi_{d,k,t}) \Big[\eta_{d,0} E_{Dir(\overline{\boldsymbol{\theta}}_{<d-1}+\boldsymbol{\alpha})}(\theta_{d,k}) \\
&+ \eta_{d,1} E_{Dir(\boldsymbol{\theta}_{d-1}+\boldsymbol{\alpha})}(\theta_{d,k}) + \eta_{d,2} E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k})\Big]
\end{aligned} \tag{2.10}$$

The conditional probability is finally obtained by computing the expectations of the four Dirichlet distributions. According to the definition of the expectation of Dirichlet Distribution, we obtain the conditional probability $P_{d,m,k}$:

$$\begin{aligned}
P_{d,m,k} \propto & \frac{n^t_{k,-(d,m)} + \phi_{d-1,k,t} + \beta}{\sum_{f=1}^{V} n^f_{k,-(d,m)} + \phi_{d-1,k,f} + \beta} \Big[\eta_{d,0}(n_{d,k,-(d,m)} + \overline{\theta}_{<d-1,k} + \alpha) \\
&+ \eta_{d,1}(n_{d,k,-(d,m)} + \theta_{d-1,k} + \alpha) + \eta_{d,2}(n_{d,k,-(d,m)} + \alpha)\Big],
\end{aligned} \tag{2.11}$$

Algorithm 1: Gibbs sampling algorithm for SOT

Input: Sequence of D documents of length N_d; number N_{iter} of training iterations

Output: Topic distribution $\boldsymbol{\theta}_d$ of each document d; word distribution $\boldsymbol{\phi}_d$ specific to topics; probabilities $\boldsymbol{\eta}_d$ of different dependency types

Initialize topic assignments randomly for all words in document set D

for *iteration = 1 to* N_{iter} **do**

  for d *= 1 to* $|D|$ **do**

    Calculate probabilities of document types $\boldsymbol{\eta}_d$ by Eq. (2.6)

    for m *= 1 to* N_d **do**

      Generate a topic $z_{d,m}$ from $\boldsymbol{P}_{d,m}$ by Eq. (2.11)

      Update $n_{d,z_{d,m}}$ and $n^{d,m}_{d,z_{d,m}}$

Compute the posterior estimates $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_d$ for each document d

where $n^{t}_{k,-(d,m)}$ is the number of word t belonging to topic k except $W_{d,m}$. $n_{d,k,-(d,m)}$ represents the number of all words in document d belonging to topic k. Based on $P_{d,m,k}$, we can obtain the topic distribution $\boldsymbol{P}_{d,m}$. From Eq. (2.11) and Eq. (2.6), we see that each word generation is more likely to be influenced by documents with similar topic distributions. Our Gibbs sampling algorithm is shown in Algorithm 1.

2.2.4 Fine-grained SOT

The frequency of topic changes varies in different kinds of datasets. Some sequences may contain multiple topic evolution. For example, for a news data sequence, there might be multiple topics of different domains (e.g. politics, economics and art) in the same period. For this kind of cases, the mean of the previous document set is uninformative and it is hard to obtain an obvious trend. To assign appropriate dependencies, it is necessary to split the original interval of length L into fine-grained subintervals and specifically compare the similarities of topic distributions for each subinterval. In this case, we extend the variable η_d to the probability of dependencies of the current document on each subinterval and its Independency. For the dth document under subinterval length l, the probability of dependency on the ith subinterval is calculated as

$$\eta_{d,i} = \frac{f(\boldsymbol{\theta}_d, \overline{\boldsymbol{\theta}}_{d,i}),)}{\sum_{s=1}^{\lceil \frac{L}{l} \rceil} f(\boldsymbol{\theta}_d, \overline{\boldsymbol{\theta}}_{d,s})) + \varepsilon}, \tag{2.12}$$

where s refers to the index of subintervals. Similar to SOT, the probability of Independency is calculated by

$$\eta_{d,\varepsilon} = \frac{\varepsilon}{\sum_{s=1}^{S} f(\boldsymbol{\theta}_d, \overline{\boldsymbol{\theta}}_{d,s})) + \varepsilon}. \tag{2.13}$$

However, to select an appropriate length for subintervals is difficult, since the complexity of topic evolution depends on the sequences. Since our goal is to assign an appropriate dependency for a document by comparing its topic distribution with the mean topic distribution for each subinterval, the uncertainty of the topics in the subinterval should be as low as possible. Therefore, we can determine the optimal subinterval length by calculating the uncertainty of $\boldsymbol{\eta}_d$ under different lengths. Information Entropy [85] is a widely used method to evaluate the uncertainty [84] for a set of distributions. Moreover, to exclude the influence of the number of sub-intervals on the entropy, we use the Mean Information Entropy [79] $\overline{H}$ in our method to evaluate the uncertainty for $\boldsymbol{\eta}_d$. Therefore, we can estimate the optimal subinterval length by calculating $\overline{H}_l$ under different lengths l. Specifically, we define $I_{d,l}$ as the inverse of the mean entropy for $\boldsymbol{\eta}_d$ under subinterval length of l ($l \in [1, L-1]$) as follows:

$$I_{d,l} = \frac{1}{\overline{H}_l(\eta_d)}, \tag{2.14}$$

where $\overline{H}_l(\boldsymbol{\eta}_d)$ is the Mean Information Entropy for $\boldsymbol{\eta}_d$ of subinterval length l:

$$\overline{H}_l(\boldsymbol{\eta}_d) = -\frac{1}{\log S}\left[\sum_{s=1}^{S} \eta_{d,s} \log \eta_{d,s} + \eta_{d,\varepsilon} \log \eta_{d,\varepsilon}\right], \tag{2.15}$$

where $S=\lceil \frac{L}{l} \rceil$ is the number of subintervals and s is its index.

Based on the above definitions, we introduce a new variable λ_d to control the subinterval length, where λ_d obeys a Categorical distribution:

$$\lambda_d \sim Cat(\boldsymbol{I}_d),$$

where $\boldsymbol{I}_d = (I_{d,1}, ..., I_{d,L-1})$.

Therefore, for such sequential data flooded with frequent topic drifts, we propose Fine-grained SOT (f-SOT) model. There are two improvements of f-SOT over SOT: (1) to generate an optimal length of the fine-grained interval for each document by variable λ_d; (2) with the generated subinterval length, to replace the Trend Dependency and Consecutive Dependency by S-pairwise dependencies, where S is the number of subintervals for the previous L documents. Specifically, we increase the number of dependencies from $2+1$ (Consecutive Dependency, Trend Dependency and Independency) to $S+1$.

In parameter estimation, except for modifying the calculation of $P_{d,m,k}$, we need to integrate new hidden variable λ_d in the posterior probability. The rest of the estimation algorithm is the same. We can obtain the posterior probability in two steps. First, we calculate the posterior probability $P_{d,m,k,l}$ with a given subinterval length l. Since we just use each specific pairwise dependency to replace the mean one, we can easily obtain its formula as

$$\begin{aligned} P_{d,m,k,l} \propto & \left[\sum_{s=1}^{S} \eta_{d,s}(n_{d,k,-(d,m)} + \overline{\theta}_{d,s,k} + \alpha) + \eta_{d,\varepsilon}(n_{d,k,-(d,m)} + \alpha) \right] \\ & \cdot \frac{n^t_{k,-(d,m)} + \phi_{d-1,k,t} + \beta}{\sum_{f=1}^{V} n^f_{k,-(d,m)} + \phi_{d-1,k,f} + \beta}, \end{aligned} \tag{2.16}$$

where $\overline{\theta}_{d,s,k}$ is the mean topic distribution for the documents in the subinterval s. Then the required posterior probability $P_{d,m,k}$ can be obtained by

$$P_{d,m,k} = \int \lambda_d P_{d,m,k,l} d\lambda_d = \sum_{l=1}^{L-1} I_{d,1} P_{d,m,k,l}. \tag{2.17}$$

From Eqs. (2.12) and (2.13), the difference of f-SOT from SOT is that the former considers the dependencies of subintervals in each given window L, which makes it

possible to cope with more complex document sequences. Nevertheless, compared with other methods (including SOT), more parameters need to be estimated in f-SOT, which might increase the risk of overfitting.

2.3 Experiments

Firstly, we evaluate the modeling accuracy by analyzing the perplexities under different datasets (TDT2, Reuters, Twitter Event Detection Dataset and Weibo-84168) and topic numbers ($K = 50$, 100 and 200). In the second part, we verify the effects of our topic evolution modeling in two aspects: (1) the outlier detection capabilities under the varying complexity of the topic changes, and (2) the quality of emerging topic detection in real datasets. The analysis of the parameter sensitivity is given in the last part.

2.3.1 Datasets

- **Reuters-21578 Corpus (REU)**[1] Reuters contains 21578 documents in 135 categories. The documents in the collection appeared on the Reuters newswire in 1987.

- **Multilingual Text and Annotations Dataset (TDT)**[2] TDT2 has 11201 on-topic documents, which are extracted from different broadcasts. The number of unique words per document is 100 in average.

- **THUCNews (THU)**[3] THUC is collected from the historical data of Sina News from 2005 to 2011. It contains 740000 news documents and is divided into 14

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/
[2] https://catalog.ldc.upenn.edu/LDC2001T57
[3] thuctc.thunlp.org

candidate categories.

- **PubMed**[1] PubMed is a corpus which contains journal citations and abstracts for biomedical literature from around the world since 1970s.
- **Weibo-84168 (WEI)**[2] Weibo is a Chinese micro-blog dataset collected in 2014, which includes short texts in 12 topics from 63641 users.
- **Twitter Event Detection Dataset (TWI)**[3] Twitter is a collection of 120 million tweets, with relevance judgements for over 500 events.

Each dataset was sorted by time stamps and all the results were calculated with eight-fold cross-validation.

2.3.2 Baseline Models

To validate the proposed models SOT and f-SOT (available on Github[4]), we test the following five methods including a traditional topic model LDA, two single-dependency based methods s-LDA [1] and DCT [56], a time decay based topic model ETT [37] as well as a time-slicing based word embedding method D-W2V [107]: (a) LDA (Latent Dirichlet Allocation), which trains a topic model on the whole training data. (b) s-LDA (Streaming-LDA), which takes into account consecutive document dependencies for the task of topic tracking [1]. (c) DCT (Dynamic Clustering model), which focuses on modeling the dependencies of previous documents on a longer time-step history [56]. (d) ETT (Emerging Topic Tracking Model), which models latent topics in the word co-occurrence space and estimates word novelty based on weighted word

[1] https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[2] https://github.com/liliverpool/Dataset
[3] http://mir.dcs.gla.ac.uk/resources/
[4] https://github.com/liliverpool/SOT.git

tables with time decay [37]. (e) D-W2V (Dynamic Word2vec), which is a time-slicing based dynamic embedding methods for temporal documents [107]. For D-W2V, we respectively set three values (5, 10, 20) as the total number of its slice and call them D-W2V-5, D-W2V-10 and D-W2V-20. We use the generated topic distributions as inputs of clustering and outlier detection experiments. For the clustering experiments, we chose the kernel k-Means[1] as the clustering method, which is one of the widely used methods for clustering tasks [19]. For the outlier detection experiments, the outlier detection method of our choice is Micro-Cluster Based Algorithm (MCOD) [48]. For all these models, both hyper parameters α and β were fixed to 0.05, the Trend Dependency factor $L = 10$ by following a previous study [56] and the outlier sensitivity factor $\varepsilon = 0.15$. The sensitivities to the parameters are given in the last part.

2.3.3 Perplexities

We first evaluate the performance of our model and the topic modeling based baselines in terms of perplexity, which is widely used as an evaluation metric in conventional topic or language modeling works [7, 16]. Intuitively, it quantifies the degree of uncertainty for a topic model of assigning topics to words of each document. The value of perplexity reflects the ability of a model to generalize to unseen data. A lower perplexity score indicates better modeling performance.The perplexity is given as below.

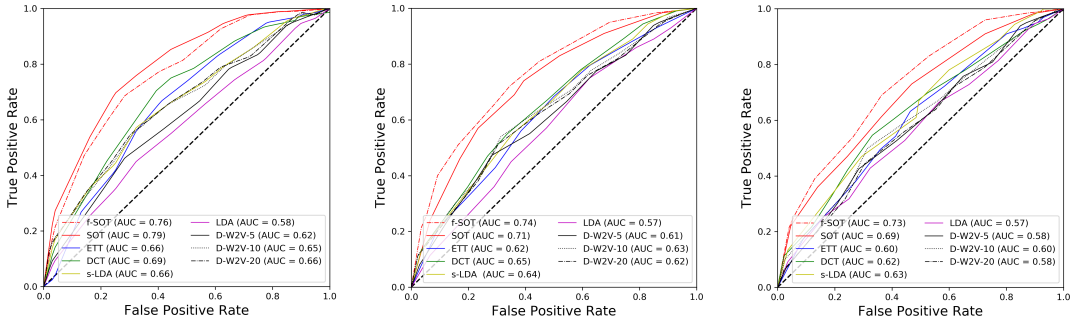
$$\text{Perplexity}(C) = \exp\left(\frac{-\sum_d \sum_n \log \sum_k \theta_{d,k} \cdot \phi_{d,k,v_{d,n}}}{V}\right), \tag{2.18}$$

where C refers to the test dataset, V is the number of total words and $v_{d,n}$ represents the nth word of document d.

[1] We used the Gaussian kernel function and set $\sigma = 0.5$.

Table 2.1: Comparison of perplexities with different latent topic numbers K. Bold fonts highlight the best results.

Models	**REU**			**TDT**			**PUB**		
	K=50	K=100	K=200	K=50	K=100	K=200	K=50	K=100	K=200
LDA	921	854	861	2578	2247	2169	2533	2066	1914
DCT	874	764	747	2314	1841	1760	2371	1863	1782
s-LDA	843	691	669	1837	1635	1572	2248	1729	1696
ETT	856	788	762	1894	1645	1749	2380	1927	1865
SOT	838	710	675	1768	1575	1558	2128	1982	1918
f-SOT	**815**	**676**	**620**	**1726**	**1464**	**1442**	**2047**	**1816**	**1742**
Models	**THU**			**WEI**			**TWI**		
	K=50	K=100	K=200	K=50	K=100	K=200	K=50	K=100	K=200
LDA	1874	1628	1661	2854	2265	2347	3359	2819	2967
DCT	1486	1325	1330	2512	2078	2126	3160	2647	2737
s-LDA	1237	1092	1152	2388	1959	2043	3088	2558	2624
ETT	1291	1134	1169	2470	2048	2106	3105	2628	2695
SOT	1138	1024	1057	2428	1982	2018	**2739**	2401	2517
f-SOT	**1115**	**971**	**986**	**2347**	**1866**	**1902**	2744	**2282**	**2369**



(a) $\mathbb{S}_{0.2}$ (b) $\mathbb{S}_{0.6}$ (c) $\mathbb{S}_{0.8}$

Figure 2.4: ROC curves in four datasets of MOP = 0.2 (a), 0.6 (b) and 0.8 (c).

We conducted experiments by varying the number of topics in set $\{50,100,200\}$. Table 2.1 shows the evolution of perplexities of different models on all test datasets with varying topic numbers. The best results are obtained with SOT or f-SOT since our models consider the dependency with other documents when generating a document topic distribution. This dependency is determined by the similarity with the topic distribution of other documents. According to Eq. (2.6), a document topic distribution is highly likely to depend on a distribution similar to itself, thus reducing its uncertainty. Moreover, the outliers are discriminated from various topic evolutions in modeling, which also contributes to the decrease of perplexities. All the testing models on Twitter and Weibo are less effective, due to the difficulty of latent topic modeling for short texts, which might cause errors when comparing similarities of topic distributions. In addition, all models performed better when the number of topics are increased from 50 to 100. For 200 latent topics, their results on the Twitter dataset began to fluctuate, and the improvements on the TDT2 and Reuters also tend to saturate; thus in the latter experiments, we set the number of latent topics to $K = 100$ and kept the other parameters unchanged.

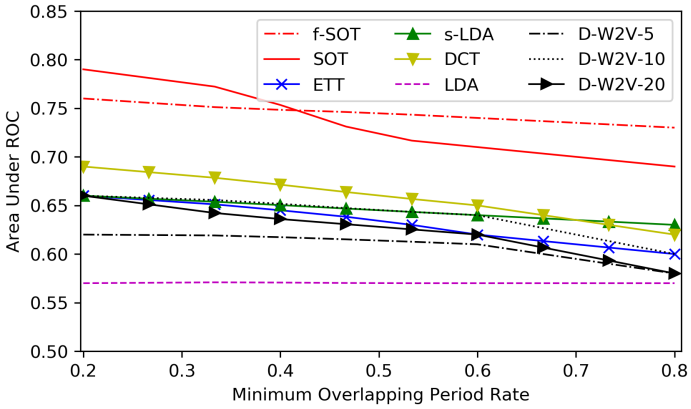
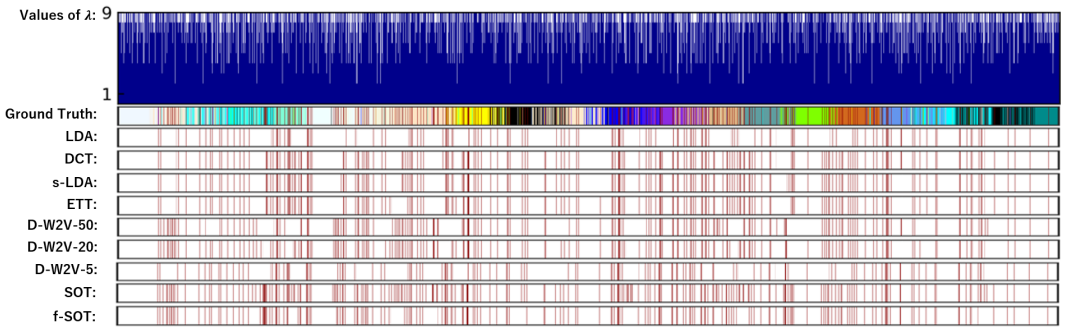


Figure 2.5: AUCs over datasets of different MOP rates.

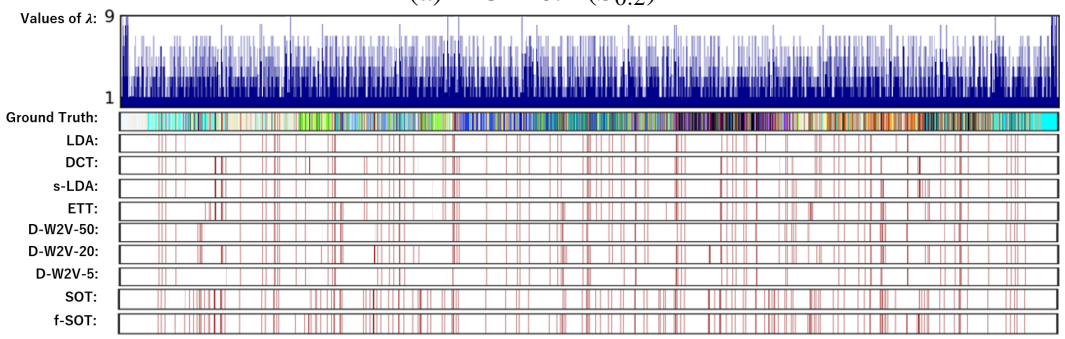
2.3.4 Outlier Detection

We conducted experiments for outlier detection based on the generated topic distributions for all the baseline methods. The testing datasets with different complexities are all sampled from TDT2. Following the outlier sampling principle of related studies about document outlier detection [44, 45], we randomly selected 20 themes of TDT2 from 4 different domains of politics, economy, society and entertainment. For each topic we chose 800 documents to construct our datasets, for a total of 16000 documents. Besides, we randomly selected 3500 documents from other different topics as outliers.

To fit the "Bursty Feature" of sequential documents, we assume that each sampled theme in the sequence belongs to a continuous sub-period, and each sub-period is overlappable. Moreover, to simulate the interleaving phenomenon of documents belonging to different sampled themes when topic change occurs, we assume that the documents in the overlapping parts of the sub-periods are randomly distributed. The constructing process consists of two phases: topic period generation and document sequence construction. For generating a sequence $\boldsymbol{S}$ with a period P_s, firstly, we randomly se-



(a) MOP=0.2 ($\mathbb{S}_{0.2}$)



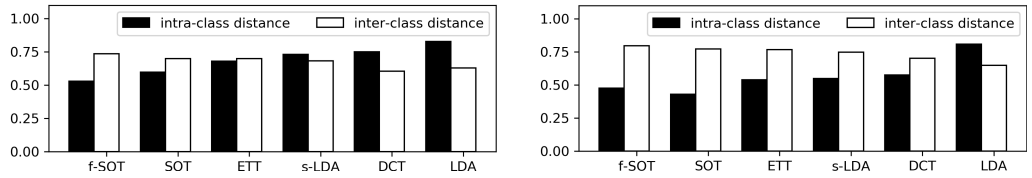
(b) MOP=0.8 ($\mathbb{S}_{0.8}$)

Figure 2.6: Values of generated subinterval length λ ($\lambda \in [1,9]$) for each document and visualization for outlier detection results in two cases of (a) MOP = 0.2 and (b) MOP = 0.8. The second line of each subfigure is the test sequences and different topics are distinguished by color. The following lines represent the positions of true positives and are labeled in red.

lected sub-period $P_{\mathbb{D}_k}$ within P_s according to a uniform distribution for document set $\mathbb{D}_k$ of each topic k, where every sub-period $P_{\mathbb{D}_k}$ is an overlapping and continuous interval. Secondly, all the documents are loaded to their corresponding sub-periods, where those in the overlapped intervals are shuffled before the loading. Therefore, the complexity of the topic evolution for a document sequence can be quantified as the ratio of the overlapping length of the subintervals to the total period. We name this ratio Minimum Overlapping Period (MOP) of topics.

For a document sequence $\boldsymbol{S}$, MOP is defined as $\text{MOP}(\boldsymbol{S}) = (\sum_k \left|P_{\mathbb{D}_k}\right| - |P_s|) / \sum_k \left|P_{\mathbb{D}_k}\right|$, where $|\cdot|$ denotes the number of documents in the period. MOP, which ranges from 0 to 1.0, is used to quantify the degree of topic evolving complexity, where the larger MOP, the more complex is the topic evolution. By changing P_s, we generate datasets of MOP from 0.2 to 0.8 to test the outlier detection performance. Note that the period P_s here does not refer to any temporal information, as we only simulate the sequences in our datasets but do not add any lag in each continuous document.

Figure 2.4 shows the Receiver Operating Characteristics (ROC) curves of the estimated models on datasets with MOP = 0.2, 0.6 and 0.8. In all the three cases, SOT and f-SOT show drastic improvements over all the baseline models. SOT outperforms other methods when MOP = 0.2 and f-SOT performs better in the other two cases. Besides, from Figure 2.5, except for LDA, the Areas Under the ROC curve (AUC) of all methods decrease as MOP increases, since there are more overlapping periods, which correspond to increasingly complicated topic evolution. LDA model neglects the dependencies of distributions between documents, and thus sequential changes made no difference to their performance. For the D-W2V methods, outliers are neglected in training; it is difficult to disambiguate these outliers from systematic topic evolution between subsequent times, in particular over short slice spans (D-W2V-20), the impact of outliers may be more notable than those in long slice spans. In addition, as shown



(a) MOP = 0.2 (b) MOP = 0.8

Figure 2.7: Intra-class and inter-topic distances on sequences with MOPs of 0.2 and 0.8.

in Figure 2.6, we visualize the topic sequences of MOP = 0.2 and 0.8 and their corresponding true positive results with 0.5 false positive rate. For sequential documents with smaller MOP, the performance are all of lower differences, while for the documents with larger MOP, our models outperform the baselines at intervals with more topics. Moreover, we see that shorter subinterval lengths are estimated in the sequential documents with more complex topic evolution. This observation also conforms with our original motivation, which is to determine the dependencies between documents through a more fine-grained pairwise comparison. However, in sequences with relatively less topic evolution, f-SOT still suffers from overfitting, which affects its accuracy in outlier detection.

Furthermore, in the cases of MOP = 0.2 and 0.8, we evaluate the average intra-topic and inter-topic similarities of topic distributions for all baseline models under different sampled themes. For intra-topic evaluation, we compute the mean Euclidean distance to the center from all documents belonging to the same theme, while the inter-topic distances are evaluated by the mean Euclidean distances between centers of all sampled themes. The shorter the distance, the higher the similarity. The normalized results are shown in Figure 2.7. By assigning appropriate weights of dependencies to each document, though there is no significant change of inter-topic similarities, our model effectively improves the intra-topic similarities, so as to emphasize the prominence of

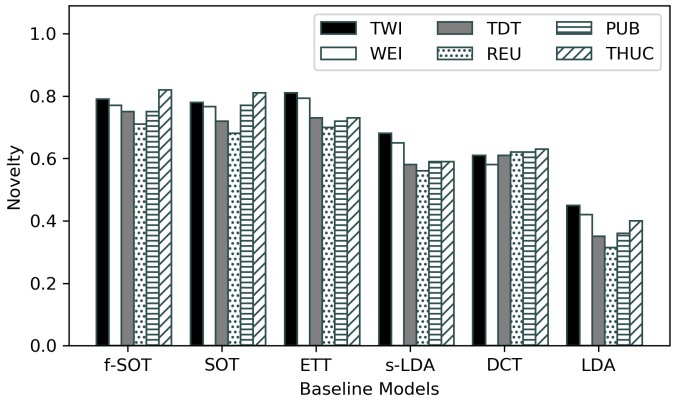


Figure 2.8: Novelties on different datasets (100 topics).

outliers in documents of various emerging topics.

2.3.5 Emerging Topic Detection

Firstly, to investigate the sensitivity of the baseline models for novel topics, we calculated novelties for our model and other topic modeling based methods in real data. Novelty [105] is used to measure the freshness of a word or a topic, which is widely applied in the tasks of emerging topic clustering and novel topic detection [25]. We adopt it to evaluate how sensitive a topic model is in detecting an emerging topic. The definition is as follows:

$$\text{Novelty}(d^{(t)}) = \frac{\left|\mathbb{W}^{(t)}\right| - \left|\mathbb{W}^{(t)} \bigcap \mathbb{W}^{(t-1)}\right|}{TK}, \tag{2.19}$$

where $|\cdot|$ denotes the cardinality of a set and $\mathbb{W}^{(t)}$ is a word set of the top-T high frequent words in document $d^{(t)}$ based on its topic distribution and word distribution. In these experiments, we choose $T = 10$.

In Figure 2.8, we show the novelty detection performance. SOT, f-SOT and ETT significantly outperform DCT and LDA. ETT shows the highest score by weighting

fading and emerging words, which largely improves the sensitivity of novel topic detection. However, by the results shown in the outlier detection experiments, both documents of outliers and emerging topics are of higher novelty for ETT, documents in emerging topics are likely to be misdetected as outliers. For SOT and f-SOT, the beginning of every emerging topic is of both weak Consecutive and Trend Dependencies, which enables documents of emerging topics be seperated from others. Besides, documents of emerging topics from the beginning can still depend more on each other according to their stronger Consecutive Dependencies, based on which the outliers are able to be discriminated from documents of emerging topics.

To further study their accuracies of emerging topic detection, we compared the performance of all baselines on clustering qualities according to their generated topic distributions. The experiments were conducted on six sampled sequences of 10 topic clusters from TDT, REU, PUB, THU, TWI and WEI, where 8000 documents are contained in TDT, REU, PUB and THU, as well as 15000 documents in Twitter, and Weibo. For these datasets, they naturally contain outlier documents, therefore, we generated another group of datasets (TDT*, REU*, PUB*, THU*, TWI* and WEI*) with outliers filtered out for comparative experiments. The metrics we used are Purity [61] and Normalized Mutual information (NMI) [74]. Purity is a simple clustering evaluation measure, which is defined as the proportion of the number of documents correctly clustered to the total and NMI is an information theoretic measure of the shared information between a clustering and a ground-truth classification. Both of them range from 0 to 1, the higher the better.

We calculated the average Purity and NMI values for cluster numbers from 10 to 20, and the results are reported in Table 2.2. SOT and f-SOT are superior to the others on all the datasets without outliers filtering. These results prove the rationality of our assumption, that a document is possibly independent of other documents, where

Table 2.2: Comparison of Purities and NMIs (the mean and the standard deviation). Bold fonts highlight the best results ($K = 100$).

Models	**Purity (100%)**					
	TDT/TDT*	**REU/REU***	**PUB/PUB***	**THU/THU***	**TWI/TWI***	**WEI/WEI***
LDA	0.663/0.684	0.606/0.621	0.612/0.636	0.649/0.673	0.512/0.533	0.557/0.571
DCT	0.684/0.707	0.638/0.662	0.645/0.671	0.696/0.718	0.547/0.561	0.578/0.591
s-LDA	0.692/0.718	0.642/**0.665**	0.653/0.668	0.705/**0.734**	0.542/0.554	0.572/0.586
ETT	0.675/0.696	0.604/0.628	0.646/0.663	0.683/0.704	0.525/0.539	0.554/0.575
D-W2V-5	0.656/0.679	0.596/0.628	0.623/0.641	0.655/0.676	0.504/0.521	0.542/0.558
D-W2V-10	0.672/0.694	0.632/0.652	0.631/0.647	0.681/0.703	0.538/0.558	0.565/0.577
D-W2V-20	0.665/0.681	0.609/0.617	0.607/0.625	0.652/0.674	0.514/0.531	0.553/0.569
SOT	0.714/0.725	**0.644**/0.653	**0.667**/**0.674**	**0.725**/0.731	0.561/0.567	0.582/0.588
f-SOT	**0.728**/**0.737**	0.641/0.649	0.661/0.668	0.722/0.728	**0.585**/**0.592**	**0.598**/**0.606**

Models	**NMI**					
	TDT/TDT*	**REU/REU***	**PUB/PUB***	**THU/THU***	**TWI/TWI***	**WEI/WEI***
LDA	0.815/0.839	0.726/0.751	0.653/0.675	0.778/0.791	0.653/0.672	0.677/0.695
DCT	0.857/0.882	0.759/0.779	0.683/0.702	0.805/0.826	0.677/0.698	0.692/0.714
s-LDA	0.877/0.902	0.772/**0.798**	0.691/0.716	0.811/**0.845**	0.672/0.686	0.688/0.695
ETT	0.831/0.855	0.743/0.767	0.653/0.686	0.792/0.807	0.679/0.693	0.682/0.693
D-W2V-5	0.805/0.824	0.717/0.733	0.665/0.689	0.767/0.792	0.645/0.656	0.669/0.681
D-W2V-10	0.847/0.875	0.749/0.775	0.671/0.694	0.789/0.813	0.671/0.684	0.683/0.692
D-W2V-20	0.817/0.833	0.726/0.742	0.643/0.671	0.762/0.791	0.654/0.661	0.671/0.679
SOT	0.884/0.891	**0.787**/0.793	**0.718**/**0.725**	**0.837**/0.841	0.675/ 0.681	0.687/0.695
f-SOT	**0.892**/**0.898**	0.784/0.791	0.715/0.724	0.834/0.841	**0.692**/**0.698**	**0.705**/**0.711**

the independent probability depends on the similarities to the topic distributions of its previous ones. For SOT and f-SOT, the case of outliers is considered and the generated topic clusters are of lower intra-topic distance, which leads to the reduction of the occurrence of "false positive" in the clustering. For other models, such as s-LDA and DCT, they assume that each document either belong to topics or an emerging topic, which leads to the inaccuracies of dependencies for documents in topics. On the other hand, for the dataset group with outlier filtering, the accuracy of all models is improved. Compared with SOT and f-SOT, the improvements of the others are more significant. Nevertheless, our proposals are still superior to the other methods in most cases. For REU* and THU* collected from two single media sources (Reuters and Sina), documents belonging to the same topic usually appear continuously with outliers filtering, and s-LDA considers the topic similarities between consecutive documents; thus, s-LDA outperforms the other methods in REU* and THU*. For the factors which affect topic modeling for sequence documents, there are frequently intertwined topics besides outliers. Although removing the interference of outliers can improve emerging topic detection performance, the single-dependency based models are still unable to deal with frequent and complex topic evolutions effectively.

For the results of the short text datasets (TWI/TWI* and WEI/WEI*), we see that the improvements in our models are still slightly better compared with the baseline models. However, compared to the performance for long document datasets, the overall performance in all models is degraded. Topic modeling or document embedding can alleviate the problem of semantic representation for documents, by mapping the high dimensional sparse text feature to the shared topic space and capturing the co-occurrence pattern of words in each document. Therefore, when the document size becomes smaller and the documents contain lower word counts, those models would be difficult to obtain accurate word co-occurrence patterns. Moreover, if the distri-

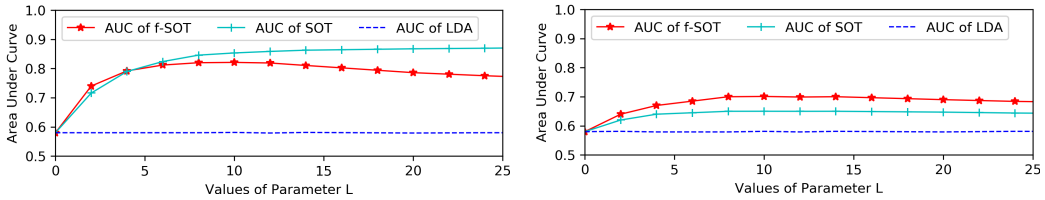
Table 2.3: Time Cost (Seconds) per Iteration on TDT Collection

Models	**Running Time (s)**					
	K=10	**K=50**	**K=100**	**K=150**	**K=200**	**K=250**
LDA	5.07	32.21	71.63	102.64	137.52	168.18
DCT	8.77	49.95	105.49	157.41	207.85	254.25
s-LDA	9.32	54.65	116.17	173.78	228.96	281.78
ETT	7.65	42.58	92.77	129.49	164.31	208.92
SOT	11.08	61.42	127.59	188.26	251.17	306.75
f-SOT	15.27	82.52	224.47	326.14	412.81	495.33

bution of documents for the topic is heavily skewed, they tend to learn more general topics supported by many documents rather than rare topics [111]. These might be the reasons of their inferior performance.

2.3.6 Running Time

For efficiency comparison, we list the average running time (per iteration) of our proposals and other baseline topic models on TDT dataset with $K = 10, 50, 100, 150, 200, 250$. The experiments were conducted on a PC with Intel i9 processor and 128GB memory. As shown in Table 2.3, we see that the time cost of all the models increases monotonously as the topic number K grows and the topic models for sequential data cost more time than the traditional topic models. The reason is that they need to estimate the inter-document topic dependencies in modeling, especially SOT, it considers three types of topic dependencies. Moreover, we see that the running time of f-SOT is always about three times of LDA over different topic numbers, because f-SOT consid-



(a) MOP = 0.2 (b) MOP = 0.8

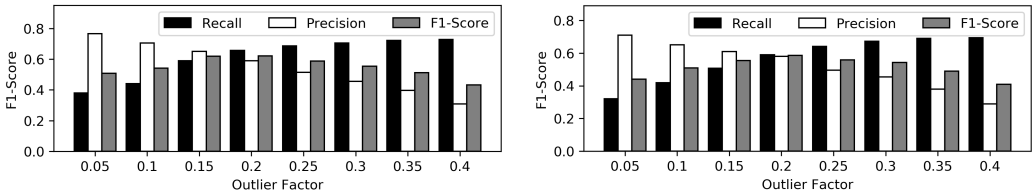
Figure 2.9: AUC under different values of L in both datasets (MOP=0.2 (a) and 0.8 (b)).

ers different sub-interval cases in each iteration. However, either the inter-document dependency types or the number of different subinterval lengths are independent of the topic numbers. Therefore, we can see that the change rate of running time for f-SOT and SOT is consistent with LDA as K increases.

2.3.7 Parameter Sensitivities

For investigating the sensitivity to different values of L in our models, we tested their outlier detection performance in datasets of MOP = 0.2 and MOP = 0.8. All parameters are fixed as the previous settings except for L and the datasets are generated the same as in the outlier detection experiments.

Figure 2.9 shows the AUCs of f-SOT, SOT and LDA with L ranging from 2 to 25. We see that in the dataset of MOP = 0.2, the AUCs of both SOT and f-SOT improve quickly with the increase of L. For f-SOT, the AUC saturates around $L = 10$, and starts to decrease when L larger than 10. Smilarly, for the results in the dataset of MOP = 0.8, the performance of the both proposed models are improved with the increase in L. However for f-SOT, saturation of the performance is observed at around $L = 5$. Moreover, the maximum values of AUCs are lower than that in the case of $MOP = 0.2$. LDA model is independent of the length L and thus exhibits stable performance.



(a) SOT (b) f-SOT

Figure 2.10: F1-Score under different outlier sensitivity factor ε in both datasets.

For SOT model, L determines the number of documents which are included in the calculation of the overall mean of the topic distributions in the previous document set. The more documents are included, the closer the overall mean is to the trend of the current topic. For the case of a dataset with frequent topic changes, a larger number of topics are aggregated within a given interval; thus, it is more difficult to quantify the overall trend accurately for datasets with high MOP. Besides, compared to SOT, f-SOT just needs to include at least one document of the same topic in the smallest possible interval. From (13) and (14) in Section 3.4, we see that if a document with the same topic has appeared in the scope of length L, continuing to grow the length only increases the risk of overfitting, especially in the case of high MOP. Therefore, in datasets with complex topic evolution, it is inappropriate to set a large value of L.

Similarly, for investigating the outlier sensitivity factor ε, we calculate the average precisions and recalls by using F1-Score [61] as the metric. The results are shown in Figure 2.10. ε determines how sensitive the model is to the outlier documents. However, excessive sensitivity results in a lower recall, which affects the performance of outlier detection as well.

2.4 Summary

In this Chapter, we proposed two topic models in topic evolution modeling based on hybrid inter-document dependencies. The first model considers Consecutive Dependency, Trend Dependency and Independency in contextual documents. For a sequence of more complex topic evolution, we improved it by considering fine-grained local dependency relations. Our experiments show that our proposals outperform baseline models, in terms of accuracy of topic modeling, clustering quality and effectiveness of outlier detection.

Chapter 3

Context-Aware Latent Dirichlet Allocation for Topic Segmentation

3.1 Overview

In this chapter, we focus on the word-level context based topic modeling for document topic segmentation task. Topic segmentation has long been studied in various topic models [6, 20, 21, 82, 97, 100], such as segHMM [6] and Bayesseg [21]. The traditional methods mainly rely on the document physical structure, which refers to the text-spans in each document, such as sentences or paragraphs [3]. They basically assume that words in the same text-span share the same topic or topic distribution. They conduct segmentation by introducing HMM structure in their topic models and modeling dependencies between consecutive text-spans. However, these approaches are unable to handle data with no structural information, which significantly limits their applicability. Moreover, in most cases, topics might evolve in long paragraphs or sections, and thus a text-span might contain multiple topics.

Recent studies have been focusing on physical structure-independent segmenta-

tion [2, 16, 83, 105]. Topic Keyword Model (TKM) [83] is a topic model based on keywords and their contexts. Its main weakness lies in handling short topic segments, which are likely to be absorbed by long topic segments due to their small numbers of keywords. Biterm Topic Model (BTM) [16] learns topics by modeling the generation of word co-occurrence patterns, which improves the sensitivity of the discovery of phrases in short text data. On the basis of the former, Bursty Biterm Topic Model (BBTM) introduces a new variable to discover bursty topics[1] [105]. These phrase-level topic modeling methods can achieve good results in discovering word co-occurrence patterns in individual short documents and require no physical structure information. However, high-frequency phrases only make up a tiny proportion of the corpus, which limits their ability to generate coherent topics in topic segmentation tasks. The main difference from our model is that they consider all distinct word pairs of each fixed-size window, while we focus on the topic-specific word pairs, which only concern the target word in the corresponding context. Copula LDA with Segmentation (SegLDA) [2] is an LDA-based model which automatically segments documents into topically coherent sequences of words. SegLDA predefines segments for each document before modeling. For each word in a segment, a topic is assigned either from the segment-specific topic distribution or the document-specific topic distribution. These distributions differentiate the main topics of a document from potential segment-specific topics, which improves the saliency of short segments. However, the two distributions are independent. Specifically, in the former distribution, a topic assignment depends only on the words within the segment, which leads to a loss of much context information in the original document.

In addition, context information is also utilized in other topic models to solve var-

[1] In their study [105], a topic is considered to be bursty in a time slice if it is heavily discussed, but not in most of the other slices.

ious specific problems in document semantic analysis [71, 106], such as Contextual Topic Model (CTM) [106] and Contextual Latent Dirichlet Allocation (Contextual-LDA) [71]. CTM considers the dependencies of topics between each sentence in document summarization while Contextual-LDA uses the topic position of each physical structure-based segment for key information detection. Different from them, we focus on solving the problem of topic segmentation by considering topic-specific word pairs in contexts.

In this chapter, we propose a new generative model, Context-Aware Latent Dirichlet Allocation (C-LDA), for document segmentation. In the topic assignment, we consider both the topic distributions and the topic-specific occurrence of word pairs in contexts. Our model enjoys two substantial merits over the state-of-the-art methods: (1) a word is generated by both the document-specific topic distribution and the topic distribution associated with each word and its context; (2) it is independent of physical structures.

3.2 Context-Aware Topic Modeling

3.2.1 Context Word Pairs-Topic Distribution

For conventional LDA and its extended models, topic assignment for each word mostly relies on topic distribution and word distribution. Although the constraints of topic distribution can alleviate the uncertainty in the topic assignment, it is still insufficient to handle documents containing multiple main topics. For example, a document on the study of modern football and the geographical distribution of England, should at least belong to two topics (geography and sports). We study the topic assignment of the word "Liverpool" in a specific location and consider its 3 related topics: sports,

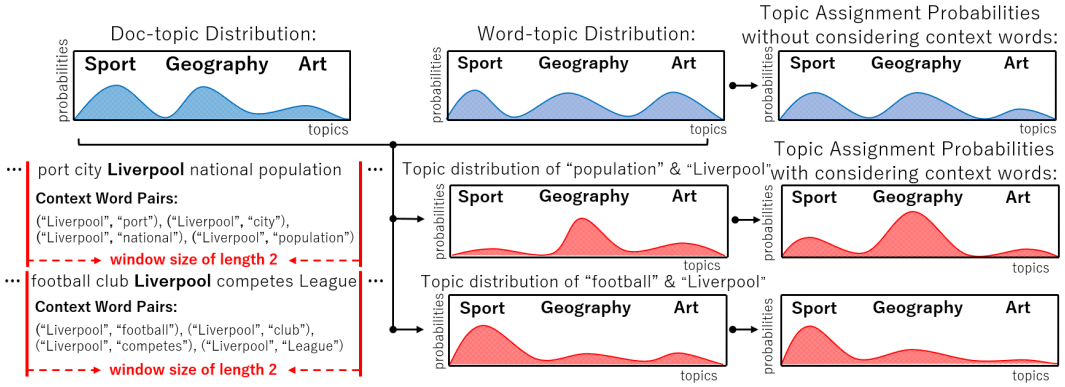


Figure 3.1: Schematic illustration of a topic assignment for word "Liverpool" with and without considering its context words (respectively labeled by red and blue). We see that if "Liverpool" co-occurs with word "football" in the same context, it is more likely to be assigned to the topic of "sports", while "geography" if co-occurs with "population".

geography, and art. As shown in Figure 3.1, for traditional topic models, although the topic distribution reduces the probability of being assigned to the topic of "art", there is still a large uncertainty between "sports" and "geography". However, by considering the frequency of co-occurrence of context words on various topics, this uncertainty can be further reduced, which also coincides with the distributional hypothesis[1] [24, 30].

Therefore, in our model, we give each word w a context window of length L and define a set of words within the window as context words $\vec{c}_w$. For the topic assignment of w, we consider the topics of word pairs $\boldsymbol{b}_w$ which consist of w and $\vec{c}_w$. $\boldsymbol{b}_w$ is defined as:

$$\boldsymbol{b}_w \triangleq \{(w, w') | w' \in \vec{c}_w\}.$$

Following LDA [7], we also assume that the topic distribution $\boldsymbol{\lambda}_w$ of all the sets of

[1]The Distributional Hypothesis is that words that occur in the same contexts tend to have similar meanings [24].

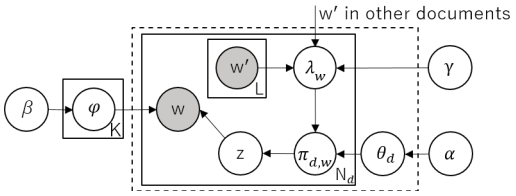


Figure 3.2: Graphical model for Context-Aware LDA.

word pairs follows a Dirichlet distribution and name it Context Word Pairs-Topic Distribution (CWTD):

$$\boldsymbol{\lambda}_w \sim Dir(\boldsymbol{\gamma}).$$

$\boldsymbol{\lambda}_w$ depends on the topic distribution of the word pairs of $\boldsymbol{b}_w$ in all other documents. By the definition of Dirichlet distribution [7], the expectation can be calculated as:

$$E_{Dir(\boldsymbol{\gamma})}(\lambda_{w,k}) = \frac{n^{\boldsymbol{b}_w}_{k,-(d,l)} + \gamma_k}{\sum_{s=1}^{K}(n^{\boldsymbol{b}_w}_{s,-(d,l)} + \gamma_s)}, \tag{3.1}$$

where $n^{\boldsymbol{b}_w}_{s,-(d,l)}$ is the total number of word pairs which are in $\boldsymbol{b}_w$ and belong to topic k in all documents without containing the lth word of document d. In the topic assignment, we reorganize the topic distribution $\boldsymbol{\theta}_d$ of a document based on the context of each word and name the reorganized topic distribution as Context-Aware Topic Distribution (CTD), denoted by $\boldsymbol{\pi}_{d,w}$. Therefore, the topic $Z_{d,w}$ for word w in d follows a Categorical distribution which is from the Dirichlet distribution $\boldsymbol{\pi}_{d,w}$ with the prior of both the topic distribution $\boldsymbol{\theta}_d$ and the CWTD $\boldsymbol{\lambda}_w$:

$$\boldsymbol{\pi}_{d,w} \sim Dir(\boldsymbol{\theta}_d + \boldsymbol{\lambda}_w), Z_{d,w} \sim Cat(\boldsymbol{\pi}_{d,w}).$$

3.2.2 Context-Aware Latent Dirichlet Allocation

As Figure 3.2 shows, we introduce four variables $\pi_{d,w}$, λ_w, w' and γ based on traditional LDA, where $\pi_{d,w}$ represents the CTD for word w in document d, λ_w is the corresponding CWTD with prior of γ and w' refers to a context word of w. Besides, θ_d represents the topic distribution of document d with prior α and ϕ_k is the word distribution of topic k with prior β. For a dataset of D documents with a vocabulary of size V and latent topics indexed in $\{1,...,K\}$, C-LDA is associated to the following generative model.

1. Generate the word-topic distribution ϕ_k for each topic k: $\phi_k \sim Dir(\beta)$.
2. For each document d:
   (a) Generate the topic-word distribution θ_d of document d: $\theta_d \sim Dir(\alpha)$.
   (b) For each word w in d (index by l):
      i. Get context word pairs b_w and generate the CWTD λ_w based on Eq. (3.1): $\lambda_w \sim Dir(\gamma)$.
      ii. Generate the CTD $\pi_{d,w}$ of word w according to θ_d and λ_w: $\pi_{d,w} \sim Dir(\lambda_w + \theta_d)$.
      iii. Choose a topic $Z_{d,l}$ assignment according to $\pi_{d,w}$: $Z_{d,l} \sim Cat(\pi_{d,w})$.
      iv. Generate $w_{d,l}$ based on the topic $Z_{d,l}$ and ϕ_k: $w_{d,l} \sim Cat(\phi_{Z_{d,l}})$.

The topic distribution and the context words are combined to further reduce the uncertainty of the topic assignment. As we explain in Section 3.4, this reduction ensures a high probability that consecutive words are assigned to the same topic.

Algorithm 2: Gibbs sampling algorithm

Input: A set $\boldsymbol{D}$ of documents with length N_d ($d \in \boldsymbol{D}$); number of iterations N_{iter}; number of topics K

Output: For each document $d \in \boldsymbol{D}$, topic distribution θ_d; for each topic k, word distribution ϕ_k ($1 \leq k \leq K$); word co-occurrence matrix $\boldsymbol{\Lambda}$

```
Initialize topic assignments randomly for all words in D
for iteration = 1 to N_iter do
    for d = 1 to |D| do
        for l = 1 to N_d do
            Generate a topic Z_{d,l} from P_{d,l} according to Eq. (3.2).
        Update θ_d, φ_k and Λ
return φ_k for each topic k, θ_d for each document d and Λ.
```

3.2.3 Parameter Estimation

We also use Gibbs sampling [29] to estimate parameters. In our sampling procedure, we need to calculate the conditional probability of topic assignment $P_{d,l,k} = P(Z_{d,l} = k|W_{d,l}, \boldsymbol{Z}_{d,-(d,l)}, \boldsymbol{W}'_{d,l}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$ for each word, where $W_{d,l}$ represents the lth word in d. $\boldsymbol{Z}_{d,-(d,l)}$ refers to the topic assignments for all words in d except for word $W_{d,l}$. $\boldsymbol{W}'_{d,l}$ are the context words of $W_{d,l}$. The result of $P_{d,l,k}$ is computed as follows (See Appendix A in Supplementary for detailed derivation):

$$P_{d,l,k} \propto \left[(n^{\boldsymbol{b}_w}_{k,-(d,i)} + \gamma_t) + (n_{d,k,-(d,l)} + \alpha_k) \right] \frac{n^t_{k,-(d,l)} + \beta_t}{\sum_{f=1}^{V} (n^f_{k,-(d,l)} + \beta_f)}, \tag{3.2}$$

where $n_{d,k,-(d,l)}$ is the number of words in d which belongs to topic k without $W_{d,l}$, $n^t_{k,-(d,l)}$ represents the number of word t of topic k without $W_{d,l}$. Compared with the

conditional probability of traditional topic models, such as LDA (as Eq. (3.3)), we see the difference is the probability of topic k for each word, which is affected by the frequency of its context word pairs on topic k in other documents.

$$P'_{d,l,k} \propto (n_{d,k,-(d,l)} + \alpha_k) \frac{n^t_{k,-(d,l)} + \beta_t}{\sum_{f=1}^{V} (n^f_{k,-(d,l)} + \beta_f)}. \tag{3.3}$$

According to Eq. (3.2), we obtain the conditional probabilities of topic assignment $P_{d,l,k}$ of each word in document d, so as to compute their corresponding topic distribution $\boldsymbol{P}_{d,l}$. Our sampling algorithm is shown in Algorithm 2. The word co-occurrence matrix $\boldsymbol{\Lambda}$ recording the number of word pairs in each topic is utilized to compute $\boldsymbol{\lambda}$, where the first two dimensions of $\boldsymbol{\Lambda}$ are all the unique words and the third dimension records the accumulated shared topic counts.

3.2.4 Topic Coherency Ratio

To further study how C-LDA affects the coherency and saliency in modeling, we calculate the joint probability of consecutive words which share the same topic in two cases: with and without considering context word pairs. For consecutive words $\boldsymbol{W}_{d,i:j}$ from W_i to W_j in document d, we denote the joint probability of sharing topic k by $P(\boldsymbol{W}_{d,i:j}, k)$ in the former case and the one in the latter case by $P'(\boldsymbol{W}_{d,i:j}, k)$. Taking their logarithms and computing their ratios as well as removing constant terms, we obtain the result as shown in Eq. (3.4). We retain the fraction of the right-hand side and name it Topic Coherency Ratio (TCR) as Eq. (3.5), denoted by R_t (See Appendix

B in Supplementary for detailed derivation).

$$\frac{\log P(\boldsymbol{W}_{d,i:j},k)}{\log P'(\boldsymbol{W}_{d,i:j},k)} \propto 1+\frac{\sum_{w\in\boldsymbol{W}_{d,i:j}}\log n_{k,-\boldsymbol{W}_{d,i:j}}^{\boldsymbol{b}_w}}{\sum_{w\in\boldsymbol{W}_{d,i:j}}\log n_{k,-\boldsymbol{W}_{d,i:j}}^{w}} \tag{3.4}$$

$$R_t(\boldsymbol{W}_{d,i:j},k) \triangleq \frac{\sum_{w\in\boldsymbol{W}_{d,i:j}}\log n_{k,-\boldsymbol{W}_{d,i:j}}^{\boldsymbol{b}_w}}{\sum_{w\in\boldsymbol{W}_{d,i:j}}\log n_{k,-\boldsymbol{W}_{d,i:j}}^{w}}. \tag{3.5}$$

For a set of consecutive words, the TCR is a ratio of occurrence number in the same topic between the context word pairs and words. The ratio ranges from $[0,1]$ and reflects the intensity of coherency for a set of consecutive words[1]. A higher ratio corresponds to a stronger coherency. Based on Eq. (3.4), we see $P(\boldsymbol{W}_{d,i:j},k)$ is always greater than $P'(\boldsymbol{W}_{d,i:j},k)$, which proves that C-LDA is more likely to generate coherent topic segments than other conventional topic models, including LDA and most of its extended versions[2]. For short segments consisting of a tiny proportion of words in a document, they can still be assigned to topic k with a higher probability than others if they contain frequent word pairs in topic k. Thus C-LDA ensures both better coherency and saliency in topic segmentation.

Since the number K of topics is a given empirical value, it is inevitable to generate redundant topic segments in each document. Although we might be able to specify a good K value beforehand, the difference in the number of topics contained in each document also leads to the inevitability of generating redundant segments. Therefore, merging redundant segments with frequent ones is indispensable, where the key is to judge whether the resulting segment has a higher coherency than the original ones. TCR is a coherency measurement based on the ratio of word pairs and words instead

[1]For the words in $\boldsymbol{W}_{d,i:j}$ belonging to topic k, if and only if they all occur as context word pairs of topic k in all the documents, TCR gets the maximum value 1, while it gets the minimum value 0 if and only if none of them occurs in a context.

[2]The fraction on the right-hand side is always positive.

of relying solely on their frequencies. This property ensures the coherency of segments are independent of their lengths; thus, we design a TCR based Redundant Topic Merging (RTM) algorithm to optimize the generated topic segments. The steps of RTM are: for each topic segment, we consider three cases: (1) merging with the previous segment; (2) merging with the next segment; (3) non-merging. For these three cases, TCRs are calculated separately and the case with the highest ratio is selected. We repeat the above steps until the number of segments stays unchanged.

3.3 Experiments

We evaluate our model by a series of experiments. The results were obtained with eight-fold cross-validation on a machine with Intel i9 processor and 128GB memory. The hyper-parameters (α, β, γ) were all fixed to 0.05.

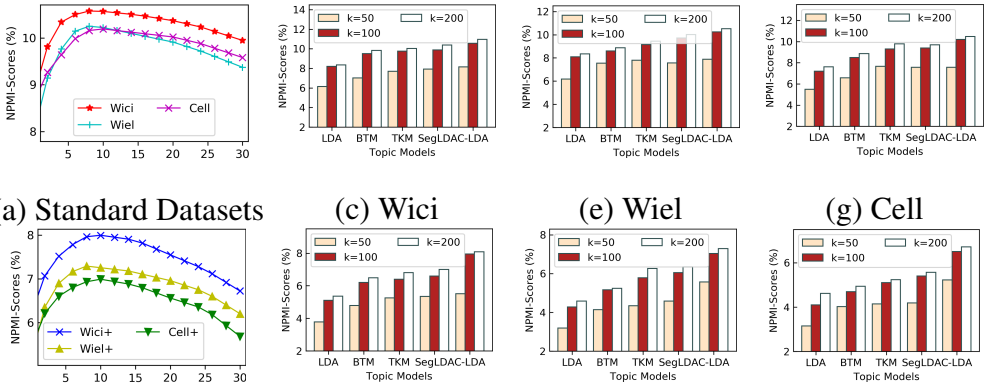
We tested our model on three standard datasets[1] (**Wikicities** (Wici), **Cellphones Reviews** (Cell) and **Wikielements** (Wiel)) and three extended datasets based on the former three. **Wikicities** contains Wikipedia articles about the world 100 largest cities by population, **Cellphones Reviews** contains 100 cellphone reviews and **Wikielements** contains 118 English Wikipedia articles about chemical elements. Labeled topic segments of the 3 standard datasets are all of the similar lengths (about 3000 words per document) and uniformly distributed; thus, to simulate the cases of more diverse topic structures, we increase their original total number of documents to 2000 and generated various sizes of topic segments for each document. The detailed generating steps for a document are: (1) select the number of segments based on a uniform distribution from 10 to 50; (2) for each segment, set its length from a uniform distribution of 10 to 100 and randomly assign it to a topic from the topic labels; (3) choose sentences of the

[1] `http://groups.csail.mit.edu/rbg/code/mallows/`

corresponding assigned topics from the labeled documents to fill the segments until all segments are loaded.

We compare C-LDA (available on Github[1]) against four topic models: LDA [7], BTM [16], TKM [83] and SegLDA [2]. BTM is a topic model based on word co-occurrence modeling. TKM is a method to generate coherent topics by considering the influence of keywords on their contexts. SegLDA is a LDA-extended model for topic segmentation by introducing an independent topic distribution for each predefined segment.

We use Normalized Point-wise Mutual Information (NPMI)[2] to measure the topic coherence scores [73]. It assumes that a topic is more coherent if the most probable words in the topic co-occur more frequently in the corpus [50]. NPMI scores are in $[-1,1]$ and a higher value indicates that the topic distributions are semantically more coherent. The performance of topic segmentation is evaluated with two metrics: PK [3] and Window Diff (WD) [4]. They both refer to error rates which are calculated by comparing the inferred segmentation with the gold-standard (ground truth) for each window based on moving a sliding window over the document. Lower scores refer to better segmentation performance.



(a) Standard Datasets (c) Wici (e) Wiel (g) Cell

(b) Extended Datasets (d) Wici+ (f) Wiel+ (h) Cell+

Figure 3.3: NPMIs of different L values (a-b) and different topic numbers k (c-h).

3.3.1 Topic Coherence

Firstly, we calculated NPMIs of C-LDA under different window sizes L (from 1 to 30) with topic number $K = 100$. The results are shown in Figure 3.3 (a-b). We see that, in both standard and extended datasets, NPMIs increase sharply until around $L = 10$ then begin to decline. Moreover, we see there is a sharp decline in the extended datasets when $L > 15$. This might be because of their more complex topic structures and longer window sizes are more likely to contain irrelevant content. Therefore, we set $L = 10$ in the rest of our experiments.

The results of NPMIs (with $K = 50, 100, 200$) for all baseline models are shown in Figure 3.3 (c-h). We see that C-LDA shows the best results on all six datasets and

[1] https://github.com/liliverpool/C-LDA.git

[2] NPMI$(k) = \sum_{1 \le i < j \le T} \frac{1}{-\log P(w_i, w_j)} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$, where $P(w_i, w_j)$ and $P(w_i)$ are the occurrence probabilities of word pair (w_i, w_j) and word w_i, respectively.

[3] $P_k(\text{ref}, \text{hyp}) = P(\text{false}|\text{refer}, \text{hyp}, \text{same}, k)P(\text{same}|\text{refer}, k) + P(\text{miss}|\text{refer}, \text{hyp}, \text{diff}, k)P(\text{diff}|\text{refer}, k)$, where "refer" is the ground truth and "hyp" is the generated segments. k is usually the half of the average gold-standard segment size (k=15 in our experiments). More details are in [4].

[4] $WD(\text{ref}, \text{hyp}) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(\text{ref}_i, \text{ref}_{i+k}) - b(\text{hyp}_i, \text{hyp}_{i+k})| > 0)$, where $b(i, j)$ represents the number of boundaries between positions i and j in the text and N is the number of sentences in the document [73].

Table 3.1: Topic segmentation results. PK and WD scores are in %. Bold fonts indicate best scores yielded by the models except for C-LDA-R and * indicates the best scores among all the models.

K	Models	PK						WindowDiff						Time Cost (hours)					
		Wici.	Wiel.	Cell.	Wici$^+$	Wiel$^+$	Cell$^+$	Wici.	Wiel.	Cell.	Wici$^+$	Wiel$^+$	Cell$^+$	Wici.	Wiel.	Cell.	Wici$^+$	Wiel$^+$	Cell$^+$
50	BTM	35.9	33.6	41.2	42.2	38.7	47.1	38.2	34.5	41.0	45.6	42.1	49.8	9.2	7.2	7.7	4.9	4.2	4.2
	TKM	28.6	23.9	37.8	33.8	28.5	43.2	28.7	33.4	38.7	35.8	31.8	46.4	**1.1***	**0.7***	**0.8***	**0.4***	**0.3***	**0.3***
	SegLDA	26.1	22.7	**35.2**	30.5	27.2	38.9	**27.1***	**25.6**	35.8	**33.4***	28.3	39.3	4.5	3.1	3.3	1.7	1.4	1.5
	C-LDA	**25.3**	**22.2**	35.3	**29.9***	**26.3**	**37.6**	27.7	25.7	**34.1***	33.7	**28.2**	**38.1***	1.9	1.2	1.5	0.9	0.8	0.8
	C-LDA-R	24.8*	20.6*	34.9*	30.3	26.2*	37.5*	27.3	24.8*	33.5	33.8	27.9*	38.5	2.0	1.3	1.6	1.0	0.9	0.9
100	BTM	32.5	30.2	37.5	40.1	36.5	44.3	35.7	33.4	40.2	41.4	39.8	45.7	15.7	12.6	11.5	8.6	8.2	8.5
	TKM	26.7	21.2	30.6	31.3	27.4	37.2	29.9	24.6	36.6	32.8	29.8	41.7	**2.1***	**1.5***	**1.7***	**0.9***	**0.7***	**0.8***
	SegLDA	23.2	20.4	31.3	27.4	24.1	34.8	28.5	23.9	33.5	29.8	24.6	36.3	8.8	6.5	7.6	3.1	2.4	2.6
	C-LDA	**22.1**	**19.7**	**29.8**	**25.8**	**23.2**	**31.7**	**27.5**	**22.6**	**32.2**	**27.4**	**24.5**	**33.9**	4.2	3.2	3.8	2.4	2.3	2.4
	C-LDA-R	21.9*	19.2*	27.6*	24.5*	22.6*	30.4*	25.2*	21.6*	30.7*	26.8*	23.7*	32.4*	4.3	3.3	3.9	2.5	2.4	2.5

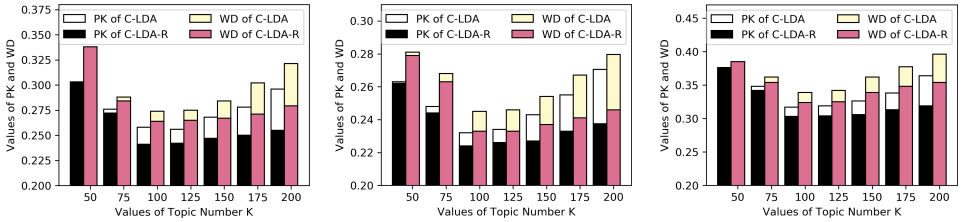
more significant improvements in datasets with more complex topic structures (Wici+, Wile+, Cell+), which proves the validity of our model for generating coherent topics. A possible reason is that C-LDA combines the frequency of context word pairs for each topic in modeling, while the other models (such as TKM) either consider only the word frequency in each topic or the frequency of all word pairs in individual documents (such as BTM). Moreover, semantic expressions in a document are usually coherent and segmented, e.g., paragraphs and sections, and thus, considering the context in a topic assignment can clarify the semantics of the word, so as to reduce the risk of splitting a coherent semantic segment.

3.3.2 Topic Segmentation

The results in topics of $K = 50$ and $K = 100$ are shown in Table 3.1, where the C-LDA-R is the C-LDA with RTM optimization algorithm. We see that C-LDA and C-LDA-R

perform the best in all cases of $K = 100$ and dominate in most cases when $K = 50$, which validates their performance for coherency and saliency of different segments in topic segmentation tasks.

BTM aims to generate all the distinct word pairs within a fixed window given a topic. Therefore, its effect on the topic coherency is achieved by increasing the joint probability of each word pair and topics. The high frequent word pairs in a corpus are of high joint probabilities. However, in a corpus, the majority are ordinary words but not word pairs, and their topic assignments are still of high uncertainty. Besides, the computation of all distinct word pairs significantly increases its training time. TKM improves the coherency of topic segmentation by considering the influence of keywords on the topic assignment of surrounding words and costs the least time. However, short topic segments with insufficient keywords are likely to be absorbed by long topic segments, which is a possible reason of its low performance. In some cases of insufficient topic number ($K = 50$), SegLDA outperforms other methods. However, as K increased from 50 to 100, its performance growth is inferior to C-LDA. For SegLDA, the topics for words in a segment can be assigned from the segment-specific topic distribution, which improves the saliency of topic segments. However, assigning topics without considering the original document can lead to a loss of context information and degrade the accuracy of topic modeling. That is, a word is possibly assigned to an incorrect topic even if it is not absorbed by others. C-LDA considers both the contextual word pairs and topic distribution. Based on the reorganized topic distribution CTD, it reduces the uncertainty of the topic assignment and increases the joint probability of consecutive words sharing the same topic at the expense of increasing the time consumption. Moreover, comparing the results of the original and their extended datasets, we see our method has stronger robustness to more complex topic structures, which also leads to better applicability.



(a) Wici$^+$ (b) Wiel$^+$ (b) Cell$^+$

Figure 3.4: PK and WindowDiff scores in terms of the number K of topics in (a) Wikicities$^+$, (b) Wikielements$^+$ and (c) CellphoneReviews$^+$. The stacked part above each bar is the improvement from RTM algorithm.

For C-LDA-R, we see that the effect of the RTM algorithm is limited in the case when $K = 50$ since it is insufficient to cover all the occurred topics. When $K = 100$, RTM effectively improves the performance of topic segmentation. To further study the effect of RTM, we calculated the changes of PK and WindowDiff with different numbers K of topics (from 25 to 200). The experiments were conducted on the 3 extended datasets and the results are shown in Figure 3.4. We see the measures of both C-LDA and C-LDA-R decrease quickly with the increase in length of K until $K = 100$. For C-LDA, the performance starts to decrease around $K = 150$, while for C-LDA-R, it tends to saturate as K keeps on increasing. The improvement by RTM becomes increasingly remarkable with the increase of K, which also proves the robustness of C-LDA-R for redundant topics. In addition, the time complexity of RTM for each document is $O(L\sum_{S' \in S} |S'|)$, where L is the context window size, S is the list of segments for a document and $|S'|$ represents the length of each segment S' in S. The time consumption of the RTM is acceptable, since L is set less than 30. Besides, the optimization process of each document is independent, which is easy for parallelization.

3.4 Summary

We proposed a new generative model for topic segmentation. By combining topic distribution and context word pairs-topic distribution, our model improves the certainty of the topic assignment and ensures high coherency and saliency in topic segmentation. Besides, we designed an optimization algorithm to merge redundant topic segments for each document. Our experiments show that our proposal outperforms baseline models, in terms of the segmentation scores of PK and WD in topic segmentation.

Chapter 4

Adaptive and Hybrid Context-Aware Fine-Grained Word Sense Disambiguation in Topic Modeling Based Document Representation

4.1 Overview

Document representation has long been studied in various areas [51] and widely used in real-life applications, e.g., public sentiment analysis [70], chat-bot [17] and e-learning [12]. Topic modeling and word embedding are two important paradigms for this task. The former takes a global view of the word distributions across the corpus to assign a topic to each word occurrence. The latter is based on a view of the local word collocation patterns observed in a text corpus. For the traditional versions of the two paradigms, such as LDA [7] and Word2vec [63], despite their significant progress

in various tasks and applications, there is a common issue that one word corresponds to one topic distribution or embedded vector, while in many cases, the semantics of a word may vary from different senses.

In recent years, lots of studies have been proposed for word sense disambiguation in the document representation task [9, 18, 23, 35]. Conventionally, they mainly rely on data enrichment, e.g., using knowledge libraries or pre-training datasets, for word sense induction, such as the WordNet [64] based Seeded-LDA [9] and SemLDA [23], the Wikipedia based Token-SDM [53] and LTTM [90]. All of them have achieved significant progress in word disambiguation performance in document representation. Moreover, the recent neural network based language models such as ELMO [72], GPT [76] and BERT [18], have rapidly improved the state-of-the-art on many NLP tasks. Despite their empirical success, the requirements for scales of pre-training datasets and computational efficiency are widely recognized issues due to their large numbers of parameters (94M for ELMO [72], 340M for BERT [18], and 1542M for GPT [76]) [92]. More importantly, for most of the data enrichment based methods, they assume that word senses are within the scope of the auxiliary text data, while senses in the auxiliary data may not constantly match the ones in a specific dataset. In contrast, we aim to discover more particular word semantic differences for a dataset related to a specific domain, in which we cannot always obtain sufficient scales of domain-specific data for the enrichment.

One solution to solve the WSD problem without data enrichment is context clustering [36, 66, 77]. DPMM [77] and EHModel [36] both obtain multi-prototype word embeddings by conducting clustering on all context word features for each word. Though useful, they generate multi-prototype word vectors in isolation, ignoring complicated correlations between word senses and their contexts [58]. MSSG [66] improves them by providing flexibility for the number of context clusters, allowing that the cluster

number varies according to the different distances of contexts in which each word occurs to their nearest sense cluster. Clustering contexts for each word can effectively divide their senses; however, it is challenging to clarify the differences between synonyms due to their similar contexts. Moreover, because of the independence between different clusters, the degree of the relationship between a sense and a specific context is ignored.

Another solution is to introduce an additional module to support the word sense disambiguation in document representation. SA-SLDA [94] integrated a Word Sense Induction (WSI) model and a topic model, and the two modules are linked by the topics corresponding to each word. CRFTM [28] captures the senses of a target word based on document-specific semantic correlated words, whose distances to the contextual words of the target word are lower than a given threshold. CGTM [104] and LF-LDA [68] are topic models using word embedding as an additional component. Topic2Vec [69] and TWE [58] introduce topic vectors in the embedding process, in which the word vector is embedded by concatenating the corresponding word and topic. NCLM [15] and TCNLM [98] incorporate the topic proportion of the text segment in which each word occurs, e.g., the document topic distribution, in the embedding framework, to distinguish different word senses. In more recent years, studies focus more on joint learning word embedding and topic modeling [26, 87, 110]. STE [87] holds the same basic idea as TWE [58] which combines both the latent topics and word embeddings, but the difference is to learn topical word embeddings in a unified manner. MMSG [26] assumes a word topic is drawn for word embeddings. Each context and the word in the context is drawn from the logbilinear model based on the topic embeddings. JTW [110] also assumes that each word embodies a finite set of senses which can be interpreted as topics, and each word sense representation can be transformed into a probabilistic distribution over topics.

Essentially, they all attempt to link each word occurrence to a specific sense through topics or topic proportions. Compared with the methods using topics, the methods based on topic proportions have better generalization. They provide more choices for a word sense in modeling than the methods limited to a given number of topics. Nevertheless, they both imply an assumption that there is a one-to-one correspondence between the word senses and the topics or the topic proportions.This assumption is somewhat an excessively strict constraint since a sense which corresponds to multiple topics or topic proportions is common, especially in the high topic dimension cases. One topic or similar topic proportions may also correspond to different senses of a word. This explicit and compulsory division for word semantics also inevitably overlooks the influence of other senses on further clarifying the differences to other words. On the other hand, many studies have shown that the two paradigms of word embedding and topic modeling are complementary in how they represent the semantics of documents; thus, improving either of them can contribute to optimizing the performance of their integrated model in document representation [26, 58, 68, 69, 87, 98, 104, 110].

In this chapter, we propose a hybrid context based word sense aware topic model (named HCT), where each sense of a word is estimated by integrating their topic distributions of both the context words in which it occurs and those of its other occurrences. Besides, we introduce the "Bag-of-Senses" (BoS) assumption that a document is a multiset of word senses, based on which HCT generates a word sense instead of the words themselves. The proposed model enjoys three substantial merits over the state-of-the-art methods: (1) no data enrichment or auxiliary module is needed, (2) it is an end-to-end model in which the topic vectors for hybrid contexts as well as their weights for each word are all considered as variables and learned jointly, and (3) the context window length of each word is learned adaptively.

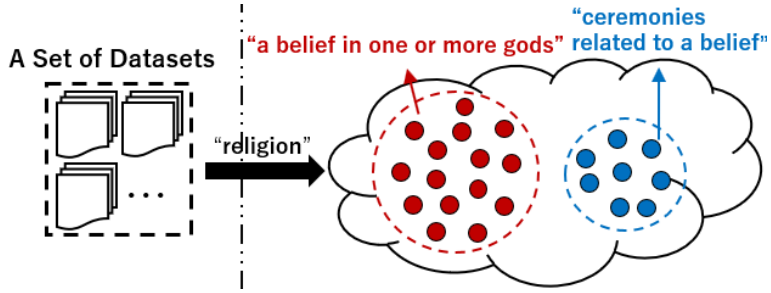
4.2 Sense Aware Topic Modeling

This section describes in detail our "Bag-of-Senses" (BoS) model. We explain how a word sense and its context window length is generated in our BoS based topic model using the hybrid context, and then how the parameters are estimated based on the Gibbs Sampling [29].

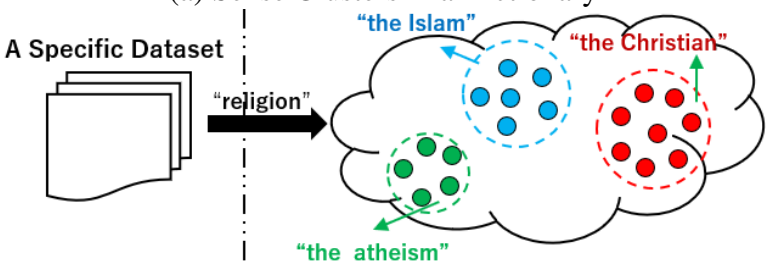
4.2.1 Bag-of-Senses

As a topic model, the basic task is, for a set of n documents $\boldsymbol{D} = \{D_0, D_1, ..., D_n\}$, to obtain the topic distribution $\boldsymbol{\theta}_{D_i}$ for each document D_i and word distribution ϕ_k for each topic k. Following most traditional topic models [7], $\boldsymbol{\theta}_{D_i}$ and ϕ_k are both assumed to follow Dirichlet distributions. The number K of topics for each document is assumed fixed and known. Each word corresponds to a K-dimensional topic vector. For the "Bag-of-Words" (BoW)[1] based models, all the word occurrences are mapped to one topic vector, whereas the vector cannot reflect the difference between various senses. Therefore, we propose the "Bag-of-Senses" (BoS) hypothesis: a document d in a dataset, is represented as a multiset of word senses s_w, $d = \{s_w\text{: } n_{s_w} | w \in \mathbb{W}_d\}$, where n_{s_w} is the counts of s_w in d, $\mathbb{W}_d$ refers to a sequence of word occurrences in d. Each word occurrence corresponds to a word sense and each sense corresponds to a topic vector. For example, suppose that a document D_i consists of three words "*religion*" where two of them refer to the sense of "*the Islam*" (s_1) and the other one refers to "*the Christian*" (s_2). In the BoW model, D_i is represented as $\{religion\text{: } 3\}$, whereas in BoS, it is $\{religion_{s_1}\text{: } 2, religion_{s_2}\text{: } 1\}$.

[1]The "Bag-of-Words" (BoW) is the most widely used simplifying model in document representation, which assumes a document is a multiset of words, disregarding the grammar and even the word order but keeping multiplicity [89].



(a) Sense Clusters in a Dictionary



(b) Dataset-Specific Sense Clusters

Figure 4.1: Comparison examples of the sense clusters in a dictionary (Figure 4.1 (a)) and the dataset-specific sense clusters (Figure 4.1 (b)).

4.2.2 Hybrid-Context based Sense Estimation

The primary problem is how to define the topic vector of the sense for a word occurrence. Based on the compilation rules of a dictionary that each group of similar usage corresponds to a sense cluster (Figure 4.1 (a)) [72], we can reasonably assume that the senses of each word in a specific dataset hold the similar clustered properties (Figure 4.1 (b)). Moreover, the usage differences of senses for each word are reflected by their corresponding different contexts. Therefore, we give the definitions of the Dataset-Specific Word Sense and the Context Words as follows.

Definition 1. Dataset-Specific Word Sense In BoS, the Dataset-Specific Word Sense for a word is defined as a cluster of similar usages of all its senses in a specific dataset.

Definition 2. Context Words For word w in a document, given window length h, the context words $\boldsymbol{w}_h$ of word w refers to a set of words within the window.

To ensure both differences between various senses for each polysemy as well as those between synonyms, based on Definitions 1 and 2, a sense vector $\boldsymbol{v}_{\boldsymbol{w}_h,w}$ for a word w in a specific context $\boldsymbol{w}_h$ is estimated by a hybrid of all its sense clusters, where $\boldsymbol{w}_h$ is the context words of w. Specifically, let each sense cluster s' of word w correspond to a specific vector $\boldsymbol{v}_w^{s'}$; thus, $\boldsymbol{v}_{\boldsymbol{w}_h,w}$ is obtained by a mixture as:

$$\boldsymbol{v}_{\boldsymbol{w}_h,w} = \sum_{s' \in \boldsymbol{S}} \mu_{\boldsymbol{w}_h,w}^{s'} \boldsymbol{v}_w^{s'}, \tag{4.1}$$

where $\boldsymbol{S}$ is a set of all sense clusters of word w and $\mu_{\boldsymbol{w}_h,w}^{s'}$ is its corresponding weight ($\sum_{s' \in \boldsymbol{S}} \mu_{\boldsymbol{w}_h,w}^{s'} = 1$). Now the problem is how to define $\boldsymbol{v}_w^{s'}$ in a topic space. Explicitly estimating all the sense clusters of each word is difficult, since the cluster number for each word is quite different; thus, estimating the word sense vector directly by Eq. (1) is difficult. To solve this problem, inspired by the "Distributional Hypothesis" [24] which states that words in similar contexts have similar meanings, we can assume that sense clusters can be reflected in different sets of contexts. Therefore, given a set of contexts of each sense cluster, the vector $\boldsymbol{v}_w^s$ for cluster s can be represented by the average of the vectors for the words in the set of its contexts:

$$\boldsymbol{v}_w^s = \sum_{w_h \in \boldsymbol{w}_{h,s}} \boldsymbol{v}_{w_h}^g, \tag{4.2}$$

where $\boldsymbol{w}_{h,s}$ refers to the set of words occurring in the context of all senses in cluster s and $\boldsymbol{v}_{w'}^g$ is the global topic vector of word w'.

Nevertheless, obtaining $\boldsymbol{w}_{h,s}$ is also difficult since we cannot know all the possible contexts of each sense cluster in a dataset beforehand. Therefore, we rewrite Eq. (1)

as:

$$\boldsymbol{v}_{\boldsymbol{w}_h,w} = \mu^{s}_{\boldsymbol{w}_h,w}\boldsymbol{v}^{s}_{w} + \sum_{s' \in \boldsymbol{S}_{-s}} \mu^{s'}_{\boldsymbol{w}_h,w}\boldsymbol{v}^{s'}_{w},$$

where $\boldsymbol{S}_{-s}$ is the set of all sense clusters except for s. We see that $\boldsymbol{v}_{\boldsymbol{w}_h,w}$ can be represented as a combination of one sense cluster vector and a weighted sum of other cluster vectors, while the latter can be approximately regarded as a general vector of w since it contains most of its senses. Hence, we can always find a combination of weights to let $\boldsymbol{v}_{\boldsymbol{w}_h,w}$ be represented as a weighted sum of a local sense vector $\boldsymbol{v}^{l}_{w}$ and a global topic vector $\boldsymbol{v}^{g}_{w}$, where $\boldsymbol{v}^{l}_{w}$ is only concerned about the current context $\boldsymbol{w}_h$ ($\boldsymbol{v}^{l}_{w} = \sum_{w' \in \boldsymbol{w}_h} \boldsymbol{v}^{g}_{w'}$). Hence, the sense vector $v_{\boldsymbol{w}_h,w}$ in a specific context can be calculated as:

$$\boldsymbol{v}_{\boldsymbol{w}_h,w} = \mu_{\boldsymbol{w}_h,w}\boldsymbol{v}^{l}_{w} + (1 - \mu_{\boldsymbol{w}_h,w})\boldsymbol{v}^{g}_{w}, \tag{4.3}$$

where $\mu_{\boldsymbol{w}_h,w}$ is the corresponding weight. Eq. (3) avoids obtaining all the sense clusters of each word beforehand since the calculation of $\boldsymbol{v}_w$ is independent of sense clusters.

We name the topic vector of a word Global Word Sense Vector (denoted by $\boldsymbol{v}^{g}$), the mean vector of its context words Local Word Sense Vector (denoted by $\boldsymbol{v}^{l}$), the topic vector of word sense Word Sense Vector (denoted by $\boldsymbol{v}_w$), and the weight of $\boldsymbol{v}^{l}$ Local Sense Weight (denoted by $\mu_{\boldsymbol{w}_h,w}$). Therefore, $\boldsymbol{v}_{\boldsymbol{w}_h,w}$ for a word w within context words $\boldsymbol{w}_h$ can be estimated by its $\boldsymbol{v}^{g}_{w}$ and $\boldsymbol{v}^{l}_{\boldsymbol{w}_h,w}$. Their definitions are as follows.

Definition 3. Global Word Sense Vector For a K dimensional topic space, the Global Word Sense Vector $\boldsymbol{v}^{g}_{w}$ is the probability distribution of w for the K topics.

Definition 4. Local Word Sense Vector For a word w in a context of $\boldsymbol{w}_h$, the Local

Word Sense Vector $\boldsymbol{v}^l_{\boldsymbol{w}_h,w}$ of w is a mean of $\boldsymbol{v}^g$s of its context words:

$$\boldsymbol{v}^l_{\boldsymbol{w}_h,w} = \sum_{w' \in \boldsymbol{w}_h} \boldsymbol{v}^g_{w'}.$$

Definition 5. Word Sense Vector For a word w with context words $\boldsymbol{w}_h$, its sense $\boldsymbol{v}_{\boldsymbol{w}_h,w}$ is a weighted average of its $\boldsymbol{v}^g$ and $\boldsymbol{v}^l$:

$$\boldsymbol{v}_{\boldsymbol{w}_h,w} = \mu_{\boldsymbol{w}_h,w}\boldsymbol{v}^l_{\boldsymbol{w}_h,w} + (1-\mu_{\boldsymbol{w}_h,w})\boldsymbol{v}^g_w,$$

where $\mu_{\boldsymbol{w}_h,w}$ is named Local Sense Weight.

4.2.3 Word Sense Generation

Based on the above definitions, for a BoS based topic model, a document is generated by word senses, while a specific sense consists of a word and its context words. Therefore, given topic to the ith word of document d, not only a word is generated but also its context words.

According to Definition 4, using joint probabilities to estimate the generation of context words is inappropriate because the Local Word Sense Vector of a word is defined as the mean topic vector of the context words. Therefore, we assume the set of context words $\boldsymbol{w}_h$ of word w to be a pseudo word $c_{\boldsymbol{w}_h}$ as an observed variable, and takes the average of topic vectors for all the involved words as its own vector. Following LDA [7], the topic-word distribution ϕ_k for w and the topic-pseudo word distribution $\pi_{h,k}$ for $c_{\boldsymbol{w}_h}$ follow their respective Dirichlet distributions as:

$$\phi_k \sim Dir(\boldsymbol{\beta}), \pi_{h,k} \sim Dir(\boldsymbol{\gamma}).$$

Therefore, given a topic k, the word w and its corresponding pseudo word $c_{\boldsymbol{w}_h}$ follow their respective Categorical distributions:

$$w \sim Cat(\phi_k), c_{\boldsymbol{w}_h} \sim Cat(\pi_{h,k}).$$

According to the conjugate of Dirichlet distribution and Categorical (or Multinomial) distribution [42], their expectations are calculated as follows:

$$E_{\beta}(\phi_{k,w}) = \frac{n^w_{k,-(d,i)} + \beta_w}{\sum_{f=1}^{V}(n^f_{k,-(d,i)} + \beta_f)} \tag{4.4}$$

$$E_{\gamma}(\pi_{h,k,c_{\boldsymbol{w}_h}}) = \frac{\sum_{t \in \boldsymbol{w}_h}(n^t_{k,-(d,i)} + \gamma_t)}{L\sum_{f=1}^{V}(n^f_{k,-(d,i)} + \gamma_f)}, \tag{4.5}$$

where $\phi_{k,w}$ and $\pi_{h,k,c_{\boldsymbol{w}_h}}$ refer to the probabilities of generating word w and $c_{\boldsymbol{w}_h}$ given topic k, respectively. h is the size of context window. $n^t_{k,-(d,i)}$ is the number of word t belonging to topic k without the ith word in document d. Based on Definition 5, we introduce a hidden variable $s_{\boldsymbol{w}_h,w}$, named word sense, to present the sense of word w in context $\boldsymbol{w}_h$, where $s_{\boldsymbol{w}_h,w}$ is generated from:

$$s_{\boldsymbol{w}_h,w} \sim (1 - \mu_{\boldsymbol{w}_h,w})\phi_{k,w} + \mu_{\boldsymbol{w}_h,w}\pi_{h,k,\boldsymbol{w}_h}.$$

For weight $\mu_{\boldsymbol{w}_h,w}$, based on Definition 5, $s_{\boldsymbol{w}_h,w}$ can be regarded as the probability for $v^l_{\boldsymbol{w}_h,w}$ in a mixture of topic distributions of $\boldsymbol{v}^g_w$ and $\boldsymbol{v}^l_{\boldsymbol{w}_h,w}$. Therefore, given topic k, $\mu_{\boldsymbol{w}_h,w}$ can be estimated by the Conditional Probability Formula [39]. Specifically, for the ith word w in document d, we obtain $\mu_{\boldsymbol{w}_h,w}$ by the probabilities of topic k in $\boldsymbol{v}^g_w$ and

$\boldsymbol{v}^l_{\boldsymbol{w}_h,w}$, as Eq. (6):

$$\mu_{\boldsymbol{w}_h,w} \triangleq P(v^l_{\boldsymbol{w}_h,w}|k) = \frac{P(k|v^l_{\boldsymbol{w}_h,w})P(v^l_{\boldsymbol{w}_h,w})}{P(k)} = \frac{P(k,v^l_{\boldsymbol{w}_h,w})}{P(k,\boldsymbol{v}^g_w)+P(k,v^l_{\boldsymbol{w}_h,w})}, \tag{4.6}$$

where $P(k,\boldsymbol{v}^g_w)$ refers to the probability of topic k in $\boldsymbol{v}^g$ and $P(k,\boldsymbol{v}^l_{\boldsymbol{w}_h,w})$ refers to that in $\boldsymbol{v}^l_{\boldsymbol{w}_h,w}$. Their calculations are:

$$P(k,\boldsymbol{v}^g_w) = \frac{\phi_{k,w}}{\sum_{s=1}^K(\phi_{s,w})} \propto \frac{n^w_{k,-(d,i)}+\beta_w}{\sum_{s=1}^K(n^w_{s,-(d,i)}+\beta_w)},$$
$$P(k,v^l_{\boldsymbol{w}_h,w}) = \frac{\pi_{k,\boldsymbol{w}_h}}{\sum_{s=1}^K(\pi_{s,\boldsymbol{w}_h})} \propto \frac{1/L\sum_{t\in\boldsymbol{w}_h} n^t_{k,-(d,i)}+\gamma_t}{\sum_{s=1}^K\left[1/L(\sum_{t\in\boldsymbol{w}_h} n^t_{k,-(d,i)}+\gamma_t)\right]},$$

where $n^w_{k,-(d,i)}$ is the number of word w which belongs to topic k without the ith word in document d.

4.2.4 Adaptive Context Estimation

Since using a fixed context window length for different usages of each word is inappropriate in the WSD tasks, we introduce another new variable $\eta_{d,w}$ to adaptively control the window length for each word. In BoS, the context words of each word reflects one of its sense, therefore, the context words related to the sense should be consistent with the topic assignment of the corresponding word in topic modeling. Specifically, given a topic k of a word, the probability $P(h|k)$ of selecting a context window with size h can be estimated according to the probability of k in the average topic vector of context words under each possible size, since $P(h|k)$ is reduced when the window covers words which are not related to the word sense. Similar to $\mu_{\boldsymbol{w}_h,w}$, we obtain the

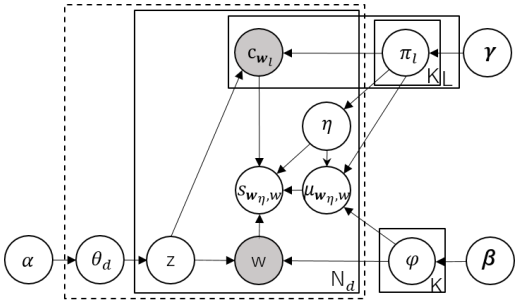


Figure 4.2: Plate notation of HCT.

probability $P(h|k)$ by the Conditional Probability Formula [39] as follows:

$$P(h|k) = \frac{P(k, v^l_{\boldsymbol{w}_h,w})}{\sum_{h=1}^{H} P(k, v^l_{\boldsymbol{w}_h,w})}, \tag{4.7}$$

and the optimized context window length η is generated by a Categorical distribution:

$$\eta \sim Cat(P(1|k), ..., P(L|k)), \tag{4.8}$$

where $h \in [1, H]$ and H is an empirical value which represents the possible maximum value of the context window length in a given dataset. We see that the higher the proportion of the context words related to the topic of the involved word, the greater the probability of generating the corresponding window length.

4.2.5 Model Description

The plate notation is shown in Figure 4.2. We introduce six new variables $\boldsymbol{\pi}$, $\boldsymbol{\gamma}$, $c_{\boldsymbol{w}_\eta}$, η, $s_{\boldsymbol{w}_\eta,w}$ and $\boldsymbol{\mu}_{\boldsymbol{w}_\eta,w}$ to traditional LDA, where $\boldsymbol{\pi}_{\boldsymbol{\eta}}$ represents the topic-pseudo word distribution with a parameter of $\boldsymbol{\gamma}$. $c_{\boldsymbol{w}_\eta}$ refers to a pseudo word for the average of context words. $s_{\boldsymbol{w}_\eta,w}$ represents the sense of word w in context $\boldsymbol{w}_\eta$. $\boldsymbol{\mu}_{\boldsymbol{w}_\eta,w}$ refers to

the Local Sense Weight. For the other variables, $\boldsymbol{\theta}_d$ represents the topic distribution of document d with parameter $\boldsymbol{\alpha}$. $\boldsymbol{\phi}_k$ is the topic-word distribution of topic k with parameter $\boldsymbol{\beta}$. w is a word in a document and z is its corresponding topic. For a dataset of D documents with a vocabulary of size V and latent topics indexed in $\{1,...,K\}$, the generative process of HCT is described as follows:

1. Generate $\boldsymbol{\phi}_k$ for each topic k: $\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta})$.

2. For each document d:

   (a) Generate $\boldsymbol{\theta}_d$ for document d: $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$.

   (b) For each word w in d (index by i):

      i. Assign topic $z_{(d,i)}$ by $\boldsymbol{\theta}_d$: $z_{(d,i)} \sim Cat(\boldsymbol{\theta}_d)$.

      ii. For each context window length $h \in [1,H]$:

         A. Obtain the context words $\boldsymbol{w_h}$.

         B. Generate the topic-pseudo word distribution $\boldsymbol{\pi}_{h,z_{(d,i)}}$:
         $\boldsymbol{\pi}_{h,z_{(d,i)}} \sim Dir(\boldsymbol{\gamma})$.

         C. Calculate $P(h|z_{(d,i)})$ of selecting the context window with size h based on Eq. (7).

      iii. Generate the optimized context window length η:
      $\eta \sim Cat(P(1|z_{(d,i)}),...,P(H|z_{(d,i)}))$.

      iv. Generate $c_{\boldsymbol{w}_\eta}$ by $\boldsymbol{\pi}_{\eta,z_{(d,i)}}$: $c_{\boldsymbol{w}_\eta} \sim Cat(\boldsymbol{\pi}_{\eta,z_{(d,i)}})$.

      v. Generate w by $\boldsymbol{\phi}_{z_{(d,i)}}$: $w \sim Cat(\boldsymbol{\phi}_{z_{(d,i)}})$.

      vi. Calculate $\mu_{\boldsymbol{w}_\eta,w}$ by Eq. (6).

      vii. Generate $s_{\boldsymbol{w}_\eta,w}$ by $z_{(d,i)}$, $\phi_{z_{(d,i)},w}$ and $\pi_{\eta,z_{(d,i)},c_{\boldsymbol{w}_\eta}}$:
      $s_{\boldsymbol{w}_\eta,w} \sim (1-\mu_{\boldsymbol{w}_\eta,w})\phi_{z_{(d,i)},w} + \mu_{\boldsymbol{w}_\eta,w}\pi_{h,z_{(d,i)},\eta,\boldsymbol{w}_\eta}$.

ϕ_k and $\pi_{\eta,k}$ share the same topic-word matrix $\boldsymbol{M}$ which records the number of occurrence for each word in different topics. Based on $\boldsymbol{M}$, ϕ_k and $\pi_{\eta,k}$ are calculated with reference to their respective Dirichlet parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Each row and column of $\boldsymbol{M}$ respectively corresponds to a topic-word distribution and a topic vector of a word.

4.2.6 Parameter Estimation

Except for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the parameters of our model are approximately estimated by Gibbs sampling [29]. In the estimation procedure, we need to calculate conditional distribution $P_{(d,i),k} = P(z_{(d,i)}=k|w_{d,i},\boldsymbol{z}_{d,-(d,i)},\boldsymbol{w}_{1,(d,i)},...,\boldsymbol{w}_{h,(d,i)},...,\boldsymbol{w}_{H,(d,i)},\boldsymbol{\mu}_{(d,i)},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})$, for each document d, where $w_{(d,i)}$ represents the ith word in d and $\boldsymbol{z}_{d,-(d,i)}$ refers to the topic assignments for all words in d except word $w_{(d,i)}$. $\boldsymbol{w}_{h,(d,i)}$ is the context words of $w_{(d,i)}$ with a window length of η ($\eta \in [1,H]$) and $\mu_{\eta,(d,i)}$ refers to the Local Sense Weight of $w_{(d,i)}$. $P_{(d,i),k}$ is computed as follows:

$$
\begin{aligned}
P_{(d,i),k} &\propto P(z_{(d,i)} = k, s_{\boldsymbol{w}_{\eta,(d,i)},w_(d,i)} = s_t|\boldsymbol{w}_{1,w_(d,i)},...,\boldsymbol{w}_{H,w_(d,i)},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) \\
&= P(z_{(d,i)} = k, w_{(d,i)} = t, c_{\boldsymbol{w}_{1,(d,i)}} = c_{\boldsymbol{t}_1},...,c_{\boldsymbol{w}_{H,(d,i)}} = c_{\boldsymbol{t}_H}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) \\
&= P(z_{(d,i)} = k|\boldsymbol{\alpha}) \int (1-\mu_{\eta,(d,i)})P(w_{(d,i)} = t|\boldsymbol{\beta}) \\
&\quad + \mu_{\eta,(d,i)}P(c_{\boldsymbol{w}_{\eta,(d,i)}} = c_{\boldsymbol{t}_\eta}|\boldsymbol{\gamma})P(\eta|k)d\eta \\
&= \int P(z_{(d,i)} = k|\boldsymbol{\theta}_d)P(\boldsymbol{\theta}_d|\boldsymbol{\alpha})d\boldsymbol{\theta}_d \\
&\quad \int P(\eta|k)\Bigg[(1-\mu_{\eta,(d,i)})\int P(w_{(d,i)} = t|\phi_k)P(\phi_k|\boldsymbol{\beta})d\phi_k \\
&\quad + \mu_{\eta,(d,i)}\int P(c_{\boldsymbol{w}_{\eta,(d,i)}} = c_{\boldsymbol{t}_\eta}|\boldsymbol{\pi}_{\eta,k})P(\boldsymbol{\pi}_{\eta,k}|\boldsymbol{\gamma})d\boldsymbol{\pi}_{\eta,k}\Bigg]d\eta
\end{aligned}
$$

$$= \int P(z_{(d,i)} = k|\boldsymbol{\theta}_d) P(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) d\boldsymbol{\theta}_d$$
$$\sum_{\eta \in [1,L]} P(\eta|k) \Bigg[(1 - \mu_{\eta,(d,i)}) \int P(w_{(d,i)} = t|\phi_k) P(\phi_k|\boldsymbol{\beta}) d\phi_k$$
$$+ \mu_{\eta,(d,i)} \int P(c_{\boldsymbol{w}_{\eta,(d,i)}} = c_{t_\eta}|\boldsymbol{\pi}_{\eta,k}) P(\boldsymbol{\pi}_{\eta,k}|\boldsymbol{\gamma}) d\boldsymbol{\pi}_{\eta,k} \Bigg]$$

Based on the definition of Dirichlet distribution, conditional distribution $P_{(d,i),k}$ can be simplified as

$$P_{(d,i),k} \propto E_\alpha(\theta_{d,k}) \sum_{\eta \in [1,H]} P(\eta|k) \Big[(1 - \mu_{\eta,(d,i)}) E_\beta(\phi_{k,t}) + \mu_{\eta,(d,i)} E_\gamma(\pi_{h,k,\boldsymbol{w}_{h,(d,i)}}) \Big], \quad (4.9)$$

where $E_{\boldsymbol{\alpha}}(\theta_{d,k})$ refers to the expectation of the probability for topic k in document d, which can be estimated by

$$E_{\boldsymbol{\alpha}}(\theta_{d,k}) \propto (n_{d,k,-(d,i)} + \alpha), \quad (4.10)$$

where $E_{\boldsymbol{\beta}}(\phi_{k,t})$ and $E_{\boldsymbol{\gamma}}(\pi_{h,k,\boldsymbol{w}_{h,(d,i)}})$ are the expectations of the probabilities for word w_t and the pseudo word of context words $\boldsymbol{w}_{h,(d,i)}$. They are calculated by Eqs. (4) and (5). $\mu_{\eta,(d,i)}$ and $P(\eta|k)$ are estimated by Eqs. (6), (7) and (8). Based on Eq. (9), we obtain topic assignment probability $P_{(d,i),k}$ for each word in d, so as to compute their corresponding topic distribution $\boldsymbol{P}_{(d,i)}$. Detailed steps are shown in Algorithm 1.

Algorithm 3: Parameter Estimation Algorithm

Input: A set of D documents of length N_d; number N_{iter} of iterations; number K of topics; Dirichlet parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$

Output: For each document d, topic distribution θ_d; for each topic k, word distribution ϕ_k ($1 \leq k \leq K$);

Initialize topic assignments randomly, context window length as 5 for each word, and $\mu_{\boldsymbol{w}_\eta, w} = 0.5$ for all words in documents D with context words of $\boldsymbol{w}_\eta$

for *iteration = 1 to* N_{iter} **do**

  for d *= 1 to* D **do**

    for i *= 1 to* N_d **do**

      Update the context window length η by Eq. (7) and Eq. (8).

      Update $\mu_{\boldsymbol{w}_{\eta,(d,i)}, w_{(d,i)}}$ by Eq. (6).

      Assign a topic $z_{(d,i)}$ from $\boldsymbol{P}_{(d,i)}$ by Eq. (9).

    Update θ_d and topic-word matrix $\boldsymbol{M}$.

return topic-word matrix $\boldsymbol{M}$, θ_d for each document d as well as $\mu_{\boldsymbol{w}_{\eta,(d,i)}, w_{(d,i)}}$ for each word.

4.3 Experiments

We conducted both quantitative and qualitative analyses. Firstly, we use three benchmark datasets 20Newsgroups[1] (20NG), Toxic Comments[2] (T-COM) and Sanders Tweet[3] (Tweet) in the quantitative analysis for evaluating the word sense estimation qualities, document classification effects, topic modeling accuracy, the adaptive context window length effectiveness, and model efficiency. In the qualitative analysis, we use 20NG, T-COM and PubMed[4] to verify the effects of our model in fine-grained word sense detection.

20NG is a collection of approximately 20,000 newsgroup documents, organized into 20 different newsgroups, each corresponding to a different topic. T-COM is a dataset of Wikipedia comments which human raters have labeled for toxic behavior, i.e., comments which are rude, disrespectful, or controversial. Tweet is a twitter sentiment corpus created by Sanders Analytics, which consists of 5513 hand-classified tweets. Each tweet was classified for one of four different topics. PubMed database contains more than 30 million journal citations and abstracts for biomedical literature from around the world since the 1970s. For all the datasets, stop words were removed in advance.

To validate the proposed model HCT, we test the following baseline methods: a traditional topic model LDA [7], four word embedding methods combined with topic modeling, JTW [110], STE [87], TWE-1 [58] and NCLM [15], where STE and JTW learn word embeddings and topics jointly, two topic models CGTM [104] and LF-LDA [68], which are combined with a Skip-gram based word embedding model, as well as two sense cluster based embedding methods EHModel [36] and MSSG [66].

[1] http://qwone.com/~jason/20Newsgroups/

[2] http://kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

[3] https://github.com/zfz/twitter_corpus

[4] https://www.nlm.nih.gov/databases/download/pubmed_medline.html

Moreover, in the quantitative analysis, we combine our model HCT with a skip-gram based word embedding framework [63] as another testing method (denoted by HCT-S). Its integration principle is similar to TWE-1 [58], where the difference is that we take the sense vector for each word occurrence rather than its topic assignment as additional input features.

4.3.1 Experimental Setup

We use Support Vector Machines (SVM) [89] to predict ground truth labels from the topic vectors of documents and used WEKA [31] for learning a classifier. These results are the average of ten-fold cross-validation. The detailed parameters for SVM are set as follows: (1) the type of SVM model is C-SVC; (2) the type of kernel function is RBF; (3) the degree in kernel function is 3; (4) the gamma in kernel function is $1/k$ (k is the number of different labels); (5) the parameter C of C-SVC is set by 1 (C determines the influence of the misclassification on the objective function).

In both of the qualitative and quantitative analyses, we use kMeans [59] to cluster the word sense vector and visualize them with t-SNE [96]. The cluster number are determined by Silhouette Coefficient[1]. We chose the cluster number (from 2 to 10) with the highest Silhouette Coefficient as the parameter for kMeans. Moreover, we set the maximum empirical value of context window length $H = 20$. The hyper-parameters (α, β and γ) were all fixed to 0.05.

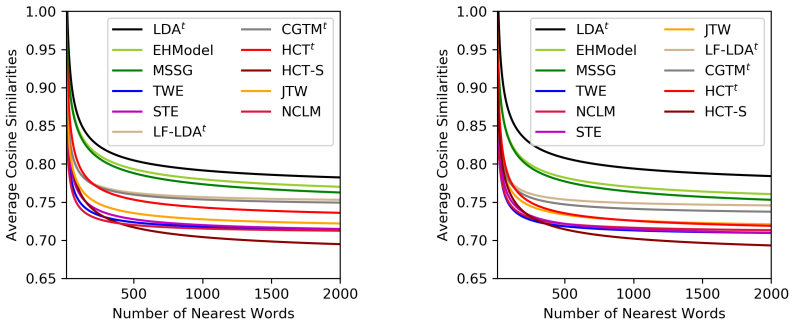
4.3.2 Quantitative Analysis

The quantitative experiments are conducted on 20NG, T-COM and Tweet, in terms of three aspects: classification effect, sense estimation quality and topic modeling accu-

[1]The Silhouette Coefficient ranges in $[-1, 1]$, where a higher value indicates that the object is better matched to its own cluster and poorly matched to other clusters [53].

Table 4.1: Comparison of the average similarities between the sense clusters of each word ($\overline{C_{sc}}$).

	K	EHModel	MSSG	TWE-1	NCLM	STE	JTW	LF-LDA^t	CGTM^t	HCT^t	HCT-S
$\overline{C_{sc}}$	100	**0.819**	0.822	0.849	0.841	0.832	0.857	0.887	0.876	0.847*	0.826
	200	0.835	**0.814**	0.837	0.833	0.821	0.815	0.867	0.864	0.835*	0.823



(a) K=100 (b) K=200

Figure 4.3: Comparison of average similarities between vectors of each word and its Top-n ($n \in [1, 2000]$) nearest words based on the cosine similarity on 20NG dataset.

racy.

4.3.2.1 Sense Estimation Quality

To investigate sense estimation qualities, we evaluate the differences among various senses for each word and those among its synonyms on 20NG. In this analysis, the differences are measured by the cosine similarity. A lower value reflects higher discrimination between different senses or words, corresponding to a better sense estimation quality. We used kMeans [96] to cluster the sense vectors, where the detailed parameter settings are as described in Section 4.1.

We calculated: (1) the average cosine similarities between sense clusters for all occurrences of each word (denoted by $\overline{C_{sc}}$, as shown in Table 4.1), as well as (2) the aver-

age cosine similarities between each word vector and its Top-n ($n \in [1, 2000]$) nearest word vectors (denoted by $\overline{C_w}$, as shown in Figure 4.3), where the word vector is calculated by the mean of all its sense vectors. * indicates the best scores yielded by topic models (labeled by ^{t}), and bold fonts indicate the best ones of all the baselines. We see that the lowest $\overline{C_{sc}}$ are achieved by the clustered based methods (EHModel and MSSG). In all cases, HCT-S is superior to most of the others, and HCT performs the best in the topic models. The word semantic divisions by contexts clustering can directly clarify the distinctions between different senses of a word. However, it may obscure the differences to other words, especially those with similar usages. There is a trade-off for the two goals in WSD task: (1) division for different senses of the same word, and (2) discrimination for different words with the similar or same sense. Ignoring the differentiation to other words while dividing the word semantics is likely to confuse their similar sense clusters, and thus reducing the differences between synonyms. Furthermore, for the word embedding based methods, their $\overline{C_w}$ are lower than the topic models as the number of nearby words involved increases. One possible reason is that the optimization targets of these two paradigms are different. The embedding models focus on optimizing word vectors, while topic models aim at optimizing the topic distributions of documents. This difference directs the embedding vectors more sufficiently to reflect the semantic similarities and differences between words. In the following experiments, we will discuss their complementary nature in the document representation task.

4.3.2.2 Document Classification

To evaluate the quality of document representation vectors, we conducted classification experiments on three benchmark datasets. We randomly sampled 12000 documents

Table 4.2: Comparison of the classification performance and NPMI on “Tweet”, “T-COM” and “20NG”.

Method	K	T-COM				20NG				TWEET			
		Precision	Recall	F-Score	NPMI	Precision	Recall	F-Score	NPMI	Precision	Recall	F-Score	NPMI
LDA^t		0.774	0.767	0.771	-12.2	0.721	0.695	0.707	-8.2	0.650	0.651	0.651	-15.3
CGTM^t		0.828	0.825	0.827	-8.4	0.841	0.835	0.838	-6.7	0.690*	0.686	0.688*	**-11.3***
LF-LDA^t		0.826	0.822	0.824	-9.6	0.839	0.831	0.835	-7.7	0.683	0.682	0.682	-12.8
EHModel		0.771	0.763	0.767	-	0.799	0.797	0.794	-	0.668	0.664	0.665	-
MSSG		0.783	0.779	0.781	-	0.814	0.812	0.813	-	0.677	0.675	0.675	-
TWE-1	100	0.819	0.824	0.821	-	0.848	0.847	0.847	-	0.682	0.681	0.682	-
STE		0.825	0.828	0.826	-9.1	0.851	0.857	0.854	-7.2	0.691	0.687	0.689	-11.7
NCLM		0.823	0.827	0.825	-	0.852	0.849	0.851	-	0.685	0.682	0.683	-
JTW		0.829	0.832	0.831	-8.6	0.858	0.861	0.860	-6.6	0.692	0.689	0.689	-12.3
HCT^t		0.838*	0.839*	0.838*	**-7.6***	0.862*	0.859*	0.861*	**-6.4***	0.689	0.687*	0.688	-11.5
HCT-S		**0.856**	**0.861**	**0.857**	-	**0.880**	**0.881**	**0.881**	-	**0.717**	**0.716**	**0.716**	-
LDA^t		0.806	0.788	0.797	-13.6	0.738	0.727	0.731	-9.4	0.651	0.653	0.653	-16.6
CGTM^t		0.835	0.832	0.832	-10.3	0.848	0.841	0.845	-7.8	0.697	0.695	0.695	-12.2
LF-LDA^t		0.831	0.830	0.830	-11.5	0.845	0.838	0.841	-8.3	0.687	0.686	0.686	-13.7
EHModel		0.808	0.801	0.804	-	0.811	0.807	0.808	-	0.671	0.675	0.672	-
MSSG		0.814	0.811	0.812	-	0.824	0.821	0.823	-	0.682	0.679	0.680	-
TWE-1	200	0.825	0.823	0.824	-	0.857	0.855	0.856	-	0.687	0.685	0.685	-
STE		0.831	0.834	0.832	-9.8	0.863	0.859	0.861	-7.7	0.702	0.697	0.697	-12.8
NCLM		0.828	0.832	0.831	-	0.859	0.857	0.858	-	0.686	0.687	0.686	-
JTW		0.837	0.836	0.836	-10.1	0.867	0.866	0.866	-8.5	0.693	0.691	0.692	-13.2
HCT^t		0.855*	0.857*	0.855*	**-9.6***	0.872*	0.875*	0.872*	**-7.1***	0.699*	0.703*	0.701*	**-11.9***
HCT-S		**0.866**	**0.871**	**0.869**	-	**0.885**	**0.887**	**0.885**	-	**0.727**	**0.731**	**0.728**	-

from 20NG, 10000 documents from T-COM, and all the four classes in Tweet. The precision and recall as well as the macro averaged F1-Score [61] (K=100, 200) are presented as the evaluation metrics for this task. The results are reported in Table 4.2, where * indicates the best scores achieved by the topic models, and bold fonts indicate the best scores achieved by all the models. Topic models are labeled by ^{t}. We see that HCT shows the best results in the topic modeling based methods in most cases, and the integrated method HCT combined with skip-gram is superior to all the other baseline models on the three datasets.

Classification performance reflects the ability to distinguish different classes of documents in their representation spaces. HCT considers the relationships between the different senses of each word in topic modeling, and thus achieves a better trade-off between the differentiation of various senses of each word and the semantic differences of synonyms. For short text datasets, the context words for each word may occupy the vast majority of the document; thus, their influence on the word sense may counteract the role of the document itself, such as the topic distribution. Therefore, our model has limited improvement in classification accuracy on short text datasets compared to other baselines. On the other hand, the integrated models, which combine both the topic modeling and word embedding, are more effective than the other baselines. The methods of jointly learning word embeddings and document topics perform better than the models integrating the two modules independently. However, they all assume that word senses under each topic dimension are different, or senses belong to one topic or topic proportion are the same. Nevertheless, it is common that a sense could belong to multiple topics, and the number of senses for each word is different. The senses of a word and its topics, or the word senses and the topic proportion of the text segment in which it occurs are not always in one-to-one correspondence. Therefore, this explicit and compulsory division for word senses is likely to decrease the accuracy of their

embedded vectors. Another significant issue is that all the baseline models neglect the degree of dependency for a word sense on its context words. However, these degrees of the dependencies of a sense varies from its usage frequency. For example, a non-standard use of a word is more dependent on its context than its standard use [47, 76]. These problems might be the leading causes of their performance bottlenecks.

Besides, the complementarity of topic modeling and word embedding improves the performance of the integrated methods. For most topic modeling based methods, embeddings are mainly used to improve the accuracy of the topic assignment for each word (CGTM and LF-LDA). This indirect influence on topic modeling cannot sufficiently reflect the context information captured by the embedding models. For the embedding based models (JTW, NCLM, TWE-1, STE, and HCT-S), the topic modeling results are inputted as additional features and directly utilized in the word vector estimation, which might be the main reason for the embedding-based integrated methods being generally better than other integrated ones in this analysis.

4.3.2.3 Topic Modeling Accuracy

As a topic modeling method, we evaluate the accuracy of the discovered topics by calculating the average normalized pointwise mutual information (NPMI) for each method. NPMI is a popular metric of the topic modeling quality by measuring the coherence of a topic based on point-wise mutual information [80]. It assumes that a topic is more coherent if the most probable words in the topic co-occur more frequently [65]. Besides, topic coherence can also reflect the matching between the topic assignment and semantics for each word, since semantic expressions in a document are usually coherent and segmented (such as paragraphs and sections) [75]. A higher NPMI score indicates that the topic distributions are semantically more coherent. Given the T most

probable words of topic k, NPMI is:

$$\text{NPMI}(k) = \sum_{1 \le i < j \le T} \frac{1}{-\log P(w_i, w_j)} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)},$$

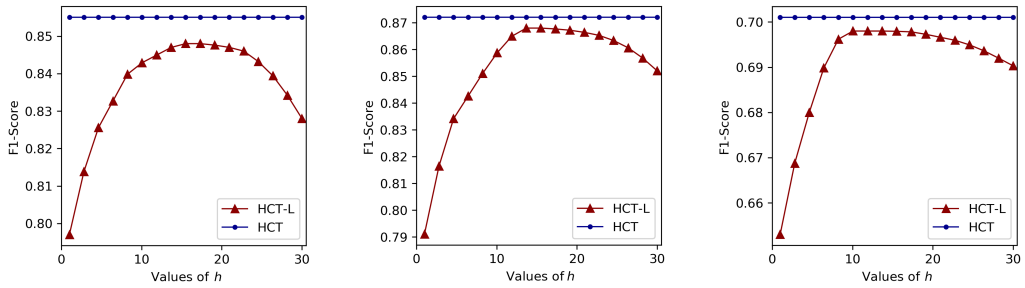
where $P(w_i, w_j)$ and $P(w_i)$ are the probabilities of word pair (w_i, w_j) and word w_i, respectively, and are both estimated from an external dataset[1].

The results of topic models (HCT, LDA, LF-LDA, and CGTM) and joint learning based integrated methods (STE and JTW) are shown in Table 4.2, where bold fonts highlight the best results. We see HCT shows the best results in most cases, which confirms that our model can generate more coherent topics than baselines. HCT generates both words and context words as well as uses the context and adaptive weights to clarify word semantics, reducing the uncertainty of the word topic assignment. The others use embedding vectors to clarify the word topic assignment [68]. They give semantically related words a better chance to share the same topic label, considering semantic similarities of word embeddings. However, as mentioned before, it is not always appropriate to give the same topic to the semantically similar words in the embedding based vector space. Moreover, the word embedding vectors are learned by all their contexts, which is also difficult to help specify a rare sense for a word in a specific context.

4.3.2.4 Effectiveness of the Adaptive Context Window Length

For investigating the effectiveness of the Adaptive Context Window Length, we tested our model in terms of their classification performance in 20NG, T-COM, and Tweet using a uniform context window length of $h \in [1, 30]$, denoted by HCT-L. In these

[1] We use the January 2020 English Wikipedia dump as the external dataset, and collected words that co-occur in a window of ± 5 (`https://dumps.wikimedia.org/enwiki/`).



(a) T-COM (b) 20NG (b) Tweet

Figure 4.4: F1-Scores of HCT and HCT-L with different values of h on T-COM, 20NG and Tweet.

experiments, we set the topic number K=200 and fixed other parameters as previous settings. The results are shown in Figure 4.4. We see that the F1-Scores of HCT-Ls increase sharply as h increases and tend to saturate when h reaches around 12 to 20. As h continues to increase, F-Score starts to gradually decline. Moreover, the optimal context window length in the short text dataset (Tweet) is smaller than those of the other two datasets (T-COM and 20NG). Although the best value of h varies with the dataset, the optimal result of each dataset of HCT-L is still inferior to the results which are obtained by the adaptive context-based HCT. These observations verify that the adaptive context window is effective in the topic modeling process and are also the reasons that the maximum empirical size is set to 20 in our experiments.

4.3.2.5 Efficiency Comparison

We list the average running time (per iteration) for 1000 iterations of our proposals and other baseline topic models on 20NG with $K = \{10,50,100,150,200,250\}$. Moreover, we also compared the efficiency of HCT with three cases of fixed context window lengths $h = 5,10,20$. The experiments were conducted on a machine with Intel i9 processor and 128GB memory.

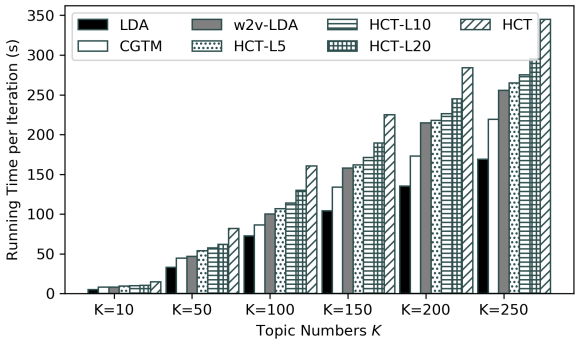


Figure 4.5: Running time per iteration (s) on 20NG of all the baselines and HCT-L with $h = 5, 10, 20$ (denoted by HCT-L5, HCT-L10 and HCT-L20, respectively).

As shown in Figure 4.5, we see that the time cost of all the models increases monotonously as topic number K grows and all the integrated topic models as well as HCT and HCT-Ls cost more time than the traditional topic model LDA. The reasons are, for the integrated models, they use an auxiliary module in topic assignments, which increases their time in computing the topic distribution of each word occurrence; while for HCT-Ls, they generate both words and their corresponding context words, which consumes more time in obtaining each word distribution given a topic. For HCT, due to its consideration of various probable context window lengths within a maximum empirical value, its increase in time consumption is more evident than HCT-Ls. We can derive that the time complexity of HCT-l and traditional LDA are both $O(N_D\overline{S}K)$, where N_D is the total number of documents in the dataset, $\overline{S} = \sum_i S_i/N_D$ is the average size of documents and S_i refers to the length of the ith document. For HCT, its time complexity is $O(N_D\overline{S}KL)$, where L is the context window length. Although the consideration of h in HCT increases its time consumption, the given maximum empirical value is typically less than 20 and independent of the topic numbers. Moreover, based on Algorithm 1 and the generating steps of HCT, we see that the processing related to h contributes a small part of the internal loop. Therefore, the time consuming for HCT

is acceptable, and we see that, as K increases, the change rate of time consumption for HCT and HCT-L is consistent with LDA.

4.3.3 Qualitative Analysis

We analyzed the effectiveness of our model in fine-grained word sense detection from two aspects: (1) the discovering of dataset-specific word sense and (2) the capturing of word sense evolution in the specific topic dimensions. Secondly, we verified how the "Bag-of-Senses" assumption positively affects the topic modeling.

4.3.3.1 Dataset-Specific Word Sense Discovery

Firstly, we verify whether our model can capture useful domain-specific senses by the estimated sense vectors. We randomly sampled 10000 documents from 20NG with 20 classes[1] and 7000 comments from T-COM covering three sensitive themes of "religion", "race" and "homosexual". We respectively select three high frequent words which are likely to cover the most related themes of each datasets according to the Longman Dictionary[2] ("*power*", "*card*", "*key*" for 20NG and "*religion*", "*race*", "*homosexual*" for T-COM) as examples and compute the Word Sense Vector $\boldsymbol{v}_{\boldsymbol{w}_\eta,w}$ of each word w within each context $\boldsymbol{w}_\eta$ by Definition 5. We used the same settings as those in the quantitative analysis for sense vectors clustering, and visualized the results by t-SNE [96]. Silhouette Coefficients for each example word with different cluster numbers are as shown in Figure 4.7. The number of clusters with the highest Silhouette Coefficient is set as the clustering parameter of the corresponding word sense vectors.

[1]These 20 classes mainly cover the themes of "electronics", "sports", "religion", "politics" and "industry" (`http://qwone.com/~jason/20Newsgroups/`).

[2]`https://www.ldoceonline.com/dictionary/`

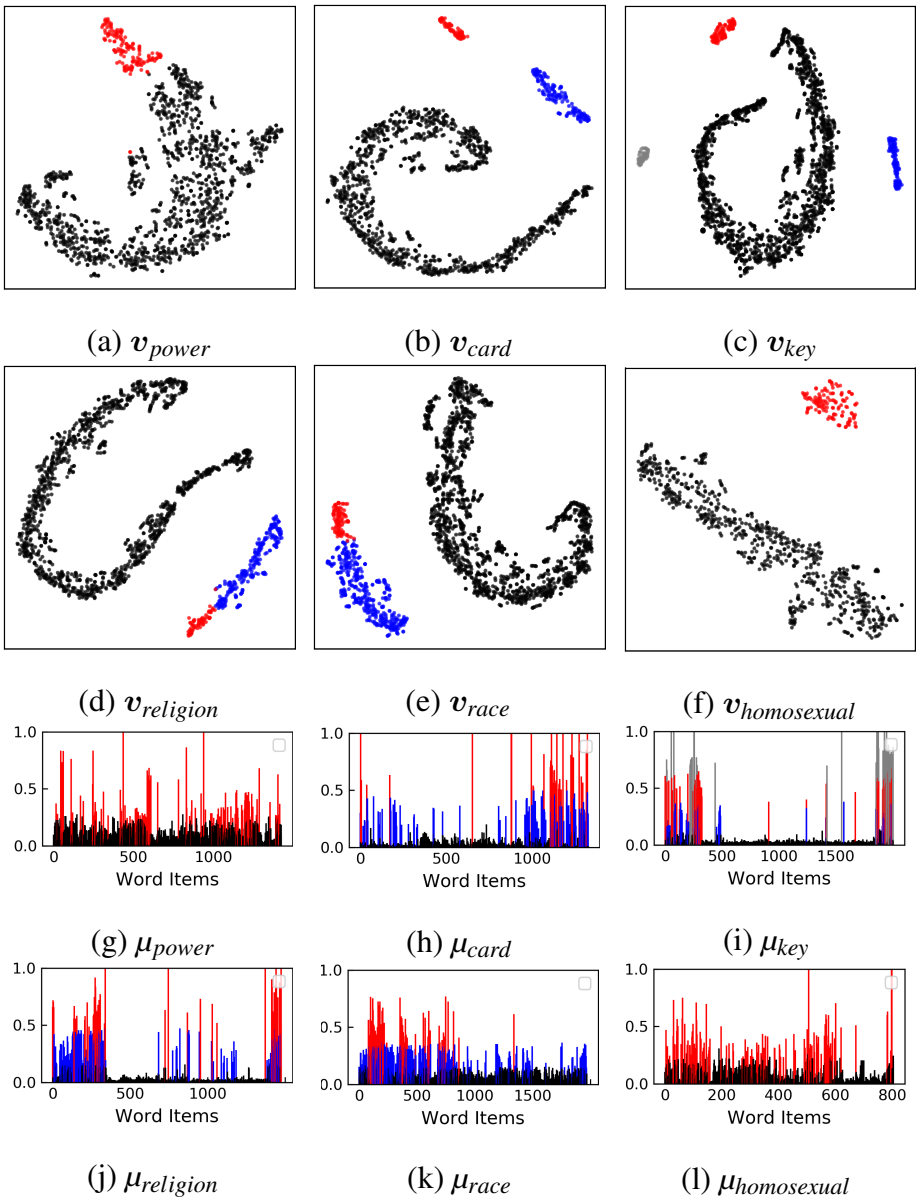
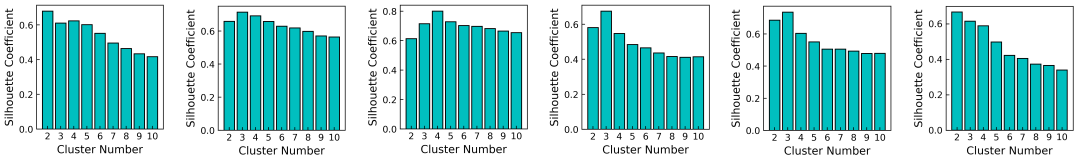


Figure 4.6: Visualization for each example word w by their Word Sense Vectors ($\boldsymbol{v}_w$) and the corresponding Local Sense Weights (μ_w) in the 20NG. Each point in (a-f) or bar in (g-l) refers to a word item in the dataset, where each color corresponds to a sense cluster.

As shown in Figure 4.6 (a-f), each point represents a sense vector, and each color refers to a sense cluster. We see that the sense vectors exhibit varying degrees of clus-

tered properties. This observation verifies our hypothesis in Definition 1. For further study of the semantics for each cluster, we then counted the high frequent context words for each cluster and presented the interpretations which are likely to be relevant to these clusters based on the Longman Dictionary, as shown in Tables 4.3 and 4.4. From Table 4.3, we see that although not all groups of context words can be abstracted to an exact meaning, the differences between them are clear. For instance, the two sense clusters of "*power*" possibly correspond to "*a kind of energy*" ($\boldsymbol{c}_1$) and "*a supernatural ability*" ($\boldsymbol{c}_2$), respectively. The two clusters ($\boldsymbol{c}_1$ and $\boldsymbol{c}_2$) of "*card*" respectively refer to "*a computer-related equipment*" and "*a person identification certificate*". For the word "*key*", the differences between the clusters are obvious, where the senses of $\boldsymbol{c}_1$ are possibly relevant to the sense "*encryption*", the ones of $\boldsymbol{c}_2$ may refer to "*the keyboard buttons*", $\boldsymbol{c}_3$ possibly refer to "*a tool to lock or unlock a door*", and $\boldsymbol{c}_4$ may represents "*a kind of password or serial number*". Combining Tables 4.3 and 4.4, we see that the interpretations for the above senses might be found in the knowledge library, while the following three examples show more particular and fine-grained senses for a specific dataset. For example, our model captured three entities that the word "*religion*" refers to: "*the Christianity*" ($\boldsymbol{c}_1$), "*the Islam*" ($\boldsymbol{c}_2$) and "*the communist*" ($\boldsymbol{c}_3$). Moreover, we can also recognize the positions or tendencies represented by different clusters according to obviously uncomfortable or discriminatory context words, such as $\boldsymbol{c}_2$ and $\boldsymbol{c}_3$ of the word "*race*", as well as $\boldsymbol{c}_2$ of "*homosexual*". These results confirm our assumptions about the word sense vector, e.g., Definition 5, and the effectiveness of obtaining domain-specific senses.

Besides, we further analyzed the relationships between the Local Sense Weight $\mu_{\boldsymbol{w}_\eta,w}$ and $\boldsymbol{v}_{\boldsymbol{w}_\eta,w}$. Figure 4.6 (g-l) shows the weights for each example word, where the colors of bars correspond to those of the clusters in Figure 4.6 (a-f). We observe that clusters with fewer sense vectors have higher weights than others, and vectors



(a) "*power*" (b) "*card*" (a) "*key*" (b) "*religion*" (a) "*race*"(b) "*homosexual*"

Figure 4.7: Silhouette Coefficients for each example words with different clustering numbers of kMeans.

that belong to the same cluster correspond to similar weights. These observations signify that $\mu_{\boldsymbol{w}_{\eta},w}$ reflects a difference between a sense cluster and its corresponding general one. The higher the weights, the more different from its general sense and the more dependent on its context words. These phenomena also confirm a viewpoint of a lexicography sect about the formation of word senses, that corpus citations of a word fall into one or more distinct but related clusters. Each of these clusters, if large enough and distinct enough from others, forms a distinct word sense [47].

4.3.3.2 Word Sense Evolution Detection

Compared with the embedding-based methods, one advantage of the topic models is the explainability of each generated vector dimension. Specifically, according to the topic-word distribution, the Top-n most probable words of each topic can be obtained, based on which the general explanation of each topic dimension can be inferred. Therefore, the other group of experiments are conducted on PubMed, to verify our model on the effectiveness of capturing the evolution of word senses over time in specific topic dimensions.

We sampled two sets of 5000 abstracts from 1975 to 2018, which are respectively related to two cases: (1) "*Fish-oil*" and "*Raynaud*"; (2) "*Alzheimer*" and "*Indomethacin*". Both cases include the evolution of word senses, where "*Fish-oil*" was

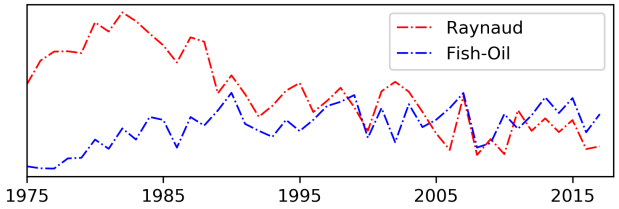
Table 4.3: Context words for each sense cluster of the example words. Bold fonts indicate the high-frequency context words which help clarify the semantic difference. The color for each cluster symbol $\boldsymbol{c}$ corresponds to that of each cluster in Figure 4.6.

Word	Cluster	Selected 5 Words in Top-10 High Frequent Ones in Context
power	$\boldsymbol{c}_1$	"people", "**supply**", "**connector**", "**nuclear**", "**battery**"
	$\boldsymbol{c}_2$	"**god**", "**lord**", "**christ**","**jesus**", "believe"
card	$\boldsymbol{c}_1$	"**video**", "**drive**", "system", "**graphic**", "vga"
	$\boldsymbol{c}_2$	"people","**identify**", "**nationality**", "**number**", "**authority**"
	$\boldsymbol{c}_3$	"key", "**tool**", "guess", "hold", "**game**"
key	$\boldsymbol{c}_1$	"escrow", "**system**", "public", "**encryption**", "**number**", "**security**"
	$\boldsymbol{c}_2$	"**character**", "**application**", "**code**", "**program**", "**system**"
	$\boldsymbol{c}_3$	"**home**", "**car**", "**door**", "**lock**", "available"
	$\boldsymbol{c}_4$	"**drive**", "**machine**", "**number**", "printer", "**series**"
religion	$\boldsymbol{c}_1$	"god", "**jewish**", "**christian**", "**judaism**", "faith"
	$\boldsymbol{c}_2$	"**islam**", "god", "faith", "politics", "**muslim**"
	$\boldsymbol{c}_3$	"god", "eastern", "**socialist**" , "**communist**", "ethnicity"
race	$\boldsymbol{c}_1$	"people", "**religion**", "ethnicity", "**language**", "**human**"
	$\boldsymbol{c}_2$	"**nazi**", "**holocaust**", "victims", "sex", "family"
	$\boldsymbol{c}_3$	"**white**", "ethnic", "**black**", "**asian**", "**nationalism**"
homosexual	$\boldsymbol{c}_1$	"**gay**", "**sex**", "children", "female", "male"
	$\boldsymbol{c}_2$	"man", "**fuck**", "**shit**", "dog", "**ass**"

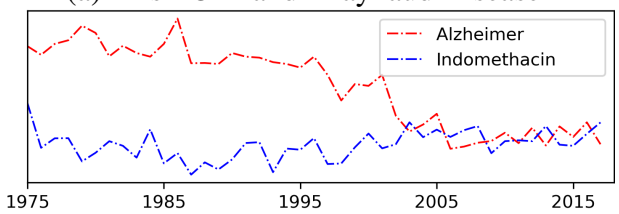
found to prevent "*Raynaud*" disease in 1986 [93] and "*Indomethacin*" has gradually been used as a non-steroidal anti-inflammatory drugs in the treatment of "*Alzheimer*" disease since 1990s [41, 91]. The detailed experiment steps are: (1) we estimated all the sense vectors of the four words, (2) their sense vectors were sorted by time-stamps

Table 4.4: Interpretations in the Longman Dictionary for the generated sense clusters. c_i(s) in each row represents the possibly related cluster(s).

Word	**Cluster(s)**	**Interpretation**
power	c_1	"*energy that make a machine work*"
card	c_1	"*a piece of equipment in a computer*"
	c_2	"*a small piece of plastic or paper that contains information about a person*"
key	c_2	"*the buttons on a computer keyboard*"
	c_3	"*a specially shaped piece of metal to lock or unlock a door, start a car etc*"
religion	c_1, c_2	"*a belief in one or more gods*"
race	c_1, c_2, c_3	"*one of the main groups that humans can be divided into by their colour of skin or other physical features*"
homo-sexual	c_1, c_2	"*someone, especially a man, is sexually attracted to people of the same sex*"



(a) “Fish-Oil” and “Raynaud Disease”



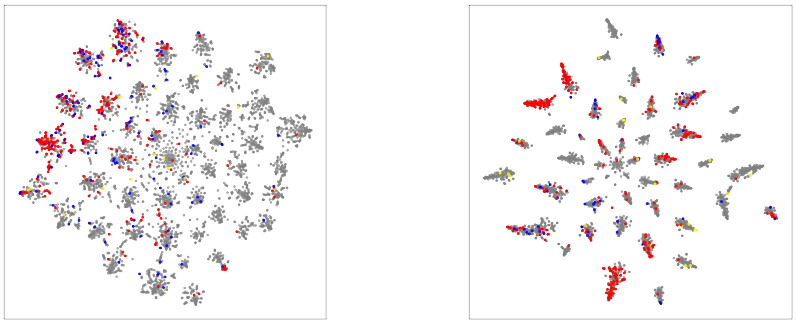
(b) “Indomethacin” and “Alzheimer Disease”

Figure 4.8: Visualization for topic vectors of word pairs “*Fish-Oil*” and “*Raynaud*” (a) and “*Indomethacin*” and “*Alzheimer*”, with only the topic dimensions whose top-10 high-frequency topic words contain word families of “prevent” and “treat”.

and divided into groups by year, respectively, (3) we calculated the average of the sense vectors for each word in years, (4) for the four sequences of the average sense vectors, we selected the topic dimensions whose top-10 high probable topic words contain word families of “prevent” and “treat” and use t-SNE to visualize them into one dimension on the timeline from 1975 to 2018. The results are as shown in Figure 4.8. We see that the two pairs “*Fish-oil*” and “*Raynaud*”, “*Alzheimer*” and “*Indomethacin*” begin to coincide in the late 1980s and 1990s, respectively, which confirms the sense evolution of words “*Fish-oil*” and “*Indomethacin*” in the history and show that HCT is sensitive to the new emerging word senses over time in modeling.

4.3.3.3 Effectiveness of “Bag-of-Senses”

The essential difference between “Bag-of-Words” and “Bag-of-Senses” in the document representation is that a document is represented as a multiset of word senses



(a) “Bag-of-Words” (b) “Bag-of-Senses”

Figure 4.9: Comparison of document vectors on 20NG of the traditional“Bag-of-Words” based topic model (LDA [7]) and the “Bag-of-Senses” based HCT. Each point corresponds to a document, where the red, blue, and yellow ones refer to the documents containing the word “*key*” with its top-3 high frequent word sense clusters, respectively. Note that, documents containing multiple senses are labeled by the color of the sense with the largest number.

instead of words, which alleviates the problem of low semantic discrimination caused by shared identical words among documents. In the quantitative analysis, we have verified the effectiveness of the ”Bag-of-Senses” based word sense vectors and document topic vectors. In this part, we visualize the generated topic vectors of 20NG and show the relationship between the word senses and the document topics.

In the visualization of the document topic vectors for 20NG, as shown in Figure 4.9, we see that compared to the traditional “Bag-of-Words” based topic model (LDA [7]), the diversity of these fine-grained senses has positive effects on document vectors, which improves the discrimination between topic clusters and keeps documents with similar topics be more compact in the topic space. The improved linear separability among documents belonging to different topics can also contribute to the performance of these topic vectors in classification tasks. Moreover, we labeled the documents that contain different senses of the word “*key*”. Due to the different semantic division

standard of words in each baseline, we took the Top-3 frequent senses for visualization. We see that most documents containing the same sense are clustered in the same topic cluster. These observations verify the validity of the hypothesis of the "Bag-of-Senses" in the improvement of document topic modeling.

4.4 Summary

In this chapter, we proposed an adaptive and hybrid context based topic model for handling the WSD problem in document representation without data enrichment. By integrating topic distributions of both the context in which a word occurs and those of its other occurrence in the sense estimation, the proposed model effectively captures domain-specific word senses and preserves the differences between synonyms. Besides, we proposed the "Bag-of-Senses" hypothesis, based on which our model generates senses instead of words. Topic modeling based on "Bag-of-Senses" is more effective in dealing with word sense disambiguation than the methods based on "Bag-of-Words". Our experiments confirm the effectiveness of our model to obtain the fine-grained word sense vectors and showed that our proposal outperforms the baseline models in terms of the sense estimation quality, the classification performance, and the topic modeling accuracy.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this thesis, we focused on studying the topic modeling methods based on two types of context information, the document-level context and the word-level context. We investigated three specific topic modeling tasks, i.e., sequential topic modeling, topic segmentation, and word sense-aware topic modeling to analyze the effect of the context information on their performance, respectively. In Chapter 2, we first considered the document-level context information in topic modeling and proposed a topic model in topic evolution modeling based on hybrid inter-document dependencies. The first model considers Consecutive Dependency, Trend Dependency and Independency in contextual documents. For a sequence of more complex topic evolution, we improved it by considering fine-grained local dependency relations. According to our experiments, we see that our proposals outperform baseline models, in terms of the accuracy of topic modeling, the clustering quality and the effectiveness of outlier detection.

In Chapters 3 and 4, we studied the word-level context based topic models from two tasks: topic segmentation and word sense-aware topic modeling. The difference

is that, for the former task, we use the context words to improve the accuracy of topic assignment for each word, while the latter focuses on integrating the context words of different senses for each identical word. For Topic Segmentation Task, we proposed a new generative model for topic segmentation. By combining topic distribution and context word pairs-topic distribution, the proposed model improves the certainty of the topic assignment and ensures high coherency and saliency in topic segmentation. Besides, by introducing the Topic Coherency Ratio, we designed an optimization algorithm to merge redundant topic segments for each document. Our experiments show that our proposal outperforms baseline models, in terms of both Perplexity and PMI-Score in topic modeling as well as the scores of PK, WD and WDE in topic segmentation.

For word sense-aware topic modeling, we proposed an adaptive and hybrid context based topic model for handling the WSD problem in document representation without data enrichment. By integrating topic distributions of both the context in which a word occurs and those of its other occurrence in the sense estimation, the proposed model effectively captures domain-specific word senses and preserves the differences between synonyms. Besides, we proposed the "Bag-of-Senses" hypothesis, based on which our model generates senses instead of words. Topic modeling based on "Bag-of-Senses" is more effective in dealing with the word sense disambiguation than the methods based on "Bag-of-Words". The experiments confirm the effectiveness of our model to obtain the fine-grained word sense vectors and showed that our proposal outperforms the baseline models in terms of the sense estimation quality, the classification performance, and the topic modeling accuracy.

5.2 Future Work

In future work, we will further investigate the effects of finer-grained context information, e.g., the order of words, and the labeled information on topic modeling. For example, since even one dataset or task may also have different perspectives for word sense division, it is intuitive and essential to study supervised word sense aware topic modeling to restrict the specific perspective of word sense division, such as document labels on word sense disambiguation. Besides, we will further optimize the parameter estimation steps and use more efficient algorithms (e.g., the Variational Inference [43]) to improve the adaptiveness of our model for more substantial scale datasets.

Appendix A. Derivation of our Gibbs Sampling

For the Gibbs sampling procedure in Section 3.2.3, we need to calculate the conditional probability of topic assignment $P_{d,l,k} = P(Z_{d,l} = k|W_{d,l}, \boldsymbol{Z}_{d,-(d,l)}, \boldsymbol{W}'_{d,l}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ for each word, where $W_{d,l}$ represents the lth word in d, $\boldsymbol{Z}_{d,-(d,l)}$ refers to the topic assignments for all words in d except for word $W_{d,l}$, $\boldsymbol{W}'_{d,l}$ are the context words of $W_{d,l}$. $P_{d,l,k}$ is computed as follows:

$$
\begin{aligned}
&P_{d,l,k} \propto P(Z_{d,l} = k, W_{d,l} = t|\boldsymbol{W}'_{d,l}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \\
&= \int\int\int P((Z_{d,l} = k|\boldsymbol{\pi}_{d,l})P(\boldsymbol{\pi}_{d,l}|\boldsymbol{\theta}_d, \boldsymbol{\lambda}_{d,l})d\boldsymbol{\pi}_{d,l}P(\boldsymbol{\lambda}_{d,l}|\boldsymbol{W}'_{d,l}, \boldsymbol{\gamma})P(\boldsymbol{\theta}_d|\boldsymbol{\alpha})d\boldsymbol{\lambda}_{d,l}d\boldsymbol{\theta}_d \\
&\quad \int P(W_{d,l} = t|\boldsymbol{\phi}_k)P(\boldsymbol{\phi}_k|\boldsymbol{\beta})d\boldsymbol{\phi}_k.
\end{aligned}
$$

Based on the definition of Dirichlet distribution, the conditional distribution $P_{d,l,k}$ can be simplified as:

$$
\begin{aligned}
P_{d,l,k} &\propto E_{Dir(\boldsymbol{\beta})}(\phi_{k,t}) \int\int E_{Dir(\boldsymbol{\theta}_d + \boldsymbol{\lambda}_{d,l})}(\pi_{d,l,k})P(\boldsymbol{\lambda}_{d,l}|\boldsymbol{W}'_{d,l}, \boldsymbol{\gamma})P(\boldsymbol{\theta}_d|\boldsymbol{\alpha})d\boldsymbol{\lambda}_{d,l}d\boldsymbol{\theta}_d \\
&= E_{Dir(\boldsymbol{\beta})}(\phi_{k,t}) \int\int \frac{\lambda_{d,l,k} + \theta_{d,k}}{\sum_{s=1}^{K}(\lambda_{d,l,s} + \theta_{d,s})}P(\boldsymbol{\lambda}_{d,l}|\boldsymbol{W}'_{d,l}, \boldsymbol{\gamma})P(\boldsymbol{\theta}_d|\boldsymbol{\alpha})d\boldsymbol{\lambda}_{d,l}d\boldsymbol{\theta}_d.
\end{aligned}
$$

Since the sums of $\lambda_{d,l,k}$ and $\theta_{d,k}$ as well as the integrals of $\boldsymbol{\phi}_k$ and $\boldsymbol{\lambda}_{d,l}$ are constants, $P_{d,l,k}$ can be further simplified as:

$$
\begin{aligned}
P_{d,l,k} \propto & E_{Dir(\boldsymbol{\beta})}(\phi_{k,t}) \int\int (\lambda_{d,l,k} + \theta_{d,k})P(\boldsymbol{\lambda}_{d,l}|\boldsymbol{W}'_{d,l}, \boldsymbol{\gamma})P(\boldsymbol{\theta}_d|\boldsymbol{\alpha})d\boldsymbol{\lambda}_{d,l}d\boldsymbol{\theta}_d \\
= & E_{Dir(\boldsymbol{\beta})}(\phi_{k,t}) \left[\int \lambda_{d,l,k}P(\boldsymbol{\lambda}_{d,l}|\boldsymbol{W}'_{d,l}, \boldsymbol{\gamma})d\boldsymbol{\lambda}_{d,l} + \int \theta_{d,k}P(\boldsymbol{\theta}_d|\boldsymbol{\alpha})d\boldsymbol{\theta} \right] \\
= & E_{Dir(\boldsymbol{\beta})}(\phi_{k,t}) \left[E_{Dir(\boldsymbol{\gamma})}(\lambda_{d,l,k}) + E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k}) \right].
\end{aligned}
$$

According to the definition of expectation of Dirichlet Distribution, we have:

$$E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k}) = \frac{n_{d,k,-(d,l)} + \alpha_k}{\sum_{s=1}^{K}(n_{d,s,-(d,l)} + \alpha_s)},$$

$$E_{Dir(\boldsymbol{\beta})}(\phi_{k,t}) = \frac{n^t_{k,-(d,l)} + \beta_t}{\sum_{f=1}^{V}(n^f_{k,-(d,l)} + \beta_f)},$$

where $n_{d,k,-(d,l)}$ is the number of words in document d which belongs to topic k without $W_{d,l}$, $n^t_{k,-(d,l)}$ represents the number of word t which belongs topic k without $W_{d,l}$. Substituting the results of $E_{Dir(\gamma)}(\boldsymbol{\lambda}_w)$, $E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k})$ and $E_{Dir(\boldsymbol{\beta})}(\phi_{k,t})$ into Eq. (1) and removing constant terms, we obtain the conditional probability $P_{d,l,k}$ as below:

$$P_{d,l,k} \propto \left[(n^{\boldsymbol{b}_w}_{k,-(d,i)} + \gamma_t) + (n_{d,k,-(d,l)} + \alpha_k)\right] \frac{n^t_{k,-(d,l)} + \beta_t}{\sum_{f=1}^{V}(n^f_{k,-(d,l)} + \beta_f)}.$$

Appendix B. Derivation of Topic Coherency Ratio

We provide here the complete derivation of the Topic Coherency Ratio in Section 3.2.4. For consecutive words $\boldsymbol{W}_{d,i:j}$ from W_i to W_j in document d, we denote the joint probability of sharing topic k by $P(\boldsymbol{W}_{d,i:j}, k)$ in the former case and the one in the latter case by $P'(\boldsymbol{W}_{d,i:j}, k)$. According to Eq. (1), the two joint probabilities can be respectively computed by:

$$P(\boldsymbol{W}_{d,i:j}, k) \propto \prod_{w \in \boldsymbol{W}_{d,i:j}} E_{Dir(\boldsymbol{\beta})}(\phi_{d,w}) \left[\sum_{s=1}^{S} I_{d,l,s} E_{Dir(\gamma)}(\lambda^s_{d,l,k}) + E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k})\right],$$

$$P'(\boldsymbol{W}_{d,i:j}, k) \propto \prod_{w \in \boldsymbol{W}_{d,i:j}} E_{Dir(\boldsymbol{\beta})}(\phi_{d,w}) E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k}).$$

Taking their logarithms and computing their ratio as well as removing constant terms:

$$\begin{aligned}
&\frac{\log P(\boldsymbol{W}_{d,i:j},k)}{\log P'(\boldsymbol{W}_{d,i:j},k)} \\
&= \frac{\log \prod_{w\in\boldsymbol{W}_{d,i:j}} E_{Dir(\boldsymbol{\beta})}(\phi_{d,w}) \cdot \left[E_{Dir(\boldsymbol{\gamma})}(\lambda_{d,l,k}) + E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k})\right]}{\log \prod_{w\in\boldsymbol{W}_{d,i:j}} E_{Dir(\boldsymbol{\beta})}(\phi_{d,w}) \cdot E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k})} \\
&= \frac{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log E_{Dir(\boldsymbol{\beta})}(\phi_{d,w}) + \sum_{w\in\boldsymbol{W}_{d,i:j}} \log \left[E_{Dir(\boldsymbol{\gamma})}(\lambda_{d,l,k}) + E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k})\right]}{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log E_{Dir(\boldsymbol{\beta})}(\phi_{d,w}) + \log E_{Dir(\boldsymbol{\alpha})}(\theta_{d,k})}.
\end{aligned}$$

Since the topic distribution is constant for words in the a document, we obtain:

$$\begin{aligned}
&\frac{\log P(\boldsymbol{W}_{d,i:j},k)}{\log P'(\boldsymbol{W}_{d,i:j},k)} \\
&\propto \frac{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log E_{Dir(\boldsymbol{\beta})}(\phi_{d,w}) + \sum_{w\in\boldsymbol{W}_{d,i:j}} \log E_{Dir(\boldsymbol{\gamma})}(\lambda_{d,l,k})}{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log E_{Dir(\boldsymbol{\beta})}(\phi_{d,w})} \\
&= 1 + \frac{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log E_{Dir(\boldsymbol{\gamma})}(\lambda_{d,l,k})}{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log E_{Dir(\boldsymbol{\beta})}(\phi_{d,w})}.
\end{aligned}$$

According to Eqs. (1), (3) and (4):

$$\frac{\log P(\boldsymbol{W}_{d,i:j},k)}{\log P'(\boldsymbol{W}_{d,i:j},k)} \propto 1 + \frac{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log n_{k,-(d,i)}^{\boldsymbol{b}_w}}{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log n_{k,-(d,l)}^{w}}.$$

Removing the constant term, we obtain the Topic Coherency Ratio:

$$R_t(\boldsymbol{W}_{d,i:j},k) = \frac{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log n_{k,-(d,i)}^{\boldsymbol{b}_w}}{\sum_{w\in\boldsymbol{W}_{d,i:j}} \log n_{k,-(d,l)}^{w}}.$$

Appendix C.Redundant Topic Merging (RTM) Algorithm

The pseudo code in Section 3.2.4 is shown as below.Each segment S_i is a set of words in the ith segment and T_i represents its topic. To express the algorithm concisely, we implicitly handle the cases of the list items at the head and the tail in S_{i+1} or S_{i-1}.

Algorithm 4: Redundant topic merging algorithm

Input: List S of topic segments in a document and their corresponding topic sequence T; the size L of context window

Output: Optimized segment list S'

```
Initialize S' as a null list
n = |S|
while |S'| ≠ n do
    Clear S'
    for i = 1 in |S| do
        if R_t(S_{i-1} ∪ S_i, T_{i-1}) > R_t(S_{i+1} ∪ S_i, T_{i+1}) and
          R_t(S_{i-1} ∪ S_i, T_{i-1}) > R_t(S_i, T_i) then
            Append S_i ∪ S_{i-1} to S'
        else if R_t(S_{i+1} ∪ S'_i, T_{i+1}) > R_t(S_{i-1} ∪ S_i, T_{i-1}) and
          R_t(S_{i+1} ∪ S_i, T_{i+1}) > R_t(S_i, T_i) then
            Append S_i ∪ S_{i+1} to S'
        else
            Append S_i to S'
    n = |S|
    S ← S'
return S'
```

Bibliography

[1] H. Amoualian, M. Clausel, E. Gaussier, and M.-R. Amini. Streaming-LDA: A Copula-based Approach to Modeling Topic Dependencies in Document Streams. In *Proc. SIGKDD*, pages 695–704, 2016.

[2] H. Amoualian, W. Lu, and M. Gaussier. Topical Coherence in LDA-based Models through Induced Segmentation. In *Proc. ACL*, volume 1, pages 1799–1809, 2017.

[3] G. Balikas, H. Amoualian, M. Clausel, E. Gaussier, and M. R. Amini. Modeling Topic Dependencies in Semantically Coherent Text Spans with Copulas. In *Proc. COLING*, pages 1767–1776, 2016.

[4] D. Beeferman, A. Berger, and J. Lafferty. Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3):177–210, 1999.

[5] D. M. Blei and J. D. Lafferty. Dynamic Topic Models. In *Proc. ICML*, pages 113–120, 2006.

[6] D. M. Blei and P. J. Moreno. Topic Segmentation with an Aspect Hidden Markov Model. In *Proc. SIGIR*, pages 343–348, 2001.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[8] S. Bowman and C. Willis. We Media: How Audiences are Shaping the Future of News and Information. *The Media Center, American Press Institute*, 2015.

[9] J. Boyd-Graber, D. Blei, and X. Zhu. A Topic Model for Word Sense Disambiguation. In *Proc. EMNLP-CoNLL*, pages 1024–1033, 2007.

[10] J. N. Cappella and K. H. Jamieson. *Spiral of Cynicism: The Press and the Public Good*. Oxford University Press, 1997.

[11] C. M. Carlo. Markov Chain Monte Carlo and Gibbs Sampling. *Lecture Notes for EEB*, 581, 2004.

[12] M. Casillo, F. Clarizia, G. D'Aniello, M. De Santo, M. Lombardi, and D. Santaniello. CHAT-Bot: A Cultural Heritage Aware Teller-Bot for Supporting Touristic Experiences. *Pattern Recognition Letters*, 131:234–243, 2020.

[13] V. Chandola, A. Banerjee, and V. Kumar. Anomaly Detection: A Survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[14] D. S. Chaplot and R. Salakhutdinov. Knowledge-Based Word Sense Disambiguation Using Topic Models. In *Proc. AAAI*, 2018.

[15] Y. Chaudhary, H. Schütze, and P. Gupta. Explainable and Discourse Topic-Aware Neural Language Understanding. In *Proc. ICML*, 2020.

[16] X. Cheng, X. Yan, Y. Lan, and J. Guo. BTM: Topic Modeling Over Short Texts. *IEEE Transactions on Knowledge & Data Engineering*, (1):1–1, 2014.

[17] F. Clarizia, F. Colace, M. De Santo, M. Lombardi, F. Pascale, and A. Pietrosanto. E-learning and Sentiment Analysis: A Case Study. In *Proc. ICIET*, pages 111–118, 2018.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL*, pages 4171–4186, 2019.

[19] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel *k*-Means: Spectral Clustering and Normalized Cuts. In *Proc. SIGKDD*, pages 551–556, 2004.

[20] L. Du, J. K. Pate, and M. Johnson. Topic Segmentation with an Ordering-Based Topic Model. In *Proc. AAAI*, 2015.

[21] J. Eisenstein and R. Barzilay. Bayesian Unsupervised Topic Segmentation. In *Proc. EMNLP*, pages 334–343, 2008.

[22] Y. Em, F. Gag, Y. Lou, S. Wang, T. Huang, and L.-Y. Duan. Incorporating Intra-Class Variance to Fine-Grained Visual Recognition. In *Proc. ICME*, pages 1452–1457, 2017.

[23] A. Ferrugento, H. G. Oliveira, A. Alves, and F. Rodrigues. Can Topic Modelling Benefit from Word Sense Information? In *Proc. LREC*, pages 3387–3393, 2016.

[24] J. R. Firth. Ethnographic Analysis and Language with Reference to Malinowski's Views. *Man and Culture: An Evaluation of The Work of Bronislaw Malinowski*, pages 93–118, 1957.

[25] J. G. Fiscus and G. R. Doddington. Topic Detection and Tracking Evaluation Overview. In *Topic Detection and Tracking*, pages 17–31. 2002.

[26] J. Foulds. Mixed Membership Word Embeddings for Computational Social Science. In *Proc. AISTATS*, 2018.

[27] A. Gama, M. Pechenizkiy, and A. Bouchachia. A Survey on Concept Drift Adaptation. *ACM Computing Surveys (CSUR)*, 46(4):44, 2014.

[28] W. Gao, M. Peng, H. Wang, Y. Zhang, Q. Xie, and G. Tian. Incorporating Word Embeddings into Topic Modeling of Short Text. *Knowledge and Information Systems*, 61(2):1123–1145, 2019.

[29] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. PAMI*, 6(6):721–741, 2009.

[30] Y. Goldberg and O. Levy. Word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. *arXiv preprint arXiv:1402.3722*, 2014.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[32] W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *arXiv preprint arXiv:1605.09096*, 2016.

[33] Z. Hao, G. Kim, and E. P. Xing. Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data. In *Proc. SIGKDD*, pages 1425–1434, 2015.

[34] Q. He, K. Chang, E.-P. Lim, and J. Zhang. Bursty Feature Representation for Clustering Text Streams. In *Proc. ICDM*, pages 491–496, 2007.

[35] A. Huang, D. Milne, E. Frank, and I. H. Witten. Clustering Documents Using A Wikipedia-based Concept Representation. In *Proc. PAKDD*, pages 628–636, 2009.

[36] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proc. ACL*, pages 873–882, 2012.

[37] J. Huang, M. Peng, H. Wang, J. Cao, W. Gao, and X. Zhang. A Probabilistic Method for Emerging Topic Tracking in Microblog Stream. *World Wide Web*, 20(2):325–350, 2017.

[38] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic Tracking Model for Analyzing Consumer Purchase Behavior. In *Proc. IJCAI*, volume 9, pages 1427–1432, 2009.

[39] H. Jeffreys et al. *Scientific Inference*. Cambridge University Press, 1973.

[40] J. W. Jenks. The Guidance of Public Opinion. *American Journal of Sociology*, 1(2):158–169, 1895.

[41] K. Jha, G. Xun, Y. Wang, V. Gopalakrishnan, and A. Zhang. Concepts-Bridges: Uncovering Conceptual Bridges Based on Biomedical Concept Evolution. In *Proc. SIGKDD*, 2018.

[42] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*, volume 165. Wiley New York, 1997.

[43] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999.

[44] R. Kannan, H. Woo, C. C. Aggarwal, and H. Park. Outlier Detection for Text Data. In *Proc. SDM*, pages 489–497, 2017.

[45] R. Kannan, H. Woo, C. C. Aggarwal, and H. Park. Outlier Detection for Text Data: An Extended Version. *arXiv preprint arXiv:1701.01325*, 2017.

[46] A. Kilgarriff. Dictionary Word Sense Distinctions: An Enquiry into Their Nature. *Computers and the Humanities*, 26(5-6):365–387, 1992.

[47] A. Kilgarriff. I Don't Believe in Word Senses. *Computers and the Humanities*, 31(2):91–113, 1997.

[48] M. Kontaki, A. Gounaris, A. N. Papadopoulos, K. Tsichlas, and Y. Manolopoulos. Continuous Monitoring of Distance-based Outliers Over Data Streams. In *Proc. ICDE*, pages 135–146, 2011.

[49] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, 2015.

[50] S. Lamprier, T. Amghar, B. Levrat, and F. Saubion. On Evaluation Methodologies for Text Segmentation Algorithms. In *Proc. ICTAI*, volume 2, pages 19–26, 2007.

[51] Q. Le and T. Mikolov. Distributed Representations of Sentences and Documents. In *Proc. ICML*, pages 1188–1196, 2014.

[52] S. Lefkimmiatis, P. Maragos, and G. Papandreou. Bayesian Inference on Multiscale Models for Poisson Intensity Estimation: Applications to Photon-limited Image Denoising. *IEEE Transactions on Image Processing*, 18(8):1724–1741, 2009.

[53] L. Li, B. Roth, and C. Sporleder. Topic Models for Word Sense Disambiguation and Token-Based Idiom Detection. In *Proc. ACL*, 2010.

[54] S. Liang, Z. Ren, E. Yilmaz, and E. Kanoulas. Collaborative User Clustering for Short Text Streams. In *Proc. AAAI*, pages 3504–3510, 2017.

[55] S. Liang, Z. Ren, E. Yilmaz, and E. Kanoulas. Collaborative User Clustering for Short Text Streams. In *Proc. AAAI*, pages 3504–3510, 2017.

[56] S. Liang, E. Yilmaz, and E. Kanoulas. Dynamic Clustering of Streaming Short Documents. In *Proc. SIGKDD*, pages 995–1004, 2016.

[57] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical Word Embeddings. In *Proc. AAAI*, 2015.

[58] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical Word Embeddings. In *Proc. AAAI*, 2015.

[59] S. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[60] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling Word Burstiness Using the Dirichlet Distribution. In *Proc. ICML*, pages 545–552, 2005.

[61] C. Manning, P. Raghavan, and H. Schütze. Introduction to Information Retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[62] N. Meghanathan, S. Boumerdassi, N. Chaki, and D. Nagamalai. Recent Trends in Networks and Communications. In *International Conferences, NeCoM 2010, WiMoN 2010, WeST 2010*, volume 90. Springer, 2010.

[63] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119, 2013.

[64] G. A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[65] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing Semantic Coherence in Topic Models. In *Proc. EMNLP*, 2011.

[66] A. Neelakantan, J. Shankar, A. Passos, and A. McCallum. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *Proc. EMNLP*, pages 1059–1069, 2014.

[67] L. L. Newman. Faith, Spirituality, and Religion: A Model for Understanding the Differences. *College Student Affairs Journal*, 23(2):102–110, 2004.

[68] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson. Improving Topic Models with Latent Feature Word Representations. *TACL*, 3:299–313, 2015.

[69] L. Niu, X. Dai, J. Zhang, and J. Chen. Topic2Vec: Learning Distributed Representations of Topics. In *Proc. IALP*, 2015.

[70] A. C. Pandey, D. S. Rajpoot, and M. Saraswat. Twitter Sentiment Analysis Using Hybrid Cuckoo Search Method. *Information Processing & Management*, 53(4):764–779, 2017.

[71] D. Peng, D. Guilan, and Z. Yong. Contextual-LDA: A Context Coherent Latent Topic Model for Mining Large Corpora. In *Proc. BigMM*, pages 420–425, 2016.

[72] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep Contextualized Word Representations. In *Proc. NAACL*, pages 2227–2237, 2018.

[73] L. Pevzner and M. A. Hearst. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

[74] D. Pfitzner, R. Leibbrandt, and D. Powers. Characterization and Evaluation of Similarity Measures for Pairs of Clusterings. *Knowledge and Information Systems*, 19(3):361, 2009.

[75] M. Purver. Topic Segmentation. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317, 2011.

[76] A. radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving Language Understanding by Generative Pre-Training. 2018.

[77] J. Reisinger and R. Mooney. A Mixture Model with Sharing for Lexical Semantics. In *Proc. EMNLP*, pages 1173–1182. ACL, 2010.

[78] J. C. Reynar. Topic Segmentation: Algorithms and Applications. *Ph.D. Thesis, Institute for Research in Cognitive Science Technical, University of Pennsylvania*, 1998.

[79] D. W. Robinson and D. Ruelle. Mean Entropy of States in Classical Statistical Mechanics. *Communications in Mathematical Physics*, 5(4):288–300, 1967.

[80] M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proc. WSDM*, pages 399–408, 2015.

[81] M. Sahlgren. The Distributional Hypothesis. *Italian Journal of Disability Studies*, 20:33–53, 2008.

[82] C. Sauper, A. Haghighi, and R. Barzilay. Content Models with Attitude. In *Proc. ACL*, pages 350–358, 2011.

[83] J. Schneider and M. Vlachos. Topic Modeling Based on Keywords and Context. In *Proc. ICDM*, pages 369–377, 2018.

[84] T. Seidenfeld. Entropy and Uncertainty. *Philosophy of Science*, 53(4):467–491, 1986.

[85] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[86] B. Shi, W. Lam, S. Jameel, S. Schockaert, and K. P. Lai. Jointly Learning Word Embeddings and Latent Topics. In *Proc. SIGIR*, 2017.

[87] B. Shi, W. Lam, S. Jameel, S. Schockaert, and K. P. Lai. Jointly Learning Word Embeddings and Latent Topics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 375–384, 2017.

[88] D. Sisodia, L. Singh, S. Sisodia, and K. Saxena. Clustering Techniques: A Brief Survey of Different Clustering Algorithms. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, 1(3):82–87, 2012.

[89] J. Sivic and A. Zisserman. Efficient Visual Search of Videos Cast As Text Retrieval. *IEEE Trans. PAMI*, 31(4):591–606, 2008.

[90] B. Skaggs and L. Getoor. Topic Modeling for Wikipedia Link Disambiguation. *ACM TOIS*, 32(3):1–24, 2014.

[91] N. R. Smalheiser and D. R. Swanson. Indomethacin and Alzheimer's Disease. *Neurology*, 46(2):583–583, 1996.

[92] S. Sun, Y. Cheng, Z. Gan, and J. Liu. Patient Knowledge Distillation for BERT Model Compression. In *Proc. EMNLP-IJCNLP*, pages 4314–4323, 2019.

[93] D. R. Swanson. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology & Medicine*, 30(1):7, 2015.

[94] G. Tang, Y. Xia, J. Sun, M. Zhang, and T. F. Zheng. Statistical Word Sense Aware Topic Models. *Soft Computing*, 19(1):13–27, 2015.

[95] K. Tudor. Religion, Faith, Spirituality, and the Beyond in Transactional Analysis. *Transactional Analysis Journal*, 49(2):71–87, 2019.

[96] L. Van der Maaten and G. Hinton. Visualizing Data using t-SNE. *JMLR*, 9(Nov):2579–2605, 2008.

[97] H. Wang, D. Zhang, and C. Zhai. Structural Topic Model for Latent Topical Structure Analysis. In *Proc. ACL*, pages 1526–1535, 2011.

[98] W. Wang, Z. Gan, W. Wang, D. Shen, J. Huang, W. Ping, S. Satheesh, and L. Carin. Topic Compositional Neural Language Model. In *Proc. AISTATS*, pages 356–365, 2018.

[99] X. Wang and A. McCallum. Topics over Time: A Non-Markov Continuous-time Model of Topical Trends. In *Proc. SIGKDD*, pages 424–433, 2006.

[100] X. Wang, A. McCallum, and X. Wei. Topical N-Grams: Phrase and Topic Discovery, with An Application to Information Retrieval. In *Proc. ICDM*, pages 697–702, 2007.

[101] Y. Wang, E. Agichtein, and M. Benzi. TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media. In *Proc. SIGKDD*, pages 123–131, 2012.

[102] W. Weaver. Translation. *Machine Translation of Languages: Fourteen Essays*, 1949.

[103] X. Wei, J. Sun, and X. Wang. Dynamic Mixture Models for Multiple Time-Series. In *Proc. IJCAI*, volume 7, pages 2909–2914, 2007.

[104] G. Xun, Y. Li, W. X. Zhao, J. Gao, and A. Zhang. A Correlated Topic Model Using Word Embeddings. In *IJCAI*, pages 4207–4213, 2017.

[105] X. Yan, J. Guo, Y. Lan, J. Xu, and X. Cheng. A Probabilistic Model for Bursty Topic Discovery in Microblogs. In *Proc. AAAI*, pages 353–359.

[106] G. Yang, D. Wen, N.-S. Chen, E. Sutinen, et al. A Novel Contextual Topic Model for Multi-Document Summarization. *Expert Systems with Applications*, 42(3):1340–1352, 2015.

[107] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong. Dynamic Word Embeddings for Evolving Semantic Discovery. In *Proc. WSDM*, pages 673–681, 2018.

[108] J. Yin and J. Wang. A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering. In *Proc. SIGKDD*, pages 233–242, 2014.

[109] Y. Zhang, A. Jatowt, S. S. Bhowmick, and K. Tanaka. The Past is not a Foreign Country: Detecting Semantically Similar Terms across Time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807, 2016.

[110] L. Zhu, Y. He, and D. Zhou. A Neural Generative Model for Joint Learning Topics and Topic-Specific Word Embeddings. *Transactions of the Association for Computational Linguistics*, 8:471–485, 2020.

[111] Y. Zuo, J. Zhao, and K. Xu. Word Network Topic Model: A Simple but General Solution For Short and Imbalanced Texts. *Knowledge and Information Systems*, 48(2):379–398, 2016.

Published Papers

1 **Wenbo Li** and Einoshin Suzuki. Hybrid Context-Aware Word Sense Disambiguation in Topic Modeling based Document Representation. In *Proc. IEEE International Conference on Data Mining (ICDM)*, pp. 332-341, 2020.

1 **Wenbo Li**, Tetsu Matsukawa, Hiroto Saigo, and Einoshin Suzuki. Context-Aware Latent Dirichlet Allocation for Topic Segmentation. In *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 475-486, 2020.

2 **Wenbo Li**, Hiroto Saigo, Bin Tong, and Einoshin Suzuki. Topic Modeling for Sequential Documents based on Hybrid Inter-Document Topic Dependency. *Journal of Intelligent Information Systems*, 56(3), pages 1-24, 2021.

4 **Wenbo Li** and Einoshin Suzuki. Adaptive and Hybrid Context-Aware Fine-Grained Word Sense Disambiguation in Topic Modeling Based Document Representation. *Information Processing and Management*, 58(4), Article 102592