博士論文

Reconstruction and Bayesian Nonparametrics Based Multi-Document Summarization

（再構成とノンパラメトリックベイズ手法に基づく複数文書要約）

馬　騰飛

Reconstruction and Bayesian Nonparametrics Based Multi-Document Summarization

Tengfei Ma

Adviser: Professor Hiroshi Nakagawa

December 2014

Abstract

Document summarization aims to extract the most important information from a single document or a cluster of documents. It plays an increasingly important role with the exponential growth of web documents. Over the past half a century, there are various approaches proposed to solve the problem from many different perspectives, most of which directly selected summary sentences using sentence ranking or greedy selection approaches. Generally the quality of a summary should be determined by three properties: relevance, diversity and coverage. However, the sentence ranking methods and greedy selection approaches hardly simultaneously consider the three properties, and they could not provide a solution which selects best overall sentences. Therefore optimizing all three properties jointly with a global sentence selection procedure has been attractive.

In this thesis, we solve the summarization problem and unify all aims from a novel perspective. We assumed that original documents should be reconstructed from the best summary with least information loss. From this assumption, we first propose a reconstruction based optimization framework for multi-document summarization. We brought in various information-theoretic measures and regarded the minimum distortion as the objective function. We defined three reconstruction models for optimization of the distortion measures, gaining state-of-the-art summarization results.

Moreover, we studied a new problem in summarization called summary length determination. Traditional summarization systems require users to pre-define a bounded length for summaries. However, how to find the proper summary length is quite a problem; and keeping all summaries restricted to the same length is not always a good choice. Following our reconstruction assumption, we developed a Bayesian nonparametric model to automatically determine the proper summary length. The model is demonstrated to own good summary qualities and to determine rational summary length. Finally, we consider the case that the real categories of documents are not known, and advanced the hybrid nested Dirichlet process to extend traditional Bayesian nonparametric topic analysis, which is a preprocessing step for document summarization. The topic analysis itself also provides visualization for abstractive summarization of the documents.

Acknowledgements

First and foremost, I would like to express my appreciation to my supervisor, Prof. Hiroshi Nakagawa, for his valuable advice and enormous support throughout my Ph.D. course. He is always kind to help guide me on both academic and personal decisions. My appreciation also goes to Prof. Issei Sato, whose insightful comments and rich experience have influenced my research a lot. I would like to thank him for his financial support as well. He provided me the RA position in the last half year and opportunities to attend conferences, such as ICML in Beijing.

I would like to thank my advisor in Peking University, Prof. Xiaojun Wan, who exposed me to the fascinating world of academic research. I have learned various research skills from him and I was deeply influenced by his attitude to do rigourous research. I am grateful to my advisor in Bioinformatics Institute, A*STAR Singapore, Dr. Cheng Li, who offered me a chance to do an overseas internship and broaden my horizon in machine learning research. I also appreciate Long Jiang and Ming Zhou, who supervised me in Microsoft Asia and provided me unforgettable experience of working in a team.

It is my pleasure to work as a member in Nakagawa Lab. I am grateful to all the members of the lab. Especially I would like to thank Bin Yang, who helped me adapt to the new environment in Japan. I also thank Junpei Komiyama and Yo Ehara for their useful discussions and personal help. In addition, I wish to thank my former colleagues in Peking University, Microsoft Research Asia and A*STAR Singapore for their friendship and sharing insights. I would like to thank the University of Tokyo for the scholarship which makes my Ph.D. study possible.

Finally, I owe my deepest gratitude to my family and girlfriend for their encouragement and love. My research and study would not have been possible without their support.

To my parents.

Contents

Abstract . . . iii

Acknowledgements . . . v

List of Tables . . . xi

List of Figures . . . xii

1 Introduction 1

1.1 Thesis Contribution . . . 5

1.2 Thesis Overview . . . 5

2 Document Summarization Overview 7

2.1 Taxonomy of Summarization . . . 8

2.1.1 Single-document V.S. Multi-document . . . 8

2.1.2 Extraction-based V.S. Abstraction-based . . . 9

2.1.3 Supervised V.S. Unsupervised . . . 9

2.1.4 Generic V.S. Query-Focused . . . 10

2.1.5 New Types . . . 10

2.2 Common Approaches to Document Summarization . . . 11

2.2.1 Sentence Scoring and Sentence Ranking . . . 11

2.2.2 Greedy Selection V.S. Optimization . . . 13

2.3 Evaluation . . . 14

2.3.1 Intrinsic Evaluation . . . 14

2.3.2 Extrinsic Evaluation 16

2.4 Real Systems 16

2.5 Relevant Evaluation Workshops 18

3 Multi-document Summarization using Minimum Distortion 23

3.1 Motivation and Problem Formulation 24

3.2 Distortion Measures 26

3.2.1 Hamming Distortion 26

3.2.2 Squared Error Distortion 27

3.2.3 Information Divergence (KLD) 27

3.2.4 Jensen-Shannon Divergence (JSD) 28

3.2.5 Jensen-Shannon Divergence with Smoothing (JSDS) 28

3.2.6 Other Distortion Measures 29

3.3 Cluster-based Reconstruction Models 30

3.3.1 P-median Clustering Model 30

3.3.2 Facility Location Model 32

3.4 Linear Representation Model 33

3.4.1 Motivation 33

3.4.2 The Approach of the Linear Representation Model 35

3.5 Experiments 37

3.5.1 Data Sets 37

3.5.2 Evaluation Metric 38

3.5.3 Experimental Results 38

3.6 Conclusions and Problems 42

4 Topic Models and Bayesian Nonparametrics 44

4.1 Probabilistic Topic Models 44

4.1.1 Probabilistic Latent Semantic Indexing 45

4.1.2 Latent Dirichlet Allocation 46

4.1.3 Other Topic Models 46

4.2 Topic Models for Document Summarization 46

4.3 Bayesian Nonparametric Methods 47

4.3.1 Dirichlet Process Mixture Models 47

4.3.2 The Hierarchical Dirichlet Process 48

4.3.3 LC-HDP 50

4.3.4 The Nested Dirichlet Process 51

4.3.5 Beta Processes and Latent Feature Analysis 52

4.3.6 Bayesian Nonparametric Topic Models 53

4.3.7 Bayesian Nonparametric Methods in Document Summarization 54

4.4 Hybrid Nested Dirichlet Processes for Topic Modeling 54

4.4.1 Introduction 54

4.4.2 The Hybrid Nested/Hierarchical Dirichlet Process 58

4.4.3 Application to Topic Modeling 60

4.4.4 The hNHDP model for Topic Modeling 61

4.4.5 Inference 64

4.4.6 Experiments 66

4.4.7 Conclusions 76

5 Bayesian Nonparametric Summarization and Summary Length Determination 78

5.1 Background 78

5.2 Related Work 80

5.2.1 Research on Summary Length 80

5.2.2 BNP Methods in Document Summarization 81

5.3 BNP Summarization 82

5.3.1 Framework of BNP Summarization 82

5.4 Variational Inference 85

5.5 Experiments 89

5.5.1 Evaluation of Summary Qualities 89

5.5.2 A New Evaluation Metric 94

5.5.3 Necessity of Varying Summary Length 96

5.6 Conclusion and Future Work 99

6 Conclusion 101

A Proof of the Distortion Bound 1 103

B Proof of the Distortion Bound 2 105

Bibliography 107

List of Tables

2.1 Summarization of all tasks in the evaluation conferences (mostly cited from [52]). 22

3.1 Experimental Results on DUC2004 Data. 40

3.2 Experimental Results on DUC2002 Data. 40

3.3 Comparison with Other Models on DUC2004. 41

3.4 Comparison with Other Models on DUC2002. 42

4.1 Comparison of different models. 56

4.2 Some example topics extracted from *6conf* data. Each column is a topic; for each topic the top 15 words are shown. The numbers in brackets are the typical cluster numbers of each conference shown in Figure 4.8. 74

4.3 Description of datasets for clustering. 74

5.1 Average summary length (number of words) on different datasets . . 97

5.2 Comparison of summary lengths on Separate Dataset. 98

5.3 Comparison of summary lengths on Combined1 Dataset. 98

5.4 Comparison of summary lengths on Combined2 Dataset. 99

List of Figures

1.1 An illustration of the summary length problem. 4

2.1 Ultimate Research Assistant. 18

2.2 iResearch Reporter. 19

2.3 Newsblaster. 19

2.4 Yahoo! news digest. 20

3.1 Transmission model. 26

4.1 Graphical model representations of (a) HDP and (b)LC-HDP. 50

4.2 Graphical model representation of hNHDP. 57

4.3 Clustering results on toy dataset1. (a)-(d) show the clustering assignments at different iterations: (a) real cluster assignments, (b)initial random assignments, (c) assignments after one iteration, and (d) final assignments (after 30 iterations). 67

4.4 Word differentiation on toy dataset1. 68

4.5 Results of document modeling on NIPS data. 70

4.6 Results of document modeling on 6conf data. 71

4.7 Comparison of hNHDP and hNHDP-knowncategory. 72

4.8 Visualization of the clustering results on *6conf* data. The red lines indicate the real labels, while the blue points indicate the clustering assignments. 75

4.9 Clustering comparison of hNHDP and NDP. 76

5.1 Rouge-1 values on DUC2004 dataset. 90

5.2 Rouge-2 values on DUC2004 dataset. 91

5.3 Rouge-L values on DUC2004 dataset. 92

5.4 Rate-dist value V.S. summary word length. 94

5.5 Comparison of BNP Summarization with other systems using rate-dist measure. 96

5.6 The distribution of summary word length. 98

5.7 Compression ratio versus document word length. 99

Chapter 1

Introduction

Document summarization aims to generate a short text from one or more document(s), which conveys the most important information of the original text. With the rapid growth of documents on the Internet, summarization has proved to be an essential task in the area of web data mining. For example, it can be used for news services to compress a group of news articles to a short summary, helping readers to grasp the essential points in a short time.

Generally, document summarization can be categorized as abstraction-based or extraction-based. An abstraction-based summary can be seen as a reproduction of the original document(s) in a new way, while the extraction-based summarization focuses on extracting sentences directly from the original document(s). In this thesis, we consider generic extraction-based summarization for multiple documents.

Though there is no precise definition about what summary is a good summary, researchers usually follow some common standards [63, 45][1]:

- **Relevance**: A good summary should contain the most important information, i.e. the extracted sentences should be relevant to main topics of the original documents.

[1] Sometimes the names may be different, but the main idea is always same. For example, in [63] they considered length limitation instead of coverage.

- **Diversity**: The sentences in the summary should be non-redundant.
- **Coverage**: The summary should cover as more topics in the original documents as possible.

Early extractive summarization is based on some heuristic features of the sentences such as their positions in the text, the frequency of the words they contain, or some key phrases indicating the importance of the sentences [56]. More advanced techniques consider the rhetorical structure [61] and semantic relationships [31]. Researchers also leverage these features in some machine learning models [43, 102]. However, these techniques seem to ignore or belittle the redundancy and coverage of the summary. How to optimize all the three properties jointly remained a problem. A classic approach is the maximal marginal relevance (MMR) [10]. It introduced the MMR measure which combines query relevance and information novelty in topic-driven summarization, so the relevance and the redundancy are simultaneously considered in this model. Due to its simplicity, the MMR style algorithms are widely adopted in document summarization. However the greedy selection procedure in these algorithms makes them not effective for optimal content selection of the entire summary. A typical problematic scenario for greedy sentence selection is shown in [63]. The greedy selection procedure tends to select a long and highly relevant sentence first. This sentence is regarded as most informative, however, long sentences often contain not long relevant information but also some noise which is not relevant to the main story. As a result, the amount of space in the summary remaining for other sentences would be limited by including the noise in the selected sentence. As sentences in news are generally very long, it is very common to see the failure of greedy selection.

Instead of the sentence scoring (or ranking) and greedy selection approaches, we design an optimization framework which globally select the best overall summary by optimizing some proper objective functions. We only need to include the objectives and constraints of summarization in the objective functions, such as maximizing informativeness, minimizing repetition, and conforming to required summary length. Different from the simple linear combination of relevance and redundance constraints as in previous global inference algorithms [63] which hardly represent the coverage, we unify all three objectives of summarization in just one unit and attain an information-theoretic objective function. Our optimization framework is based on a novel perspective: data reconstruction. We assume that a good summary should reconstruct the original document as good as possible, for it should cover most of the important information in original documents. Based on this assumption, we develop several reconstruction models and generate summaries that has the least information "distortion". The reconstruction assumption is also adopted by others later [35], but their optimization method is quite different.

The advantage of the reconstruction-based summarization framework is then shown in solving a new problem called summary length determination. Generally, before a summarization system generates summaries, we have to know the required summary length. On the one hand, it facilitates efficient implementation and enables comparison of different systems. On the other hand, in some cases it is not reasonable to require all summaries to have the same length. Figure 1.1 shows a simple illustration of this idea. Summary 1 and Summary 2 are generated from documents focusing on the same event. However, obviously Summary 1 contains more opinions and it should be longer than Summary 2. Furthermore, even in a definite-length summarization system, how to define a proper length is also difficult. This thesis employs a Bayesian nonparametric method to solve this problem.

Bayesian Nonparametric models have been widely used in machine learning and data mining. It provides a Bayesian framework for model selection and adaptation, where the sizes of models are allowed to grow with data size. It is efficient to address the problem such as choosing the number of clusters mixture components or latent factors. Our reconstruction-based summarization framework is easily extended to

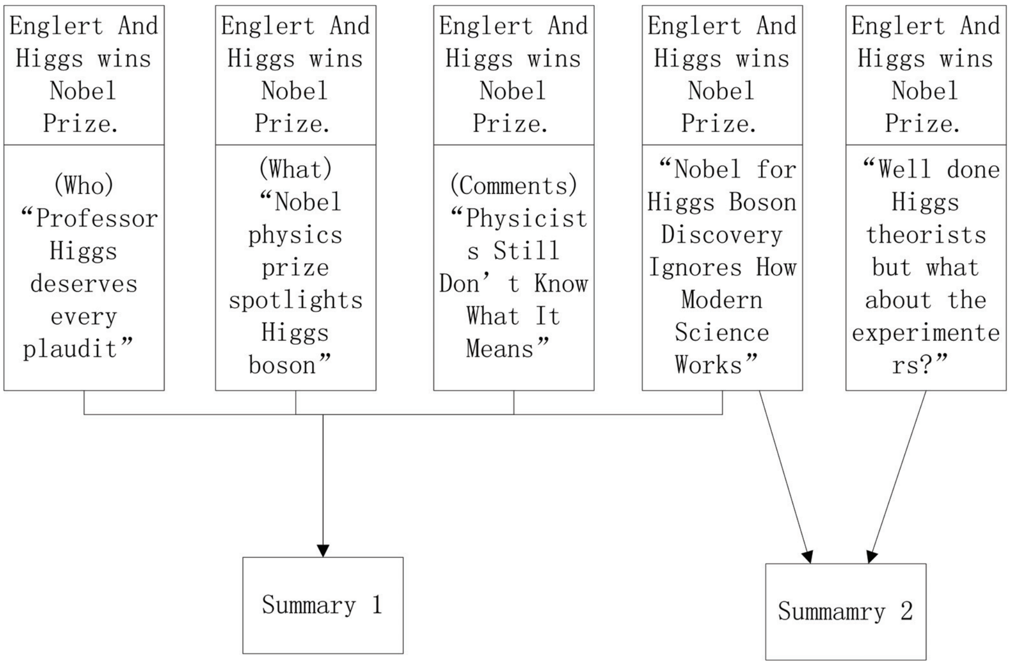

Englert And Higgs wins Nobel Prize.
(Who) “Professor Higgs deserves every plaudit”
Englert And Higgs wins Nobel Prize.
(What) “Nobel physics prize spotlights Higgs boson”
Englert And Higgs wins Nobel Prize.
(Comments) “Physicists Still Don’t Know What It Means”
Englert And Higgs wins Nobel Prize.
“Nobel for Higgs Boson Discovery Ignores How Modern Science Works”
Englert And Higgs wins Nobel Prize.
“Well done Higgs theorists but what about the experimenters?”
Summary 1
Summamry 2

Figure 1.1: An illustration of the summary length problem.

a Bayesian nonparametric model where we employ the Beta process for sentence selection. The new model could infer a proper number of summary sentences.

A problem in the Bayesian nonparametric document summarization is the sparsity of words. As summaries are very short compared to the original documents, lots of words are lost in summaries. If we use directly words frequencies or TF-IDFs as the representation of sentences, the reconstruction error must be large. It will largely impact the length determination. So topic models are utilized to represent the sentences and documents, in order to overcome the sparsity problem. In this process, we also considered an improvement to current topic models. We improved the popular Hierarchical Dirichlet Process based topic models (HDP-LDA) to deal with the situation that we do not know document category information. We propose a new Bayesian nonparametric prior for topic analysis, the hybrid nested hierarchical Dirichlet process (hNHDP). Other than improving the topic analysis results, the new model itself is an alternative to summarization. It provides a visualization of document structures and specific topics.

1.1 Thesis Contribution

The objective of this thesis is to address the problems in multi-document summarization. This leads to various reconstruction-based summarization models as explained before. The contributions of the thesis are summarized as follows:

- We proposed a novel optimization framework for multi-document summarization based on data reconstruction.

- We designed a new objective function for summarization: minimum distortion. Then we experimented with various distortion measures and compared them.

- We advanced the summary length problem, which has been rarely studied in document summarization.

- We extended the reconstruction-based framework for summarization to a Bayesian nonparametric system which determines proper summary length automatically.

- We present a new Bayesian nonparametric topic model to improve the current HDP-LDA, which plays an important role in Bayesian nonparametric document summarization.

1.2 Thesis Overview

The rest of the thesis is organized as follows:

- **Chapter 2**: In this chapter, we review the current research state of document summarization. We introduce the taxonomy, common approaches, evaluation etc.

- **Chapter 3**: This chapter describes the reconstruction-based optimization approach to summarization. The minimum distortion is proposed as the objective function. Then we design the p-median model, facility location model and linear representation model for ”reconstruction” and minimizing the distortion.

- **Chapter 4**: In this chapter we introduce the topic models and Bayesian nonparametric methods and their relationships with document summarization. The two techniques will be used in the next Chapter. We also review the Bayesian nonparametric topic models. Then we show our own contribution to this area where we propose a new model, hybrid nested Dirichet process for topic modeling.

- **Chapter 5**: This chapter addresses the problem of summary length determination. We integrate the Beta process into the reconstruction model, getting a Bayesian nonparametric model for summarization. This model could automatically determine the proper summary length. It uses the techniques explained in Chapter 4. It is also an extension to Chapter 3, for they are all based on data reconstruction.

- **Chapter 6**: This chapter summarizes the main contributions of this thesis.

Chapter 2

Document Summarization Overview

With the explosive growth of the internet, people are overwhelmed by the massive available online data. Document summarization, as an approach to solve the problem of information overload, has attracted a lot of interest in the area of natural language processing (NLP). Document summarization is the process of reducing text documents in order to create a brief summary that retains the most important information in the original texts. It could effectively save reading time as well as help users quickly find specific information.

According to the aim of document summarization, an ideal document summarization system should include but not limited to the following features:

- Conciseness. It is the most important feature of summaries. A good summary should be short to facilitate quick reading.

- Informativeness. Summaries should contain major points of the original documents.

- Good readability and clear structure.

Besides, the summaries should not be redundant or contain unrelated noise information. These features could also be summarized into three properties that we introduced before: **relevance**, **diversity** and **coverage**.

Document summarization has been studied over half a century. It has been addressed from many different perspectives, and many new types of summarization systems occurred. In the next section we will introduce the current state of summarization research, such as summarization types, real systems, relevant conferences.

2.1 Taxonomy of Summarization

Based on the number of original documents, summarization could be categorized into single-document and multi-document types. It is a popular but not the only kind of classification. Considering the output, we can divide summarization into extraction-based and abstraction-based types. According to the summarization method, there are supervised and unsupervised summarization . At last, the emergence of some new scenarios led to many new types of summarization (update summarization, opinion summarization etc.).

2.1.1 Single-document V.S. Multi-document

Generally, a summary can be produced from a single document or multiple documents. The former is called single-document summarization and the latter is multi-document summarization.

Research on document summarization can date back to 1950s [56] and has been greatly developed in recent years. Most early work focused on single-document summarization of technical documents. They measured the significance of sentences by features such as word frequency [56], sentence position [4] and the presence of cue words [21]. Then top ranking sentences are selected to form the auto-abstract.

Multi-document summarization [65] gained interests since mid 1990s, most applications being in the domain of news articles. Compared to single-document summarization, the advantage of multi-document summarization is that it could include different opinions from multiple perspectives. However it became more difficult and complex because of the thematic diversity within a large set of documents. In this thesis, we focus on the multi-document summarization.

2.1.2 Extraction-based V.S. Abstraction-based

Extraction-based (or extractive) summarization generates summaries by selecting salient sentences in original documents, while abstraction-based (or abstractive) summarization involves paraphrasing sections of the source document. Abstraction-based approaches could compress the original sentences [86], regenerate new sentences and re-ordering them [39].

Abstraction is conceptually better than extraction, for it allows to build more condensed and coherent texts. However, automatically generating texts is much more difficult, and the technique has not been mature enough. Nowadays the majority of summarization system remain extractive due to its feasibility. Recently, as the TAC workshops take more emphasis on the readability of summarization systems, and automatical linguistic evaluation methods has occurred [78], abstraction-based summarization would draw more and more attentions. In our work at TAC2010 [38], we also included simple sentence-editing methods and a new sentence ordering technique to improve readability.

2.1.3 Supervised V.S. Unsupervised

When machine learning approaches are used for summarization, we can classify the summarization approaches into supervised [43, 14] and unsupervised [10, 80]. It is easily understood that the supervised type contains training data while the unsupervised does not. The difficulty of supervised extractive summarization is the labeling of training data. The training data should be manually created by labeling sentences as "in summary" or "not in summary". However, this is not typically how people create summaries. So the natural summaries could not be directly used for training.

2.1.4 Generic V.S. Query-Focused

The generic summaries serve as surrogate of the original text and cover all aspects of the source text. Query-focused summaries [10], also called topic-focused summaries, focus on only the query or topic that users required. The query-focused summarization could also serve as a part of question-answering system.

2.1.5 New Types

In recent years, new types of summarization have appeared to meet the needs of various new scenarios. Update summaries [90, 16] concentrate on the novelty of the summaries. Users are assumed to already read some background information and they need novel information about the same event or topic. Opinion summarization [16], also called sentiment-based summarization, combines sentiment analysis with summarization regarding to the case that we are concerned about the opinions or reviews. Moreover, if we deal with texts in different languages, we may consider the multi-lingual [23] or cross-lingual summarization [96]. A multi-lingual system could deal with several languages, but the output summaries have always the same language as the input documents. The cross-lingual summarization corresponds to another case that input and output languages are different. For example, if we want to generate an English summary, but we only have Japanese documents or English translations of original Japanese documents, then the system is cross-lingual.

2.2 Common Approaches to Document Summarization

2.2.1 Sentence Scoring and Sentence Ranking

Most classic summarization systems are based on sentence scoring or sentence ranking, which are intuitive ways to obtain important sentences. Early systems determined the relevance of a sentence by means of word frequency [56] counts or only cue words/phrases [21]. Later *tf*idf* [62], mutual information [72] are used to improve the word frequency based method. Other common approaches include centroid based approaches, graph-based approaches, machine learning approaches. We will also introduce the topic model based sentence scoring in Chapter 4. .

Centroid-based Approaches

The centroid-based method [80] leverages the cluster centroids and it has been one of the most popular baselines for extractive summarization methods. A centroid is a pseudo center of a cluster of documents. It is defined as $c_j = \sum_{d \in C_j} d/|C_j|$, where C_j is the cluster of documents which describe the same topic, $|C_j|$ is the number of documents, and d is the $tf * idf$ representation of a document. The sentences that contain more words from the centroid of the cluster are considered as more salient. The MEAD toolkit[1] is an implementation of the centroid-based method that scores sentences based on sentence-level and inter-sentence features, including the cluster centroids, the position and length.

Graph-based Approaches

Graph-based ranking algorithms has been shown to be superior to centroid-based summarization. Usually a graph is constructed by establishing links between nodes

[1] `http://www.summarization.com/mead/`

(normally sentences or entities). The links are defined using similarities or other semantic relations. Sentences that are related to many other sentences are likely to be central and would have high weight for selection in the summary. Then, the system could normalize the weights of edges and calculate the sentence scores by performing a random walk on the graph (for example, in LexRank [22]). Incorporating syntactic and shallow semantic information in the graph building could improve the performance [13]. Furthermore, Wan et al. [97] developed an affinity graphs by differentiating intra-document and inter-document links between sentences, and finally penalizing redundant information.

Machine Learning based Summarization

Machine learning algorithms provide another way to score the sentences. A wide range of machine learning techniques have been applied to document summarization. The binary classifiers are studied in [43] which calculates the probability that a sentence is classified as a summary sentence. Hidden Markov Models are also connected with summarization by judging the likelihood that each sentence should be contained in the summary [14]. Neural networks [88], and support vector regression [74] are also used in summarization.

The advantage of using machine learning for document summarization is that it is of great freedom to incorporate all kinds of features, such as position, lexical, syntactic. It also allows to test the performance of the features and then selects the most suitable ones. However, many machine approaches need a big training corpus, which impedes the popularity. Labeling the corpus is very costly and utilizing the labels is also difficult because annotator agreement is often low.

2.2.2 Greedy Selection V.S. Optimization

Most summarization approaches choose content sentence by sentence. They sequentially select the most informative sentences after scoring or ranking the sentences. However, they have to check for redundance of the chosen sentences, and they could not guarantee the best coverage of important information. Global optimization approaches can be used to solve some new formulations of the summarization task, in which the best overall summary is selected.

A typical method using greedy selection is the Maximal Marginal Relevance (MMR) approach [30]. In this approach, the sentences are selected one by one to optimize a function which considers the relevance between sentence and queries (or original documents) as well as the redundance of the summaries.

The greedy selection approach is easy to be implemented and to be improved by modifying the optimization functions. However, the approach often result in bias in the selecting process as we introduced in the introduction. It could not effectively select the globally optimal summary. Optimization based algorithms, on the contrast, generate the overall best summaries. They could integrate all summarization aims or constraints in their objective functions and then select sentences together to optimize the function. Considering the features of a good summary, the objective function may represent the informativeness, redundance, and other special constraints (e.g. length limitation, query relevance). Exactly solving the global optimization is NP-hard [63]. However, global inference can be approximately solved by Linear Integer Programming [63] and dynamic programming [103] . Global optimization approaches to sentence selection have been shown to outperform greedy selection algorithms in several evaluations [83].

2.3 Evaluation

Accurately evaluating the quality of a summary spurs the improvement of summarization systems, so it has always been a critical task. Generally summary qualities could be evaluated from two aspects [40]. A general idea is to directly judge the linguistic quality and informativeness of the summaries. This approach is called intrinsic evaluation. The other approach is the extrinsic evaluation, where summaries are assessed by their helpfulness for a specific task.

2.3.1 Intrinsic Evaluation

Intrinsic evaluation usually compares summaries to some ideal reference data. Although it needs annotation of the corpora, it facilitates automatical evaluation and comparison of summarization systems. In the DUC and TAC evaluation workshops, summaries are evaluated mainly by intrinsic evaluation methods.

Human Evaluation

Early DUC conferences used the Summary Evaluation Environment (SEE) interface to manually compare peer summaries to the ideal. Assessors measured contents by marking all sharing units and rating the linguistic quality. Then a weighted score of the model units are defined and calculated to show the performance of all systems. For topic-focused summarization, the ”Responsiveness” metric is also used to reflect to what extent the summary satisfies the user’s information need.

In DUC 2001 to 2004, the manual evaluation was based on comparison with a single human-written model which may not cover all information. The pyramid method [71] addresses the problem by using multiple human summaries to create a gold-standard and by expoiting the frequency of information in the human summaries in order to assign importance to different facts. The pyramid gold-standard is based on a comparison between human-written summaries in terms of Summary Content Units (SCUs). The SCUs in peer summary are compared against an existing pyramid to evaluate how much information agrees between the peer summary and manual summary.

ROUGE

The advantage of human evaluation is its accuracy and comprehensive judgement (especially for the linguistic quality evaluation). However, it needs a lot of annotation, thus costly. Lin and Hovy developed the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [48, 46] for automatical evaluation, which has been used in later DUC conferences and most summarization work. The ROUGE-N measure is indeed an n-gram recall between a candidate summary and a set of reference summaries:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{RefSum\}} \sum_{n\text{-}gram \in S} Count_{match}(n\text{-}gram)}{\sum_{S \in \{RefSum\}} \sum_{n\text{-}gram \in S} Count(n\text{-}gram)}$$

where n stands for the length of the n-gram, and $Count_{match}(n\text{-}gram)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. $Count(n\text{-}gram)$ is the number of n-grams in the reference summaries.

ROUGE has been demonstrated a good automatic evaluation metric because it obtains good correlations with manual scores for content selection [54]. Within all ROUGE-N metrics, ROUGE-1 and ROUGE-2 are most popular. Moreover, ROUGE-L which measures the longest common subsequence and ROUGE-SU which measures the skip-bigram plus unigram-based co-occurrences are also widely applied.

Information-theoretic Measures

Most automatic evaluation measures are established based on the co-occurrence statistics to measure the content overlaps between system summaries and ideal summaries.

However, Lin et al [47] proposed a different approach form the information theoretic perspective. They introduced the new method based on the Jensen-Shannon divergence of distributions between automatic summary and reference summaries and achieved comparable performance with ROUGE.

A later extension is Louis and Nenkova's work [54] which still used the information theoretic measures but does not need to create human summaries. It is a big improvement because they only need to compare the summaries and the original documents. This idea is also employed in our thesis, both in Chapter 3 and Chapter 5.

2.3.2 Extrinsic Evaluation

Instead of direct analysis of the summaries, extrinsic evaluation assesses the impacts of summaries on other tasks, including categorization [60], information retrieval [81], and question answering [67]. For example, the SUMMAC evaluation [60] established large-scale, developer-independent evaluation of summarization systems in several relevance assessment tasks, such as document categorization. [64] designed a fact gathering task to demonstrate the helpfulness of news summaries generated by Newsblaster. Users are asked to answer related questions about an issue in the news and to generate reports by gathering facts from summaries or news articles. Then summaries are evaluated by the report scores and user satisfaction.

2.4 Real Systems

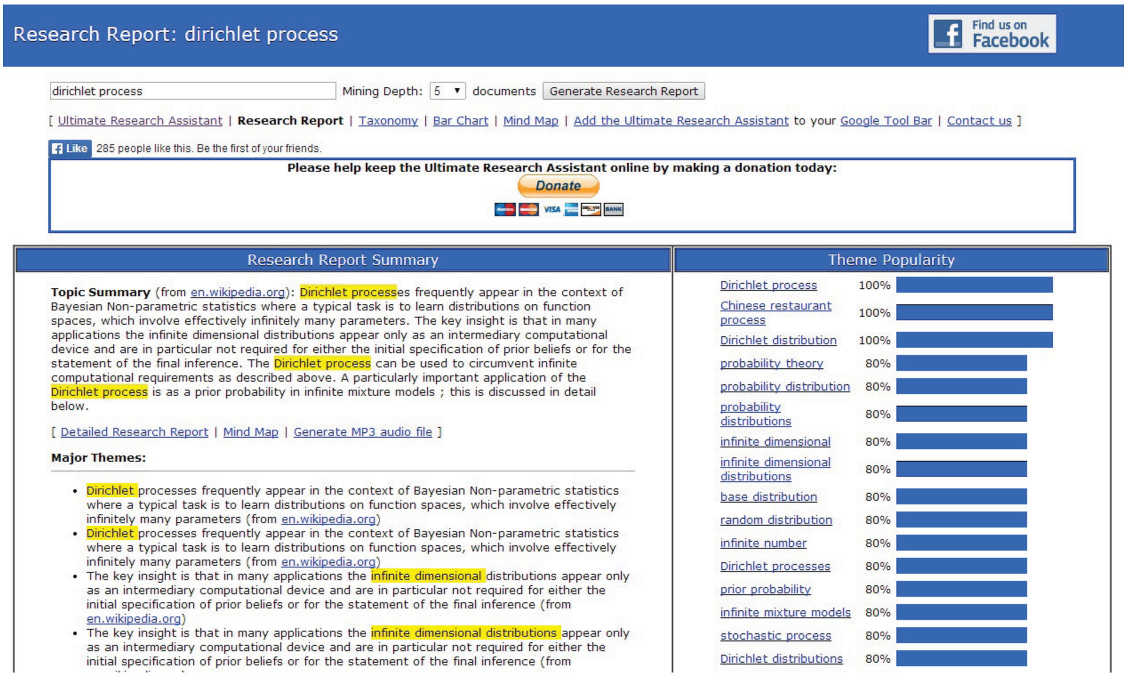
The summarization technology are coming into our life. More and more real-life summarization systems have been available in the domain of news articles and research papers. Some of them are listed as follows:

- **Ultimate Research Assistant (Figure 2.1)** The Ultimate Research Assistant is a research summarization system that combines information retrieval and text mining. It performs text mining (e.g. concept extraction, text summarization, visualization techniques) on search results of a research topic; and generates a concise research report summarizing the topic to help users perform online research.

- **iResearch Reporter (Figure 2.2)** iResearch Reporter is similar to the Ultimate Research Assistant but it is a commercial system. It could provide research report for individual professionals as well as content management solutions for business owners. It passes users' queries on to Google search engine, retrieves multiple relevant documents, and produces categorized, easily readable summary reports. Compared to the Ultimate Research Assistant, the summary is longer and contains more detailed information. The basis elements in the final report are snippets (text passages) which are derived from original documents and arranged meaningfully.

- **Newsblaster (Figure 2.2)** Newsblaster is a news summarization system developed by Columbia University. This system automatically collects, clusters and summarizes news from several web sites.

- **Yahoo! News Digest (Figure 2.4** Summarization has been available in mobile devices.Yahoo! News Digest, a mobile application which derives from the former Summly, helps people stay "quickly informed" on the day's big topics by sending out twice daily updates or digests. Basically it is a multi-document summarization system, using several sources to create news stories. The novelty is that it contains multi-modal contents, including videos, texts, maps and pictures. It also provide background information for the news at the end.



Research Report: dirichlet process
Find us on Facebook
dirichlet process
Mining Depth: 5 documents
Generate Research Report
[Ultimate Research Assistant | Research Report | Taxonomy | Bar Chart | Mind Map | Add the Ultimate Research Assistant to your Google Tool Bar | Contact us]
Like
Donate
Research Report Summary
Theme Popularity

285 people like this. Be the first of your friends.

Please help keep the Ultimate Research Assistant online by making a donation today:

Topic Summary (from en.wikipedia.org): Dirichlet processes frequently appear in the context of Bayesian Non-parametric statistics where a typical task is to learn distributions on function spaces, which involve effectively infinitely many parameters. The key insight is that in many applications the infinite dimensional distributions appear only as an intermediary computational device and are in particular not required for either the initial specification of prior beliefs or for the statement of the final inference. The Dirichlet process can be used to circumvent infinite computational requirements as described above. A particularly important application of the Dirichlet process is as a prior probability in infinite mixture models ; this is discussed in detail below.

[Detailed Research Report | Mind Map | Generate MP3 audio file]

Major Themes:

- Dirichlet processes frequently appear in the context of Bayesian Non-parametric statistics where a typical task is to learn distributions on function spaces, which involve effectively infinitely many parameters (from en.wikipedia.org)
- Dirichlet processes frequently appear in the context of Bayesian Non-parametric statistics where a typical task is to learn distributions on function spaces, which involve effectively infinitely many parameters (from en.wikipedia.org)
- The key insight is that in many applications the infinite dimensional distributions appear only as an intermediary computational device and are in particular not required for either the initial specification of prior beliefs or for the statement of the final inference (from en.wikipedia.org)
- The key insight is that in many applications the infinite dimensional distributions appear only as an intermediary computational device and are in particular not required for either the initial specification of prior beliefs or for the statement of the final inference (from

Dirichlet process	100%
Chinese restaurant process	100%
Dirichlet distribution	100%
probability theory	80%
probability distribution	80%
probability distributions	80%
infinite dimensional	80%
infinite dimensional distributions	80%
base distribution	80%
random distribution	80%
infinite number	80%
Dirichlet processes	80%
prior probability	80%
infinite mixture models	80%
stochastic process	80%
Dirichlet distributions	80%

Figure 2.1: Ultimate Research Assistant.

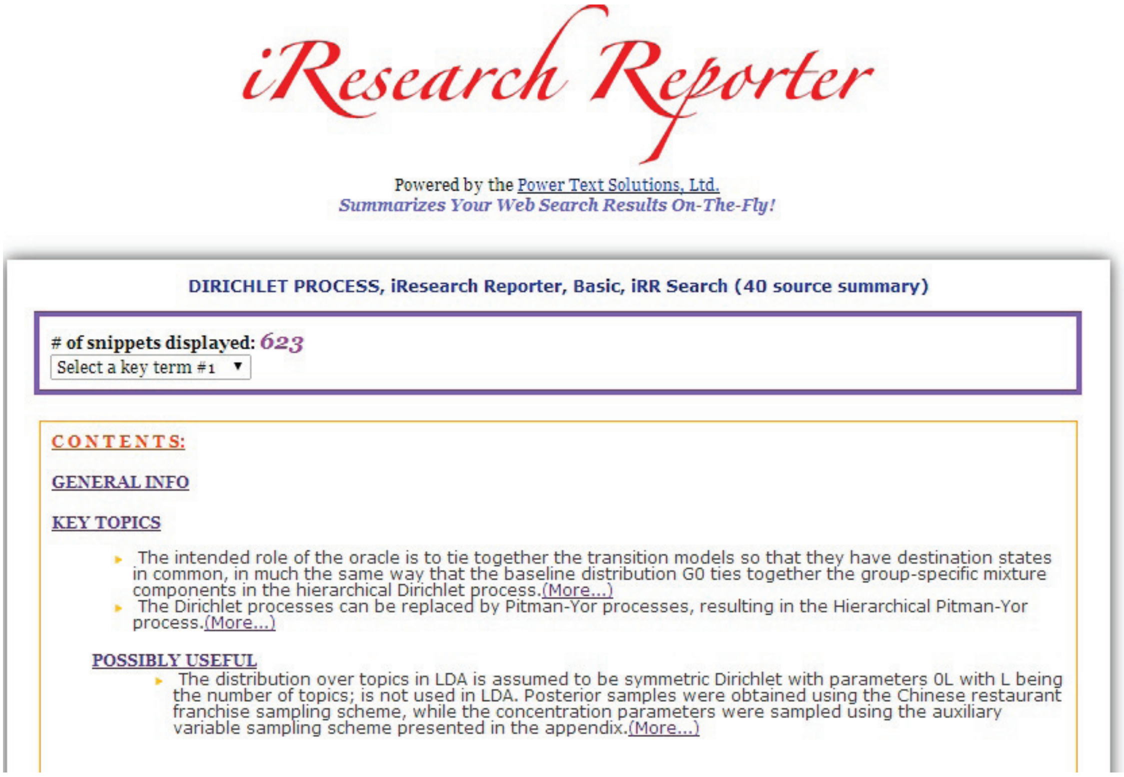
2.5 Relevant Evaluation Workshops

Evaluating the quality of a summary is a difficult but important task. The evaluation workshops for document summarization contributed a lot to the development of the technique by providing a platform to evaluate and compare summarization systems.

The TIPSTER Text Summarization Evaluation[2] (SUMMAC, 1998) is known as the first large-scale evaluation of automatic text summarization systems. In the conference, summaries are tested in categorization and question-answering tasks in order to analyze their informativeness.

From 2001 to 2003, the National Institute for Informatics Test Collection for IR (NTCIR) also involved the Automatic Text Summarization tasks. The aim is for researchers in this field to collect and share text data, and to make clear the issues of evaluation measures and methods for summarization of Japanese texts.

[2] `http://www-nlpir.nist.gov/related_projects/tipster_summac/`.



iResearch Reporter

Powered by the Power Text Solutions, Ltd.
Summarizes Your Web Search Results On-The-Fly!

DIRICHLET PROCESS, iResearch Reporter, Basic, iRR Search (40 source summary)

# of snippets displayed: 623
Select a key term #1 ▼

CONTENTS:

GENERAL INFO

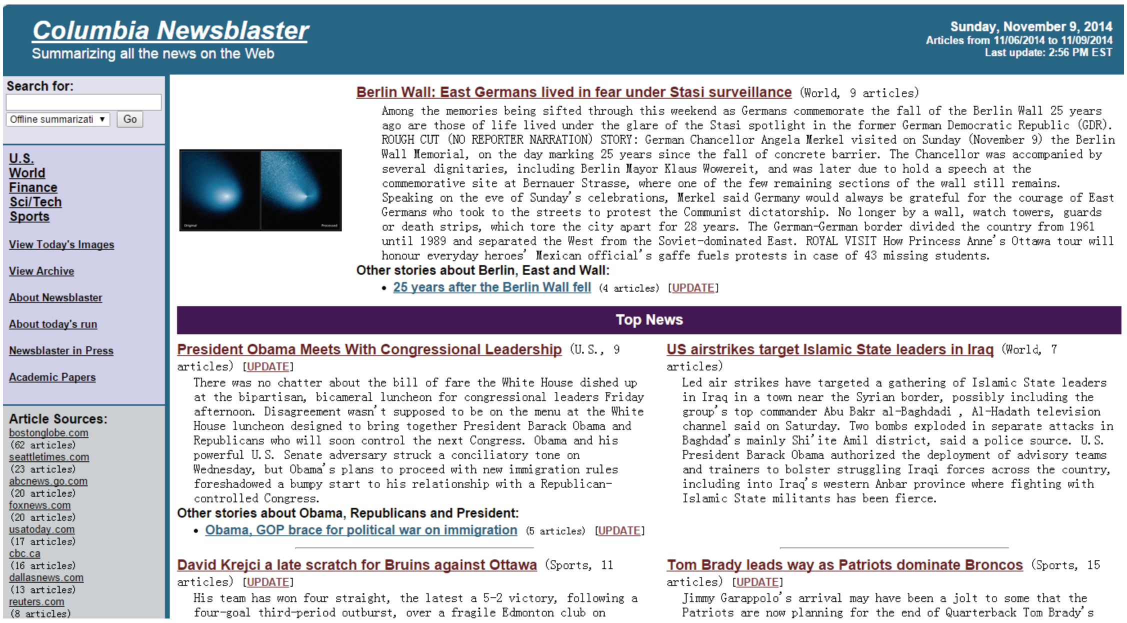
KEY TOPICS

- The intended role of the oracle is to tie together the transition models so that they have destination states in common, in much the same way that the baseline distribution G0 ties together the group-specific mixture components in the hierarchical Dirichlet process.(More...)
- The Dirichlet processes can be replaced by Pitman-Yor processes, resulting in the Hierarchical Pitman-Yor process.(More...)

POSSIBLY USEFUL

- The distribution over topics in LDA is assumed to be symmetric Dirichlet with parameters 0L with L being the number of topics; is not used in LDA. Posterior samples were obtained using the Chinese restaurant franchise sampling scheme, while the concentration parameters were sampled using the auxiliary variable sampling scheme presented in the appendix.(More...)

Figure 2.2: iResearch Reporter.



Columbia Newsblaster
Summarizing all the news on the Web

Sunday, November 9, 2014
Articles from 11/06/2014 to 11/09/2014
Last update: 2:56 PM EST

Search for:
Offline summarizati ▼ Go

U.S.
World
Finance
Sci/Tech
Sports

View Today's Images

View Archive

About Newsblaster

About today's run

Newsblaster in Press

Academic Papers

Article Sources:
bostonglobe.com (62 articles)
seattletimes.com (23 articles)
abcnews.go.com (20 articles)
foxnews.com (20 articles)
usatoday.com (17 articles)
cbc.ca (16 articles)
dallasnews.com (13 articles)
reuters.com (8 articles)

Berlin Wall: East Germans lived in fear under Stasi surveillance (World, 9 articles)

Among the memories being sifted through this weekend as Germans commemorate the fall of the Berlin Wall 25 years ago are those of life lived under the glare of the Stasi spotlight in the former German Democratic Republic (GDR). ROUGH CUT (NO REPORTER NARRATION) STORY: German Chancellor Angela Merkel visited on Sunday (November 9) the Berlin Wall Memorial, on the day marking 25 years since the fall of concrete barrier. The Chancellor was accompanied by several dignitaries, including Berlin Mayor Klaus Wowereit, and was later due to hold a speech at the commemorative site at Bernauer Strasse, where one of the few remaining sections of the wall still remains. Speaking on the eve of Sunday's celebrations, Merkel said Germany would always be grateful for the courage of East Germans who took to the streets to protest the Communist dictatorship. No longer by a wall, watch towers, guards or death strips, which tore the city apart for 28 years. The German-German border divided the country from 1961 until 1989 and separated the West from the Soviet-dominated East. ROYAL VISIT How Princess Anne's Ottawa tour will honour everyday heroes' Mexican official's gaffe fuels protests in case of 43 missing students.

Other stories about Berlin, East and Wall:
- 25 years after the Berlin Wall fell (4 articles) [UPDATE]

Top News

President Obama Meets With Congressional Leadership (U.S., 9 articles) [UPDATE]

There was no chatter about the bill of fare the White House dished up at the bipartisan, bicameral luncheon for congressional leaders Friday afternoon. Disagreement wasn't supposed to be on the menu at the White House luncheon designed to bring together President Barack Obama and Republicans who will soon control the next Congress. Obama and his powerful U.S. Senate adversary struck a conciliatory tone on Wednesday, but Obama's plans to proceed with new immigration rules foreshadowed a bumpy start to his relationship with a Republican-controlled Congress.

Other stories about Obama, Republicans and President:
- Obama, GOP brace for political war on immigration (5 articles) [UPDATE]

David Krejci a late scratch for Bruins against Ottawa (Sports, 11 articles) [UPDATE]

His team has won four straight, the latest a 5-2 victory, following a four-goal third-period outburst, over a fragile Edmonton club on

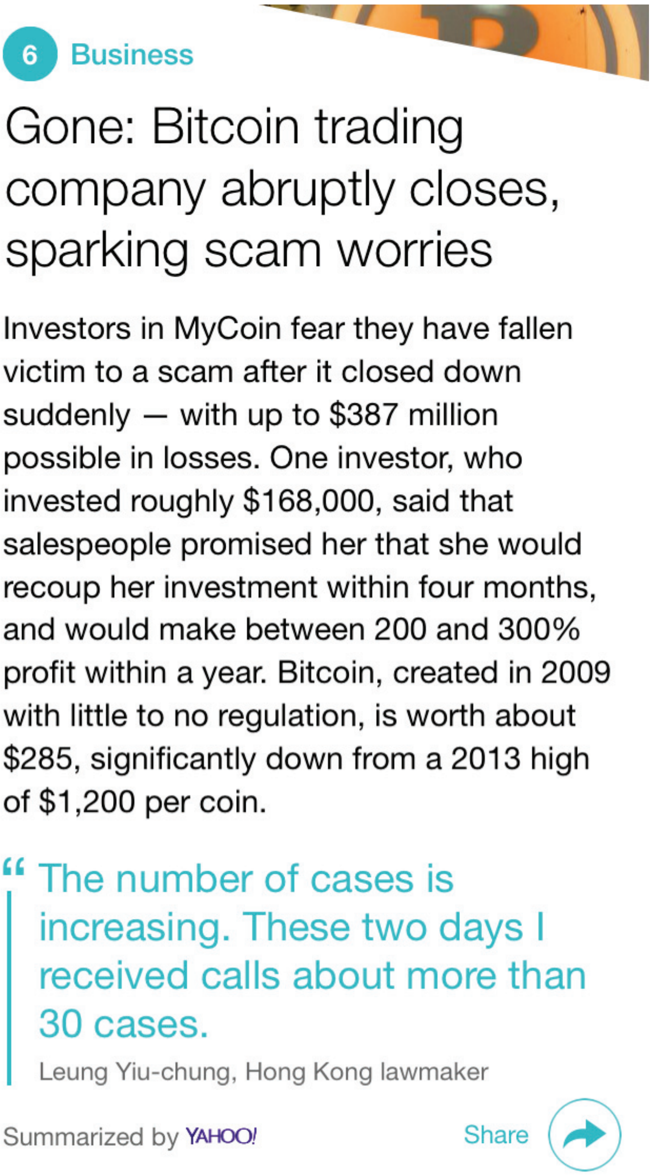
US airstrikes target Islamic State leaders in Iraq (World, 7 articles)

Led air strikes have targeted a gathering of Islamic State leaders in Iraq in a town near the Syrian border, possibly including the group's top commander Abu Bakr al-Baghdadi , Al-Hadath television channel said on Saturday. Two bombs exploded in separate attacks in Baghdad's mainly Shi'ite Amil district, said a police source. U.S. President Barack Obama authorized the deployment of advisory teams and trainers to bolster struggling Iraqi forces across the country, including into Iraq's western Anbar province where fighting with Islamic State militants has been fierce.

Tom Brady leads way as Patriots dominate Broncos (Sports, 15 articles) [UPDATE]

Jimmy Garappolo's arrival may have been a jolt to some that the Patriots are now planning for the end of Quarterback Tom Brady's

Figure 2.3: Newsblaster.



6 Business

Gone: Bitcoin trading company abruptly closes, sparking scam worries

Investors in MyCoin fear they have fallen victim to a scam after it closed down suddenly — with up to $387 million possible in losses. One investor, who invested roughly $168,000, said that salespeople promised her that she would recoup her investment within four months, and would make between 200 and 300% profit within a year. Bitcoin, created in 2009 with little to no regulation, is worth about $285, significantly down from a 2013 high of $1,200 per coin.

> “ The number of cases is increasing. These two days I received calls about more than 30 cases.
>
> Leung Yiu-chung, Hong Kong lawmaker

Summarized by YAHOO!

Figure 2.4: Yahoo! news digest.

The most famous competitions of document summarization are the series of summarization tasks in the Document Understanding Conferences (DUC) and the later Text Analysis Conferences (TAC). The DUC were held by the NIST yearly from 2001 to 2007 to progress in summarization and enable researchers participate in large-scale experiments. Every year different tasks were proposed, taking into account new challenges and requirements for document summarization. The conference provided standard data sets that are produced by experts, as well as the evaluation methods and tools. TAC can be regarded as the extension of the DUC. Initiated in 2008, TAC absorbed the DUC for text summarization and the Question-answering Track of the Text Retrieval Conference (TREC).

The tasks in DUC and TAC are changed over the years. In the early period, the conference focused on single-document summarization and generic multi-document summarization; and the data sets are collected from newswire/newspaper. To promote new research in summarization, some new challenges were proposed later, such as query-focused summarization, updated summarization, automatically evaluating summaries of peers (AESOP), and multi-lingual summarization. Besides, the data sets evolved from news to blogs and scientific articles; and the evaluation methods are also changed. In early DUC conferences, the summaries are evaluated manually using the SEE software. Then the automatic evaluation metrics are used, including ROUGE [46], Basic Elements (BE) [37] and Pyramid [71]. Recently automatic evaluation of summaries has even been a new track in the TAC conferences.

The tasks in all conferences are summarized in Table 2.1.

Conference	Summarization Task
SUMMAC	Single-document, query-focused, news
TSCb (NTCIR)	Query-focused, generic, news
TSC2 (NTCIR)	Single and multi-document, generic, news
TSC3 (NTCIR)	Multi-document, generic, news
DUC-01	Single and multi-document, generic, news
DUC-02	Single and multi-document, generic, news
DUC-03	Multi-document, query-focused, news
DUC-04	Single and multi-document, topic-oriented, news, cross-lingual
DUC-05	Multi-document, query-focused, news
DUC-06	Multi-document, query-focused, news
DUC-07	Multi-document, update, query-focused, news
TAC-08	Multi-document, update, query-focused, opinion, news & blogs
TAC-09	Multi-document, update, query-focused, news, evaluation
TAC-10	Multi-document, guided, news, evaluation
TAC-11	Multi-document, guided, multi-lingual, news, evaluation
TAC-14	Biomedical, scientific papers

Table 2.1: Summarization of all tasks in the evaluation conferences (mostly cited from [52]).

Chapter 3

Multi-document Summarization using Minimum Distortion

This chapter explains our reconstruction-based optimization approach to summarization [59]. We proposed a novel objective by borrowing a concept in information theory: distortion [15]. Our main idea is to use the distortion measures to take place of the three summary standards (relevance, diversity and coverage) which cannot be easily quantified integrally, and by minimizing the distortion our final summary can achieve the same or better effects.

We regard summarization as a data transmission system and assume that the output summary sentences represent the input document sentences. The distortion of the representation[1] is used as a measure to evaluate the summary quality. Based on different methods of the representation and the algorithms of minimizing the distortion, we propose three summarization models: p-median model, facility location model and linear representation model. First we adopt the one-to-one representation like the clustering technique (i.e. one original sentence is represented by one summary sentence). Under this assumption we get the first model - p-median model. Then the

[1] In this chapter, the ”representation” is equal to ”reconstruction”.

p-median model is improved by adding constraints or features and we propose the facility location model. Next, we jump out of the idea of clustering-like algorithms and replace the one-to-one representation with many-to-one representation (i.e. we use a linear combination of output sentences to represent one input sentence). Our final approach takes the linear representation model (many-to-many representation) combined with the facility location model, and the result exceeds most of popular summarization. In addition, we indicate that our final model can be extended to other summarization tasks.

3.1 Motivation and Problem Formulation

Given a set of sentences $\Omega = \{x_1, x_2, ...x_n\}$, where x_i denotes the ith sentence in the documents, the aim of extractive summarization is to select several representative sentences $S = \{\hat{x}_i\} \subset \Omega$.

An intuitive idea of selecting S is to rank the sentences in Ω by some measures and select the sentences with the highest ranks. The ranking system can easily integrate various features of the sentence, but it cannot sufficiently leverage the correlation with the original document(s) if we only consider the word occurrence information, for it calculates the similarity between only one sentence with the whole set. The coverage of the summary is hardly considered in the ranking model either. Other sequential selection (e.g. MMR) algorithms also cannot avoid the disadvantage.

How to develop a model and quantified measures to take advantage of the relevance, or on the contrary, the information loss, between the whole summary sentences and the whole set of original sentences is a problem deserved to be investigated. This problem also has some impact on recent evaluation metrics of the summarization [47].

As the above discussion refers to the concept of information loss, we develop an information-theoretic model, which sees the summarization as a data transmission system in Figure 3.1 [15]. Here the channel is omitted and we should only consider the information loss from the input to the output but ignore the middle process.

The sentences in Ω are represented as values of a variable , and sentences in S are seen as values of a variable . If the original documents are seen as an input of the variable X, the summary is the output of variable . Thus, in our approach the summary is a reconstruction of the original documents and every sentence in Ω can be represented by a new value $\hat{x_i} \in S$.

The representation function is defined as:

$$g : \Omega \to S.$$

Now we give a new function defined in the space $\Omega \times S$:

$$d : \Omega \times S \to \mathcal{R}^+.$$

It is called a distortion function in the information theory, and the distortion $d(x, \hat{x})$ is a measure of the cost of representing the sentence x as the sentence $\hat{x}$.To measure the sum of the cost, we use expectation of the distortion function:

$$Dis = \mathbb{E}d(X, g(X)) = \sum_{x \in \Omega} p(x) d(x, g(x)) \tag{3.1}$$

Using the rate distortion theory [15], the objective of the summarization model in Figure 3.1 can be elaborated by the Distortion Rate Function $D(R)$. When the rate R (i.e. the R in Figure 3.1, which can be thought according to the number of sentences in the summary in this case) is limited, our aim is to minimize the expectation of the distortion Dis in 3.1.

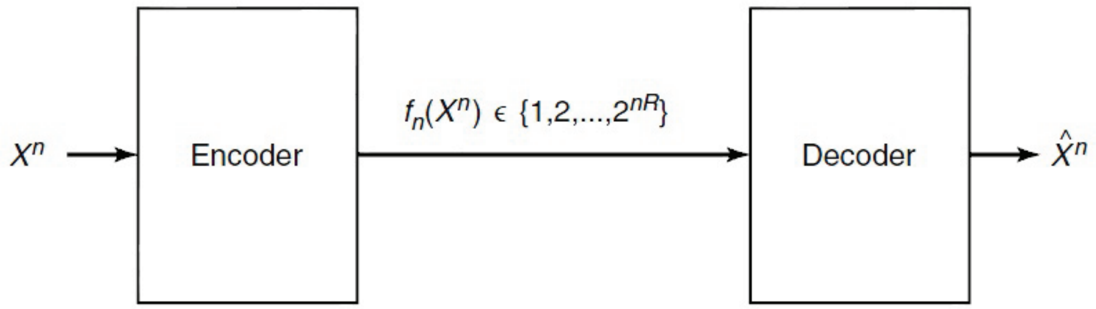


Figure 3.1: Transmission model.

3.2 Distortion Measures

While the distortion function is defined differently, the final output will be gained differently. Commonly a sentence X can be assumed as a memory-less source of words Y. This assumption makes the computation of the distortion more convenient. Actually, it is one of the reasons why we choose data transmission model with the distortion measure. Then the distortion between word sequences X^n and $\hat{X}^n$ can be extended to the following form:

$$d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(y_i, \hat{y}_i) \tag{3.2}$$

where y indicate a kind of value of a word in sentence x. In Hamming distortion y indicates the word itself, while in squared error distortion y indicates the frequency of the word. The following are some popular distortion measures.

3.2.1 Hamming Distortion

$$d(y, \hat{y}) = \begin{cases} 0; & y = \hat{y} \\ 1; & y \neq \hat{y} \end{cases} \tag{3.3}$$

Here y is the word itself. Using (3.3) in (3.2), we can see that this distortion mainly evaluate the number of common words between two sentences. In this case, the optimization of the summary can be intuitively explained as sharing the most words with the source without taking into account the weights of words.

3.2.2 Squared Error Distortion

$$d(y, \hat{y}) = (y - \hat{y})^2 \tag{3.4}$$

It is the most popular distortion measure used for continuous alphabets [15]. Although there are some disadvantages, it is widely used in image and speech coding. The distortion measure has many useful characteristics: non-negative, non-decreasing, symmetry.

If we see the sentences as points and the distortion as the distance, the optimization of (3.1) is equivalent to a p-median problem [3], which selects p centers from a set of data points so as to minimize the sum of the distances between each point to its nearest center . With respect to this assumption, we can use heuristic algorithms for p-median problem to determine which points can be chosen as the reconstruction points. The process will be elaborated in the next section.

[97] has employed the squared error distance to form the clusters , but it is based on the K-means method which is only used to form the clusters and calculate a virtual centroid, instead of real sentences in the original texts.

3.2.3 Information Divergence (KLD)

$$d(x, \hat{x}) = D_{KL}(p_x(y)||p_{\hat{x}}(y)) = \sum_y p_x(y) \log \frac{p_x(y)}{p_{\hat{x}}(y)} \tag{3.5}$$

Information divergence is also called K-L divergence (KLD), or relative entropy. It measures the expectation number of extra bits required to code when we use the distribution $p_{\hat{x}}(y)$ to replace $p_x(y)$. Every sentence x is seen as a memory-less source of words Y, and the summary is the corresponding output. It is a good measure to evaluate the degree of representation from the information-theoretic perspective. But it has a problem that it is not symmetrical, and it does meet the triangle relation. So it cannot be handled as same as the squared error distortion sometimes.

Besides, the $p_x(y) = p(y|x)$ here is finally replaced by $p(x, y)$ in our approach, because we want to add the distortion of each sentence and reflect the integral distortion between the summary and the source documents, and $p(x, y)$ is more useful to reflect the integral quality.

3.2.4 Jensen-Shannon Divergence (JSD)

As the K-L divergence is not symmetrical, the summary based on this measure usually get long sentences. However, the tasks of multi-document summarization are usually limited by the number of words instead of sentences. Thus long sentences may lead to a decrease of the total words in the summary.

One solution is adding length limits to the sentences when optimizing the expectation distortion; the other solution is to replace K-L divergence with J-S divergence, which is a symmetrical measure.

$$D_{JS}(P, Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{3.6}$$

where M is the average of the two distributions, $M = 1/2(P + Q)$.

3.2.5 Jensen-Shannon Divergence with Smoothing (JSDS)

The problem using K-L divergence is when an element in the distribution is zero. For example, if a word does not occur in $\hat{x}$ in 3.5, the $p_{\hat{x}(y)}$ will be zero, and the K-L divergence will be infinite. In this case, we should lead in some smoothing method to solve the problem. Bayes-smoothing [106, 47] is a widely used smoothing method in language models:

$$p = \frac{a_i}{a_0} \rightarrow p = \frac{a_i + \mu p(w_i|T)}{a_0 + \mu} \tag{3.7}$$

where μ is a scaling factor and $p(w_i|T)$ is the probability of word i occurring in topic T. J-S divergence can also be improved by the above smoothing method [47].

$$\begin{aligned} d(x, \hat{x}) &= D_{JSDS}(p_x(y)||p_{\hat{x}}(y)) \qquad (3.8) \\ &= \frac{1}{2}\sum_y (p(x,y)\log\frac{p(x,y)}{\frac{1}{2}(p(x,y)+p(\hat{x},y))} + p(\hat{x},y)\log\frac{p(\hat{x},y)}{\frac{1}{2}(p(x,y)+p(\hat{x},y))}) \end{aligned}$$

where $p(x, y) = \frac{OccurrenceInSentence(y)+\mu p(y|T)}{OccurenceInDocument(y)+\mu}$, and μ takes a value of 2000 following [47] and [106].

3.2.6 Other Distortion Measures

Some other distortion measures are also used for data compressing or clustering, such as various divergences introduced in [34]. And if we use the perspective of distortion, the information bottleneck method [94] adopts the loss of mutual information like a distortion measure. [34] demonstrates the rate distortion theory using information divergence distortion is equal to the information bottleneck method in clustering. However, in the document summarization, the two algorithms are not the same. A simple example is that when we represent all the source sentences using the sentence x with the highest $T(x, Y)$ as follows, the $I(\hat{X}, Y)$ will be the highest, but it is obviously not the best summary.

$$I(X, Y) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} \tag{3.9}$$

$$T(x, Y) = p(x) \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} \tag{3.10}$$

The reason is that it only respects the relation between the global information Y, but ignores the information loss between the sentences. In clustering the new representation is a class which contains the original sentence, but in summarization the representation is a new sentence which has nothing to do with the original sentence if we use the mutual information loss as the distortion.

3.3 Cluster-based Reconstruction Models

3.3.1 P-median Clustering Model

As discussed above, we use the data transmission model and the Distortion Rate Function 3.1 to solve the summarization problem. If the summary has a definite number (N) of sentences, there is no need to consider the rate region. So the optimization problem is as follows in 3.11.

$$\begin{aligned} \min_{S} Dis &= \mathbb{E}d(X, \hat{X}) \qquad (3.11) \\ &= \sum_{x \in \Omega} p(x) \sum_{\hat{x} \in S} p(\hat{x}|x) d(x, \hat{x}) \\ &= 1/N \sum_{\hat{x} \in S} \sum_{x \in \Omega} p(\hat{x}|x) d(x, \hat{x}) \\ &= 1/N \sum_{\hat{x} \in S} \sum_{x \in \mathcal{H}(\hat{x})} d(x, \hat{x}) \end{aligned}$$

Subject to:

$$
\begin{aligned}
p(\hat{x}|x) &\in \{0,1\}, \qquad (3.12)\\
S &\in \Omega,\\
|S| &\leq P,\\
\sum_{\hat{x}\in S} p(\hat{x}|x) &= 1.
\end{aligned}
$$

where $p(x) = 1/N$, P is the predefined number of summary sentences, and $\mathcal{H}(\hat{x})$ denotes the partition generated by $\hat{x}$. The problem can be solved by two heuristic approaches: the agglomerative approach and the interchange approach.

The agglomerative approach first assumes all the sentences in Ω are the representative sentences. Then one sentence is merged into a partition region in every step until the number of the sentences in the summary is N. The process is in fact a kind of hierarchical clustering, and this method can also serve as the base clustering method for traditional cluster-based summarization.

The interchange approach randomly chooses N sentences as the initial points and then starts an iteration process to replace the former point with a new point and gain a lower cost of the objective. In this approach, the problem is seen as a p-median problem, and the iteration process is local search [3]. According to [3], the worst case of this procedure has a cost $3 + 2/P$ times that of the global optimum.

Our final cluster-based model uses the result of the agglomerative approach as the initial point and then searches a local optimization result of the objective function. The process can be interpreted as Algorithm 3.3.1 and 3.3.1.

Algorithm 1 The Agglomerative Approach

```
a) Calculate all the distortions between every two sentence, and assign each sentence to a single cluster.
b) Agglomerate the two sentences with the least distortion and form a new cluster. The distance between the new cluster and another cluster is computed by the largest distortion between any two sentences of the two clusters.
c) Agglomerate the two clusters with the minimum distance and form a new cluster. Re-compute the distance between this cluster and other clusters.
d) Repeat c) until the shortest distance has reached the threshold.
e) Calculate the sum distortion of a sentence with others in a cluster. Choose the sentence with the least sum distortion as the deputy (centroid) of this cluster.
```

Algorithm 2 The Interchange Approach

```
f) Use the centroids computed by the agglomerative approach as initial points.
g) Assign each sentence to a centroid by choosing the least distortion.
h) Recalculate the centroids as e).
i) Repeat g) and h), until the centroids do not change any more.
```

3.3.2 Facility Location Model

In the former discussion, we can add length constraints when using K-L divergence as the distortion measure. Moreover, in the rate distortion model, if the rate is not a constant, according to the rate distortion theory, the objective function can be written as $I(X, \hat{X}) - \beta Dis$.

These problems provide a motivation of adding additional features into the summary cost. The above p-median problem is then converted to a facility location problem [3]. Take the length constraint as an example. If we choose a sentence into the summary, the cost has been changed to:

$$LengthPunish(\hat{x}) + \sum_{x \in \mathcal{H}(\hat{x})} d(x, \hat{x}); \tag{3.13}$$

Here we define the punishment function as

$$LengthPunish(x) = \begin{cases} \beta(Length(x) - max) & Length(x) \geq max; \\ 0 & min \leq Length(x) \leq max; \\ \beta(min - Length(x)) & Length(x) \leq min \end{cases} \tag{3.14}$$

The values of max and min depend on the datasets and are described in our experiments. Then the final objective function is changed as follows:

$$\begin{aligned} \min_{S} Dis &= \mathbb{E}(Cost(X) + d(X, \hat{X})) \qquad (3.15) \\ &= \sum_{x \in \Omega} p(x) \sum_{\hat{x} \in S} p(\hat{x}|x)(LengthPunish(\hat{x}) + d(x, \hat{x})) \\ &= \frac{1}{N} \sum_{\hat{x} \in S} \sum_{x \in \mathcal{H}(\hat{x})} (LengthPunish(\hat{x}) + d(x, \hat{x})) \end{aligned}$$

This idea complements the shortage of only using information distortion as the standard of summary selection, and it can integrate other features (such as features in the centroid-based method) or constraints in our model. The optimization algorithm is similar to p-median clustering, and we both use the simple local search method. Thus, our model gains a good extensibility without adding much complexity.

3.4 Linear Representation Model

3.4.1 Motivation

In the above cluster-based models, we assume every original sentence is represented by a new sentence. However, it is not an optimal representation. Intuitively, if a sentence is represented by more sentences instead of a single center, the information loss may be less. To formulate this idea, we use the linear combination of summary sentences to represent original sentences. Thus the distortion function is changed to:

$$d : \Omega \times \lambda(S) \rightarrow \mathbb{R}^+$$

where $\lambda(S)$ denotes the linear generative space of S.

In case of this assumption, the output of the transmission system in Fig. 1 is not changed. The change can be respected as only adopting a different transmission process. The expectation of distortion then is changed to:

$$Dis = \sum_{x \in \Omega} p(x_i) \min_{\{\hat{\lambda}_{ij}\}} d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) \tag{3.16}$$

where $\sum_j \hat{\lambda}_{ij} = 1$, and $\hat{\lambda}_{ij} \in \mathbb{R}^+$.

We prove that the reconstruction error of linear representation is the lower bound of one-to-one representation along the distortion measure of J-S divergence in Appendix A, i.e.

$$\min_{\{\hat{\lambda}_{ij}\}} d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) \leq d(x_i, \hat{x}_i). \tag{3.17}$$

So linear representation can be regarded as a better reconstruction model for summarization theoretically. However, 3.16 is hard to compute because we must calculate the set $\{\hat{\lambda}_{ij}\}$ for every sentence. Thus we consider a representation in document-level, i.e. we consider the whole input documents as a word source and the output summary is also a set of words by combining the sentences with different weights.

We expect that the distortion between the whole summary and the whole original documents is smaller than the sum of sentence-level distortion:

$$\min_S \min_{\{\lambda_i\}} \frac{1}{n} d(\sum_{x \in \Omega} x, \sum_{\hat{x}_i \in S} \lambda_i \hat{x}_i) \leq \min_S \sum_{x_i \in \Omega} \frac{1}{n} \min_{\{\hat{\lambda}_{ij}\}} d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) \tag{3.18}$$

where $\sum_i \lambda_i = n$, $\lambda_i \in \mathcal{R}^+$, n is the number of sentences in Ω.

This assumption is proved to be true in Appendix B. Thus we can directly calculate the distortion between the whole summary and the original documents without considering the representation of each sentence. In this way, our final objective function becomes:

$$Dis \propto \min_{S} \min_{\{\lambda_i\}} d(\sum_{x \in \Omega} x, \sum_{\hat{x}_i \in S} \lambda_i \hat{x}_i); \tag{3.19}$$

where $1/n$ is omitted.

3.4.2 The Approach of the Linear Representation Model

The Optimization Process

Using the above minimum objective 3.19, we develop an iterative algorithm based on an interchange process:

- a) Choose initial sentences. In our experiment, we adopt the former result of our cluster-based method (the interchange approach).

- b) Determine λ_i for corresponding $\hat{x}_i$.

$$Dis = \min_{\{\lambda_i\}} d(\sum_{x \in \Omega} x, \sum_{\hat{x}_j \in S} \lambda_{ij} \hat{x}_j)$$

- c) Remove the sentence $\hat{x}_t$ with the lowest λ_t. Add a sentence $\hat{x}_j$ to guarantee that Dis_{new} has the largest decrease.

$$Dis_{new} = d(\sum_{x \in \Omega} x, \lambda_t \hat{x}_j + \sum_{\hat{x}_i \in S, \hat{x}_i \neq \hat{x}_t} \lambda_i \hat{x}_i)$$

- d) Repeat Step b) and c) until the summary set does not change any more.

The Algorithm of Assigning λ_i to $\hat{x}_i$.

We use a gradient algorithm to assign λ_i and the proof is given in Appendix A.

- a) When the initial summary sentences are given, calculate the distortion (or with the punishment weight together) between each sentence x and the summary sentences.
- b) A sentence x is assigned to the region of the summary sentence $\hat{x}_i$, if $d(x, \hat{x}_i) = \min_{\hat{x}_j \in S} d(x, \hat{x}_j)$.
- c) Calculate the size of the $\hat{x}_i$ region, i.e. the number of sentences assigned to $\hat{x}_i$. Then the size is assigned to λ_i as its initial value.
- d) Calculate each value of $g_i = \partial d(\sum_{x \in \Omega} x, \sum_{\hat{x}_i \in S} \lambda_i \hat{x}_i)/\partial \lambda_i$.
- e) Find the largest gradient g_i and the smallest one g_j.

  $$\begin{aligned} g_i &= g_i - \delta h; \text{ if } g_i > 0. \\ g_j &= g_j + \delta h; \text{ if } g_j < n. \end{aligned}$$

  where $\delta h > 0$ and n is the number of sentences in Ω.
- f) Repeat Step d) and e) until $d(\sum_{x \in \Omega} x, \sum_{\hat{x}_i \in S} \lambda_i \hat{x}_i)$ becomes constant or larger than the last step.

In our following experiments, we take $\delta h = 0.5$. It is a tradeoff between accuracy and computational complexity.

Comparison with Soft Partition

Someone may notice that in our initial p-median model, the partition is hard, i.e. $p(\hat{x}|x) \in \{0, 1\}$(see 3.11). The model can be improved by using a soft partition (soft clustering method). In this new model, we assume $p(\hat{x}|x) \in [0, 1]$. Thus every sentence can be represented by several sentences with a serial of probabilities:

$$Dis = \sum_{x\in\Omega} p(x) \sum_{\hat{x}\in S} p(\hat{x}|x) d(x,\hat{x}) = \sum_{x\in\Omega} p(x) \sum_{\hat{x}_i\in S} \lambda_i d(x,\hat{x}_i); \tag{3.20}$$

where $0 \leq \lambda_i = p(\hat{x}_i|x) \leq 1$.

Intuitively, soft partition and the linear representation have similar effects. Now we compare the two ideas. As J-S divergence has similar characteristics with K-L divergence in these inequalities, we need only to take K-L divergence as the example. 3.20 indicates that our linear representation model can attain a smaller distortion than soft partition.

$$\begin{aligned} \sum_{\hat{x}\in S} p(\hat{x}|x) d(x,\hat{x}) &= \sum_{\hat{x}\in S} p(\hat{x}|x) \sum_{y} p(x,y) \log \frac{p(x,y)}{p(\hat{x},y)} \\ &= \sum_{y} \sum_{\hat{x}\in S} p(\hat{x}|x) p(x,y) \log \frac{p(\hat{x}|x)p(x,y)}{p(\hat{x}|x)p(\hat{x},y)} \\ &\geq \sum_{y} (\sum_{\hat{x}\in S} p(\hat{x}|x)p(x,y)) \log \frac{sum_{\hat{x}\in S} p(\hat{x}|x)p(x,y)}{sum_{\hat{x}\in S} p(\hat{x}|x)p(\hat{x},y)} \\ &= \sum_{y} p(x,y) \log \frac{p(x,y)}{\sum_{\hat{x}_i\in S} \lambda_i \hat{x}_i} = d(x, \sum_{\hat{x}_i\in S} \lambda_i \hat{x}_i); \end{aligned} \tag{3.21}$$

where $\sum_{\hat{x}\in S} p(\hat{x}|x) = 1; \lambda_i = p(\hat{x}_i|x)$.

3.5 Experiments

3.5.1 Data Sets

Document Understanding Conference (DUC) has organized yearly evaluation of document summarization. Generic multi-document summarization is one of the fundamental tasks in DUC2002 and DUC2004. In DUC 2002, 59 document sets of approximately 10 documents each were provided and generic summaries of each document set with lengths of approximately 100 words or less were required to be created. In DUC 2004, 50 document clusters were provided and a short summary with lengths of 665 bytes or less was required to be created.

3.5.2 Evaluation Metric

We use the ROUGE [46] evaluation toolkit[2], which is adopted by DUC for automatically summarization evaluation. It measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. ROUGE-N is an n-gram recall measure which is computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S\in\{RefSum\}}\sum_{n\text{-}gram\in S} Count_{match}(n\text{-}gram)}{\sum_{S\in\{RefSum\}}\sum_{n\text{-}gram\in S} Count(n\text{-}gram)}$$

where n stands for the length of the n-gram, and $Count_{match}(n\text{-}gram)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. $Count(n\text{-}gram)$ is the number of n-grams in the reference summaries.

According to [46], among the evaluation methods implemented in ROUGE, ROUGE-N (N=1, 2) is relatively simple and works well in most cases. In our work we employ ROUGE-1 and ROUGE-2 to score the summaries.

3.5.3 Experimental Results

We evaluate all the proposed models with different distortion measures on the DUC2004 dataset, and the results on the DUC2002 dataset further show the improvement of the models. Table 3.1 and Table 3.2 list the comparison results generated by our models. As the Agglomerative approach of the p-median model is most simple and its result can provide initial sentences for other approaches, we take

[2] We use ROUGEeval-1.4.2 downloaded from `http://www.haydn.isi.edu/ROUGE/`

it as the baseline. The Interchange approach (p-median) is then used to improve the Agglomerative approach. In these two approaches, four distortion measures are employed on the DUC2004 dataset: Hamming, Squared Error, K-L divergence with smoothing (KLDS) and J-S divergence with smoothing (JSDS). In the agglomerative process, the cluster threshold is empirically set to 0.9 for Hamming and Squared Error distortion; and when two sentences have less than two common words, we assign the KLDS and JSDS with a large value (1.0) and stop the clustering process according to this value. The results in Table 3.1 show that JSDS is the best measure in the interchange approach, while in the agglomerative approach different distortion measures achieve similar results. On DUC2002 data, we do not test all distortion measures but only use the best measure-JSDS to demonstrate the effectiveness of the improved models.

To limit the lengths of summary sentences when using KLDS, we add the length punishment function and solve the optimization problem using the facility location model. We also tried this model with the JSDS measure. On the DUC2004 dataset, we punish sentences whose lengths are more than 100 bytes or less than 50 bytes. And on the DUC2002 dataset, we assume the length of a good sentence is between 7 and 20 words. In the listed results, we find that in most cases the length constraint leads to performance improvement. The model can be further extended by adding more features like the positions and structure features; however, we do not investigate other features in this work. Our main aim here is to demonstrate the extendibility of our model. The final improvement comes from the usage of linear representation. As our method usually gains a local optimization, the selection of the initial sentences is crucial and it can greatly impact the final result. Fortunately, we always obtain performance improvement when using results of former runs (the interchange approach of the p-median model and the facility location model) as the initial sentences. We do not conduct experiments using distortion measures other than KLDS and JSDS

		DUC2004 Task 2	
		ROUGE-1	ROUGE-2
P-Median Model (Agglomerative)	Hamming	0.36756	0.07755
	Squared Error	0.36703	0.07813
	KLDS	0.36583	0.07571
	JSDS	0.36599	0.07495
P-Median Model (Interchange)	Hamming	0.37413	0.07845
	Squared Error	0.36791	0.07823
	KLDS	0.36791	0.07823
	JSDS	0.38235	0.08364
Facility Location Model (Length Punishment)	KLDS	0.37208	0.07868
	JSDS	0.38429	0.09107
Linear Representation(LR) Model	KLDS	0.38095	0.08262
	JSDS	0.38599	0.08345
LR Model with Length Punishment	KLDS	0.37996	0.07749
	JSDS	0.39614	0.09179

Table 3.1: Experimental Results on DUC2004 Data.

		DUC2002 Task 2	
		ROUGE-1	ROUGE-2
P-Median Model (Agglomerative)	JSDS	0.33923	0.07224
P-Median Model (Interchange)	JSDS	0.34625	0.07262
Facility Location Model (Length Punish)	JSDS	0.35262	0.07418
Linear Representaion(LR) Model	JSDS	0.35021	0.07673
LR Model with Length Punishment	JSDS	0.35884	0.07752

Table 3.2: Experimental Results on DUC2002 Data.

at this step. The linear representation model with length punishment method (i.e. we use the result of facility location model to initiate the linear representation model and add length punishment to the summary sentences.) achieves the best performance on both DUC2004 and DUC2002. This indicates the effectiveness of the two techniques, and also proves the distortion is a good standard to estimate the quality of summaries. We also compare our results with some other popular models in Table 3.3 and Table 3.4. Except for MMR [10] and KLSum [33], all the results of these models are cited from their original papers which maybe experiment only on one of

	ROUGE-1	95% confidence	ROUGE-2
Best Human	0.41828	0.40193 - 0.43463	0.10500
Worst Human	0.38902	0.36793 - 0.41011	0.08595
Team65	0.38232	0.37034 - 0.39278	0.09219
Team104	0.37436	0.36502 - 0.38568	0.08544
Team35	0.37427	0.36074 - 0.38664	0.08364
Our best model	0.39614	0.38244 - 0.41220	0.09179
Our Interchange Approach (JSDS)	0.38235	0.37028 - 0.39744	0.08364
Centroid	0.3670	0.3580-0.3767	-
Cont. LexRank	0.3758	0.3617-0.3826	-
Semi-supervised	0.329	-	0.073
MMR	0.34923	0.33740 - 0.36617	0.08010
KLSum	0.23422	0.22149 - 0.24697	0.01716

Table 3.3: Comparison with Other Models on DUC2004.

our datasets, finally we have different control groups on DUC2002 and DUC2004, and - indicates there is no reported score in this term. First, we list the best performance values of the DUC2002 and DUC2004 participants. Moreover, on DUC2004, the human summaries are also evaluated and the official ROUGE scores are given. We cite the Centroid and LexRank results provided by [22] as well the MMR and KLSum results which are generated by ourselves, in order to show the performance of the traditional greedy selection models. The language independent graph-based model (Pagerank-U) [66], the dynamic programming (Knapsack), the integer programming (ILP) [63], the semi-supervised model [100] are also included, and we use the toolkit of Information Distance [53] to experiment too. From the results in Table 3.3 and Table 3.4, we can see that our final approach (Linear Representaion Model with Length Punishment) exceeds most of popular models and the participating systems. Especially, we have achieved a result close to the human-annotated result on the DUC2004 dataset. The result of our approach is better than greedy selection based models (e.g. centroid, LexRank, MMR, KLSum). It shows that the traditional selection methods are not good enough and our optimization approach is a better choice, for our method conveys more integral information from the perspective of information theory. The

	ROUGE-1	ROUGE-2
Our best Model	0.35884	0.07752
Team26	0.35151	0.07642
Team19	0.34504	0.07936
Team28	0.34355	0.07521
Pagerank-U	0.3552	-
Information Distance	0.29216	0.05478
Knapsack	0.348	0.073
ILP	0.346	0.072
MMR	0.29519	0.05781
KLSum	0.19513	0.01230

Table 3.4: Comparison with Other Models on DUC2002.

information distance model is not very effective on the DUC2002 dataset, the reason may be that it is a model which is more suitable for topic-focused summarization. Furthermore, our model also beats other optimization (global inference) models (i.e. Knapsack and ILP). It demonstrates the superiority of our model by unifying all aims, especially the coverage, in one information-theoretic objective function.

3.6 Conclusions and Problems

In this chapter, three new summarization models are proposed based on the reconstruction assumption and they optimize an information theoretic measure: distortion. The p-median model respects the optimization as a p-median problem and conveys as more information between the whole summary and the whole original documents as possible. The facility location model adds features to the p-median model, and the linear representation model jumps out of the idea of clustering, and directly evaluate the distortion between the whole documents and all candidate sentences. The models have been demonstrated effective on the DUC2002 and DUC2004 datasets. However, there remained several problems as follows:

1. In almost all previous summarization systems, including the models proposed in this chapter, the summary length should be predefined. Summary length is an important factor in summarization, it indicates the trade-off between concision and completeness. Moreover, when quantity of information in two document sources differentiate too much, it is not reasonable to require their summaries have the same length. So, how to find a proper summary length should be considered. In Chapter 5, we explain the problem of summary length determination in detail, and extend the optimization based summarization models to be Bayesian nonparametric in order to solve the problem.

2. In this chapter, we use word vectors to represent the documents and sentences. It remains the original word information, but it ignores the relationships between words. Moreover, in our reconstruction models, the reconstruction error would be large due to word sparseness. In Chapter 4, we introduce the topic models for summarization and use them to avoid sparseness for the new model in Chapter 5.

3. The extractive summarization enables acceptable readability. However, it lacks information about the document or topic structures, especially in documents with clear categories or hierarchies. In Chapter 4, we also present an alternative method to extractive summarization. We improved a classical Bayesian nonparametric topic model, HDP-LDA, and propose a new model which could discover latent document and topic structures. The visualization of the topics and document structures provides a new perspective to understand the documents. The new topic model should also be seen as an improvement to the preprocessing step of traditional document summarization.

Chapter 4

Topic Models and Bayesian Nonparametrics

This chapter describes the fundamentals of topics models and Bayesian nonparametric methods, their relationship with document summarization. The two techniques provide the background of the next chapter, and they will be utilized to improve document summarization. First we briefly introduce the basic ideas and some typical models and applications separately. Then we discuss the connections between them and document summarization.

In addition, at the end of this chapter we introduce our new Bayesian nonparametric topic model. We integrate the advantage of both the hierarchical Dirichlet proces (HDP)s and the nested Dirichlet process (NDP), attaining the hybrid nested hierarchial Dirichlet process (hNHDP).

4.1 Probabilistic Topic Models

Topic models [6] are an increasingly useful family of algorithms for statistical analysis of document collections as well as other discrete data, such as genomic data [26] and discrete image data [24]. The aim of topic models is to uncover the latent thematic structure in documents (or other similar data). With topic models, we could better understand the documents, and easily browse, search or organize the information. They have various applications in machine learning, including information retrieval, collaborative filtering, and image classification.

The fundamental idea of topic models is to assume that each document is a mixture of latent topics, each of which is a probabilistic distribution over words. To generate a document, a distribution of topics is firstly drawn; then each word in the document is assigned randomly a topic and drawn according to the probability distribution associated with the topic. To better illustrate the models, we define the following notations. Let θ be the document-specific topic distribution, z be the assigned topic for each word w, ϕ_z be the word distribution associated with topic z, and F be the multinomial distribution which select a word from topic z. Then a topic model can be represented as a mixture model.

$$w|\Phi, z \sim F(\phi_z) \tag{4.1}$$

$$z|\theta \sim \theta$$

$$\tag{4.2}$$

4.1.1 Probabilistic Latent Semantic Indexing

Probabilistic Latent Semantic Indexing (PLSI) [36] is an early topic model (sometimes it is also called an aspect model). It follows the bag-of-words assumption that ignores the order of words. It introduces the concept of latent topic, and assumes that a document d and a word w are conditionally independent given a topic z.

$$p(d, w) = p(d) \sum_z p(w|z)p(z|d) \tag{4.3}$$

4.1.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [9] is a fully Bayesian extension of the PLSI model. The PLSI model does not make any assumption about how the topic distributions (θ) are generated. This makes it hard to be generalized to new documents. In constrast, the LDA model puts a Dirichlet prior on the topic distributions, i.e. $\theta \sim Dir(\alpha)$ where α is the hyperparameter of the Dirichlet Distribution.

4.1.3 Other Topic Models

Since the LDA, there have been a variety of topic models applying to many different situations. For example, the dynamic topic models [8] could catch the topic variance over time; the author-topic model [87] is developed to consider the author information; the multi-grain topic model [95] is present to extract comment aspects of objects in online reviews.

4.2 Topic Models for Document Summarization

In document summarization, topic models can be used for document representation [99, 33], and they can be also directly used for some special summarization tasks [18, 91]. Summarization benefits from topic model representations, which reduce the dimensionality and capture implicit semantic relations. Compared to word representations, topic representations enable better sentence scoring and sentence similarity calculation. For example, Haghighi and Vanderwende [33] show that the TopicSum method which uses topic model based representations performs much better than the SumBasic method which uses word representations. There are also special topic models developed for summarization to get the document structure(HIERSUM [33], DualSum [18], HybHSum [11]) or topical coherence [12].

4.3 Bayesian Nonparametric Methods

Bayesian nonparametric methods provide a Bayesian framework for model selection and adaptation using nonparametric models [27]. A BNP model uses an infinite-dimensional parameter space, but invokes only a finite subset of the available parameters on any given finite data set. This subset generally grows with the data set. Thus BNP models address the problem of choosing the number of mixture components or latent factors. For example, the hierarchical Dirichlet process (HDP) [92] can be used to infer the number of topics in topic models or the number of states in the infinite Hidden Markov model.

4.3.1 Dirichlet Process Mixture Models

The Dirichlet Process

The Dirichlet process is one of the best known Bayesian nonparametric priors. It has been widely used in machine learning due to its computational efficiency [25]. A Dirichlet process (DP) is a distribution over probability distributions. Given a probability measure G_0 on a measurable space $(\Theta, \mathcal{B})$, if we say G is distributed according to a DP with parameters α, H, i.e. $G \sim \mathrm{DP}(\alpha, H)$, it means the following:

$$(G(A_1), G(A_2), ..., G(A_K)) \sim Dirichlet(\alpha H(A_1), \alpha H(A_2), ..., \alpha H(A_K)) \tag{4.4}$$

for any finite partition $(A_1, A_2, ..., A_K)$ of the space $(\Theta, \mathcal{B})$.

Dirichlet Process Mixture Models

Since the probability distributions drawn from a DP are discrete, the DP related processes cannot be directly used for density estimation. Instead, they are used as a prior at the top of hierarchical models, which yields the Dirichlet mixture model

(DPM) [2]. Let w_i be an observation with a distribution $F(\theta_i)$ given factor θ_i that is *i.i.d.* drawn from a random probability measure G. Given θ_i, the observations are conditionally independent to each other. If G is Dirichlet process distributed, we can then derive the DPM as

$$\begin{aligned} w_i &\sim F(\theta_i) \text{ for } i = 1; 2; \ldots ; n \\ \theta_i &\sim G \text{ for } i = 1; 2; \ldots ; n \\ G &\sim \text{DP}(\alpha; H) : \end{aligned}$$

With respect to Dirichlet-multinomial conjugacy, $F(.)$ is usually set to be a multinomial distribution in real applications, for example, the probabilistic topic models.

4.3.2 The Hierarchical Dirichlet Process

The HDP [92] is a Bayesian nonparametric prior for modeling groups of data. It ensures that sets of group-specific DPs share the atoms. Suppose that we have observations organized into groups. Let x_{ji} denote the i^{th} observation in group j. All the observations are assumed to be exchangeable both within each group and across groups, and each observation is assumed to be independently drawn from a mixture model. Let $F(\theta_{ji})$ denote the distribution of x_{ji} with the parameter θ_{ji}, which is drawn from a group-specific prior distribution G_j. For each group j, the G_j is drawn independently from a DP, $DP(\alpha_0, G_0)$. To share the atoms between groups, the HDP model forces G_0 to be discrete by defining G_0 itself as a draw from another DP,

$DP(\gamma, H)$. The generative process for HDP is represented as:

$$
\begin{aligned}
G_0 &\sim \text{DP}(\gamma, H), \\
G_j &\sim \text{DP}(\alpha_0, G_0) \text{ for each j}, \\
\theta_{ji} &\sim G_j \text{ for each j and i}, \\
x_{ji} &\sim F(\theta_{ji}) \text{ for each j and i}. \qquad (4.5)
\end{aligned}
$$

Using the stick-breaking construction of Dirichlet processes, we can express G_0 as $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$, where δ_{ϕ_k} is a probability measure concentrated at the atom ϕ_k. The atoms are drawn from the base measure H independently, and the weights $\boldsymbol{\beta} \sim$ GEM(γ)[1] are mutually independent. Because G_0 has support at the points $\{\phi_k\}$, each G_j necessarily has support at these points as well; and can thus be written as $G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$, where the weights $\boldsymbol{\pi}_j = (\pi_{jk})_{k=1}^{\infty} \sim \text{DP}(\alpha_0, \boldsymbol{\beta})$.

When the data groups are categorized into higher-level categories, we should extend HDP to the third level. For example, let us consider documents from different corpora. In this case, a document is a group and a corpus is a category. A top-level DP generates the base measure for each corpus; draws from each of these corpus-level DPs yield the base measures for DPs associated with the documents within a corpus. Finally, draws from the document-level DPs provide the topic representation of each document (a topic is a probability distribution across words). The model allows the sharing of topics both within each corpus and between corpora. Teh et al. [92] compared three models: 2-level HDP on documents from one category, 2-level HDP on documents from different categories, and 3-level HDP on documents from different categories. The second model yielded the poorest performance, proving the need to consider the category information of groups. Unfortunately, in many cases we do not known the category information.

[1] Here GEM stands for Griffiths, Engen, and McCloskey [79]. We say $\boldsymbol{\beta} \sim$ GEM(γ) if we have $\beta_k = \beta'_k \prod_{k=1}^{k-1}(1 - \beta'_k)$ for $k = 1, ..., \infty$, where $\beta'_k \sim \text{Beta}(1, \gamma)$.

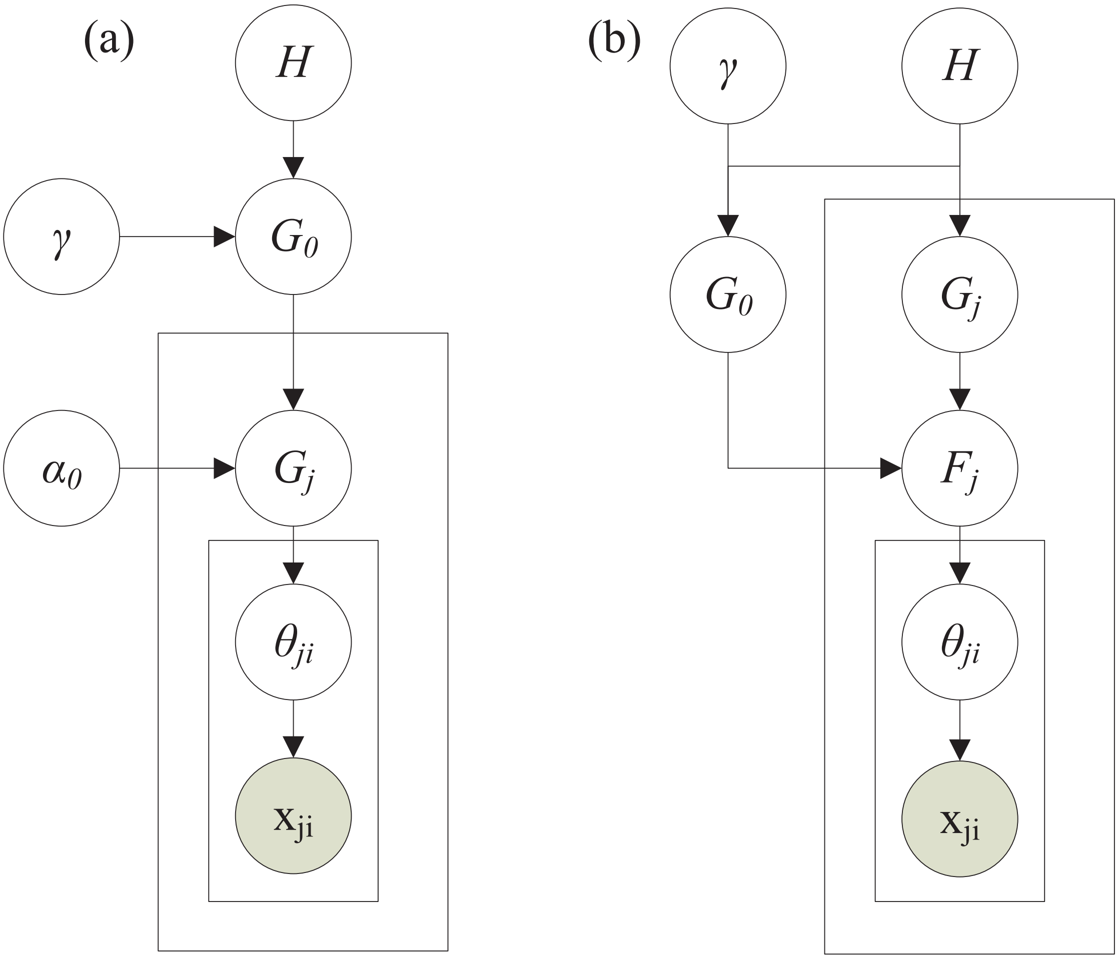


Figure 4.1: Graphical model representations of (a) HDP and (b)LC-HDP.

4.3.3 LC-HDP

Motivated by a similar problem to that of the HDP [92], Müller et. al. [68] developed another hierarchical Dirichlet process (LC-HDP). They considered a model in which a coupled set of random measures F_j is defined as

$$\begin{aligned} F_j &= \epsilon G_0 + (1-\epsilon) G_j, \\ G_j &\sim \mathrm{DP}(\gamma, H) \text{ for } j = 0, 1, ...J. \end{aligned} \tag{4.6}$$

where $0 \leq \epsilon \leq 1$ defines weights of the linear combination. This model provides an alternative approach to sharing atoms, in which the shared atoms are given the same stick-breaking weights in each of the groups. It has an attractive characteristic that it can discriminate local components, which are useful for clustering. We compare graphical model representations of the HDP [92] and LC-HDP [68] in Figure 4.1.

4.3.4 The Nested Dirichlet Process

The NDP [84] is motivated by simultaneously clustering groups and observations within groups. It induces multi-level clustering, while the HDP can cluster only observations. In the NDP model, the groups are clustered by their entire distribution. Consider a set of distributions $\{G_j\}$, each for one group. If $\{G_j\} \sim \text{nDP}(\alpha, \gamma, H)$, it means that for each group j, $G_j \sim Q$ with $Q \equiv \text{DP}(\alpha \text{DP}(\gamma H))$. This implies that we can first define a collection of DPs

$$G_k^* \equiv \sum_{l=1}^{\infty} w_{lk} \delta_{\theta_{lk}^*} \text{ with } \theta_{lk}^* \sim H, (w_{lk})_{l=1}^{\infty} \sim \text{GEM}(\gamma)$$

and then draw the group specific distributions G_j from the following mixture

$$G_j \sim Q \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*} \text{ with } (\pi_k)_{k=1}^{\infty} \sim \text{GEM}(\alpha)$$

The process ensures G_j in different groups can select the same G_k^*, leading to clustering of groups.

Although the NDP can also borrow information across groups, groups belonging to different clusters cannot share any atoms. For the NDP, the different distributions have either the same atoms with the same weights or completely different atoms and weights.

4.3.5 Beta Processes and Latent Feature Analysis

The beta process(BP) [93, 75] and the related Indian buffet process(IBP) [32] are often applied to factor/feature analysis to infer a set of factors with which data may be sparsely represented. By defining the infinite dimensional priors, these factor analysis models need not to specify the number of latent factors but automatically determine it.

Definition of BP: Here we avoid using a complete measure-space definition for the beta process, but follow the representation form given by [77].

Let B_0 be a continuous measure on a space Θ; $B_0(\Theta) = \gamma$; and α is a positive scalar. If B_k is defined as follows,

$$\begin{aligned} B_k &= \sum_{k=1}^{N} \pi_k \delta_{\theta_k} \\ \pi_k &\sim Beta(\frac{\alpha\gamma}{N}, \alpha(1-\frac{\gamma}{N})) \\ \theta_k &\sim \frac{1}{\gamma} B_0 \end{aligned} \tag{4.7}$$

then as $N \to \infty, B_k \to B$ and B is a beta process: $B \sim BP(\alpha B_0)$.

Finite Approximation: The beta process is defined on an infinite parameter space, but sometimes we can also use its finite approximation by simply setting N to a large number.

Bernoulli Process: The beta process is conjugate to a class of Bernoulli processes, denoted by $X \sim Bep(B)$. If B is discrete, of the form in (4.7), then $X = \sum_k b_k \delta_{\theta_k}$ where the b_k are independent Bernoulli variables with the probability that $b_k = 1$ equal to π_k. Due to the conjugation between the beta process priors and Bernoulli process, the posterior of B given N samples $X_i \sim Bep(B)$ is also a beta process which has updated parameters:

$$B|X_1, X_2, ..., X_M \sim BP(c + M, \frac{c}{c + M} B_0 + \frac{1}{c + M} \sum_i X_i) \tag{4.8}$$

Application of BP: Furthermore, marginalizing over the beta process measure B and taking $c = 1$, provides a predictive distribution on indicators known as the Indian buffet process (IBP) [93]. Linking the beta process and the Bernoulli process is often used in a feature analysis model to generate infinite vectors of binary indicator variables, which indicates whether a feature is used to represent a sample. In this paper, we propose a similar method to indicate which sentences are used to represent a document.

4.3.6 Bayesian Nonparametric Topic Models

A problem of the classic topic models is how to find a proper number of topics. For example, in PLSI [36] and LDA [9], we have to predefine the number of topics before we construct the model. This impedes the flexibility of the models. Bayesian nonparametric methods are suitable to solve the problem. Especially, the Dirichlet Proceess (DP) is an appropriate tool to extend finite mixture models to nonparametric models. So the topic distribution of each document is generated by a DP. Then, to guarantee the topics are shared across documents, all these document-specific DPs shares the same base measure which is an another DP. This method derives the HDP based topic model, HDP-LDA [92].

Besides the HDP-LDA, other Bayesian nonparametric priors are also utilized in the topic modeling. For example, the infinite buffet process (IBP) [32] is used to build a sparse topic model, where each topic is associated with only a subset of the vocabulary.

4.3.7 Bayesian Nonparametric Methods in Document Summarization

Recently, some BNP models are also involved in document summarization approaches [11, 17]. BNP priors such as the nested Chinese restaurant process (nCRP) [7] are associated with topic analysis in these models. Then the topic distributions are used to get the sentence scores and rank sentences as in Section 4.2. BNP here only impacts the number and the structure of the latent topics, but the summarization framework is still constant-length. Our BNP summarization model differs from the previous models. Besides using the HDP for topic analysis, our approach further integrates the beta process into sentence selection. The BNP method in our model are directly used to determine the number of summary sentences but not latent topics.

4.4 Hybrid Nested Dirichlet Processes for Topic Modeling

4.4.1 Introduction

In previous sections, we introduced the Bayesian nonparametric (BNP) models [73, 28] which have attracted a lot of attention in the machine learning and data mining community recently. In this section, we improved a well known Baysian nonparametric model, the the hierarchical Dirichlet process (HDP), to adapt it to a new environment.

Among the various BNP priors, the Dirichlet process (DP) is one of the most widely used priors owing to its efficiency of inference [69]. The DP is often associated with a mixture model, resulting in a Dirichlet process mixture (DPM) model [69]. One basic assumption of the DPM is that the observations are infinitely exchangeable. However, this assumption does not hold when data comes from multiple groups, where observations from different groups are generally not exchangeable. To model grouped data, Teh et al. [92] advanced the hierarchical Dirichlet process (HDP), which constructs multiple DPs by sharing the base measure which is drawn by another DP. With this setting, the HDP allows different groups to share mixture components. Moreover, motivated by the same problem, Müller et. al. [68] developed an alternative model which is also called HDP. They defined a random measure for each group as a linear combination of two independent samples from DPs. One is shared across the groups, while the other is idiosyncratic. In this thesis, we call Müller's model LC-HDP (Linear Combination-HDP) to distinguish it from Teh's HDP model.

The HDP has achieved great success for modeling groups of data and it has been applied to various areas such as topic modeling and hidden Markov models. It assumes that each group distribution is conditionally independent based on the same base measure. However, this assumption ignores the category information of groups. If we consider a group of data as an object; the objects are often organized into categories, such as documents in multi-corpora data and epileptic seizures (groups of channels) across patients [101]. Intuitively, objects within the same categroy should be more similar to each other than to those in other categories. These kinds of category information are useful for modeling data [44, 82], and Teh et al. [92] demonstrate that ignoring the category information would result in much worse performance. Nevertheless, in many cases the category information is difficult to get; and discovering the implicit categories is an important task [105].

Here, we consider the case that the membership of the grouped data is unknown, and we develop a hybrid nested/hierarchical Dirichlet process (hNHDP) model [58] uncovering the latent categories and taking advantage of it. We borrow the idea from the nested Dirichlet process (NDP) [84], which is able to simultaneously cluster groups and observations within groups. In the HDP model, two distributions share all atoms but they are assigned different weights to them. The LC-HDP allows distributions

Table 4.1: Comparison of different models.

	[92]	[68]	[84]	This work
Model	HDP	LC-HDP	NDP	hNHDP
Sharing atoms	$\surd$	$\surd$	-	$\surd$
Clustering data groups	-	-	$\surd$	$\surd$
Discriminating local components	-	$\surd$	-	$\surd$
Top-level base measures	H	H	H	$H_0\&H_1$

to share only part of atoms. NDP, on the other hand, leads to distributions that have either the same atoms with the same weights or completely different atoms and weights. This induces clustering in both observations and distributions.

We combine elements of the NDP and the LC-HDP in our model. We cluster the group distributions as in the NDP. However, different from the NDP, our model generates distributions sharing atoms between groups from different clusters; and we cluster the groups using only some local components. We define the distribution F_k for each cluster k as a mixture of two independently drawn DPs as in the LC-HDP: G_0 which is shared by all clusters and G_k which is cluster-specific. Through some settings, we make F_k still a realization of the DP (this is not guaranteed by the original LC-HDP). Moreover, we set different base measures (H_0 and H_1) for G_0 and G_j. Thus G_j can only include useful features for clustering. This setting is based on consideration of feature selection in data clustering. Selecting only a subset of features are enough to get good clustering performance, while including irrelevant features may even harm the clustering. The properties of the proposed model are summarized and compared with those of other models in Tabel 4.1.

We apply the hNHDP to the problem of topic modeling, which is a suitable case for illustrating the power of the prior and an alternative to document summarization as well. Our model assumes documents to be mixtures of topics and assigns documents into latent categories. It automatically identifies local words for local topics by word

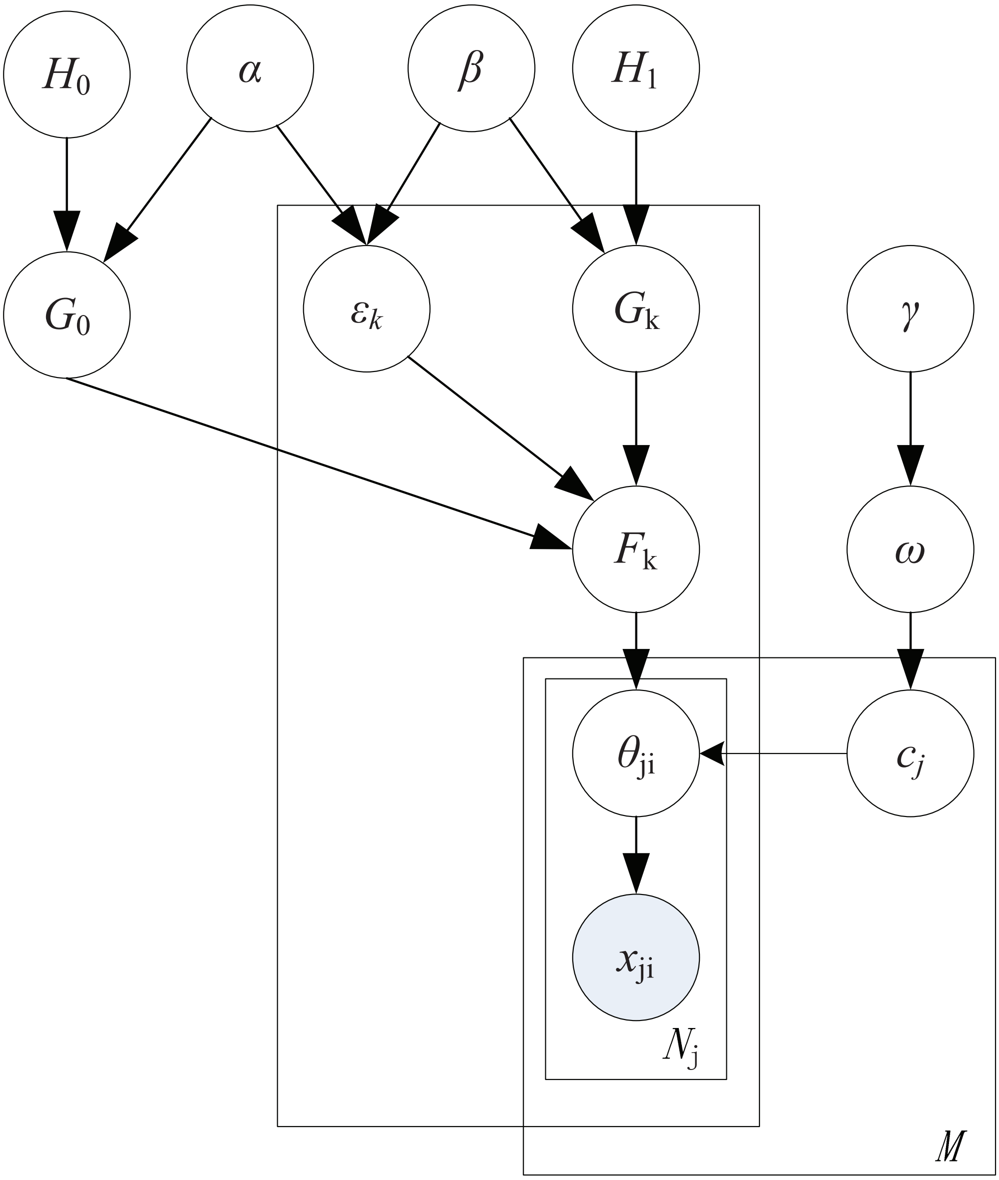


Figure 4.2: Graphical model representation of hNHDP.

differentiation. It reveals topic structures and dependencies, which are visualized in our experiments.

4.4.2 The Hybrid Nested/Hierarchical Dirichlet Process

We propose the hybrid nested/hierarchical Dirichlet process (hNHDP) mixture model for groups of data. Following the setting of HDP, assume that we have M groups of data. Each group is denoted as $\mathbf{x}_j = x_{j1}, ..., x_{jN_j}$, where $\{x_{ji}\}$ are observations and N_j is the number of observations in group j. Each x_{ji} is associated with a distribution $p(\theta_{ji})$ with parameter θ_{ji}. For example, in topic modeling the distribution p is a multinomial distribution. We now describe the generative process of observations using the hNHDP model.

As in the NDP, we first consider the set of distributions $\{F_k\}$ for different clusters. For each cluster (latent category) k, we model F_k as a combination of two components, G_0 and G_k. This setting is similar to that for the LC-HDP, but we impose some additional restrictions on the parameters. The combination weight ϵ_k is changed for each cluster, and the two components are drawn from DPs with different base measures.

$$\begin{aligned} G_0 &\sim \text{DP}(\alpha, H_0), \\ G_k &\sim \text{DP}(\beta, H_1) \text{ for each } k, \\ \epsilon_k &\sim \text{Beta}(\alpha, \beta) \text{ for each } k, \\ F_k &= \epsilon_k G_0 + (1-\epsilon_k) G_k \text{ for each } k. \end{aligned} \tag{4.9}$$

After getting the cluster-specific distributions, we assign the group distributions F'_j to the set $\{F_k\}$. This hierarchy is the same as that of the NDP.

$$F'_j \sim \sum_{k=1}^{\infty} \omega_k \delta_{F_k} \tag{4.10}$$

where $\boldsymbol{\omega} = \{\omega_k\} \sim \mathrm{GEM}(\gamma)$. This is equal to selecting a cluster label k for a group and then assigning F_k to the group as its distribution. Then we generate observations using the following process.

- For each object $\mathbf{x}_j$,
    - Draw a cluster label $c_j \sim \boldsymbol{\omega}$;
    - For each observation x_{ji}
        - $\theta_{ji} \sim F_{c_j}$;
        - $x_{ji} \sim p(x_{ji}|\theta_{ji})$.

A graphical model representation is shown in Figure 4.2.We can also define the hNHDP mixture model in another way. For each group, the observations are independently drawn from the distribution

$$\mathcal{P}_j(\cdot) = \int p(\cdot|\theta)d(F'_j(\theta))$$

where F'_j is drawn from the hNHDP prior as above and $x_{ji} \sim \mathcal{P}_j$ for each i in group j.

Model Properties

The hNHDP has some interesting properties:

- (1) F_k is still a sample from a DP.
- (2) F'_j can share atoms that are generated from G_0.

In [50], the authors proposed a new construction of DPs by three operations based on existing ones. This construction is also used to derive a coupled mixture model for groups of data [49]. Here we cite one of the operations: the superposition.

Superposition 1 *Let $D_k \sim DP(\alpha_k B_k)$ for $k = 1, ..., K$ be independent DPs and $(c_1, ..., c_k) \sim Dir(\alpha_1, ..., \alpha_k)$. Then the stochastic convex combination of these DPs remains a DP,*

$$c_1 D_1 + \cdots + c_k D_k \sim DP(\alpha_1 B_1 + \cdots + \alpha_k B_k).$$

From the set of equations in (4.9), we can infer that cluster-specific distribution F_k in the hNHDP model is still a realization of DP.

$$F_k \sim \mathrm{DP}(\alpha H_0 + \beta H_1). \tag{4.11}$$

With this form, the hNHDP can be transferred into a special NDP. However, the generative process of (4.11) is not the same as that of (4.9) because G_0 is only sampled once in (4.9). If we directly use form (4.11) for each cluster, H_0 will generate different atoms for each cluster.

Now we consider the relationship between the hNHDP and the LC-HDP. We ignore the clustering structure of the hNHDP, and focus on only the group-specific distributions $\{F'_j\}$. For each group j, F'_j can be written as $\epsilon_k G_0 + (1 - \epsilon_k) G'_j$ where $k = c_j$ is the cluster label and $G'_j = G_k$. If ϵ_k is same for all k,the hNHDP degenerates into a special LC-HDP. It also indicates that F'_j can share atoms generated from a global component G_0.

4.4.3 Application to Topic Modeling

Category information is useful for modeling complex data. For example, in the area of topic modeling, the discriminative LDA [44] and the labeled LDA [82], which utilize the side information of documents, have better predictive performance than the general unsupervised LDA (latent Dirichlet allocation [9]). Let us return to the HDP. When it is used for topic modeling (HDP-LDA), one document is regarded as a group. If we have documents in multiple corpora (each corpus is a category), the HDP is extended to a 3-level model to integrate the category information of documents. As we introduced before, Teh et al. [92] demonstrated that a 3-level HDP that considers the category information of documents performs better than a 2-level HDP that treats documents from different corpora in the same way. All these studies proved the advantage of discriminating documents from different categories in text modeling. This stimulated us to take consideration of the document structure and to utilize it.

4.4.4 The hNHDP model for Topic Modeling

We consider the case where the category information of documents is unknown and develop an hNHDP model for topic modeling in this case. We assume that word is an observation and that a document is a group. We generate the parameter distribution F'_j for each document j using the generative process in Section 4.4.2, where the distribution $p(\theta_{ji})$ is set as a multinomial distribution with parameter θ_{ji}. The base measures H_0 and H_1 are set as Dirichlet distributions over words.

In clustering analysis, feature selection is a very important task. By selecting a subset of efficient features, feature selection can improve the text clustering efficiency and performance [51]. Feature selection has already been used in the Dirichlet mixture models for clustering [42, 104]. Moreover, word selection has also been successfully used in sparse topic models [98]. So we also want to integrate feature selection into our model to reduce the dimension of topics and improve the clustering performance. Thus we develop the following process of word differentiation.

Assuming that the size of the vocabulary is V, we bring in a binary vector $\mathbf{q^1} = (q^1_1, ..., q^1_V)$ to select discriminative words and separate the vocabulary into two disjoint sets. If $q^1_v = 1$, the word v is regarded as discriminative and included only in local topics. Otherwise, v is regarded as global and included only in global topics. In this way we get two disjoint base measures, H_0 and H_1, for the hNHDP.

- For each word v
  - $q_v^0 \sim \text{Bernoulli}(\pi)$.
  - $q_v^1 = 1 - q_v^0$.
- $H_0 = \text{Dir}(\eta\mathbf{q^0})$;
- $H_1 = \text{Dir}(\eta\mathbf{q^1})$.

Here, $\mathbf{q^0}$ and $\mathbf{q^1}$ are binary vectors $\mathbf{q^0} = (q_1^0, ..., q_V^0)$ and $\mathbf{q^1} = (q_1^1, ..., q_V^1)$. For the parameter π, we set $\pi \sim \text{Beta}(\alpha, \beta)$ to conform with the hyperparameters of ϵ_k.

In practice, we may not require that the q_v^0 is uncertain if we already know the feature words; and it is also possible that $q_v^1 \neq 1 - q_v^0$. However, these cases are not discussed in this paper.

Finite Approximation

In Bayesian statistics, the Dirichlet-multinomial allocation (DMA) has often been applied as a finite approximation to the DP [105, 104]. It takes the form $G_N = \sum_{l=1}^{N} \pi_l \delta_{\theta_l}$, where $\pi = (\pi_{\mathbf{1}}, ..., \pi_{\mathbf{N}})$ is an N-dimensional vector distributed as a Dirichlet distribution $Dir(\alpha/N, ...\alpha/N)$. In our inference step, we approximate ω in (4.10) by a finite Dirichlet distribution

$$\omega \sim \text{Dir}(\gamma/K, ...\gamma/K). \tag{4.12}$$

The G_0 and G_1 are also approximated by the DMA.

$$\begin{aligned} G_0 &= \sum_{l=1}^{L} w_{l0}\delta_{\theta_{l0}} \\ G_k &= \sum_{l=1}^{L} w_{lk}\delta_{\theta_{lk}} \end{aligned} \tag{4.13}$$

where $(w_{10}, ..., w_{L0}) \sim \text{Dir}(\alpha/L, ...\alpha/L)$ and $(w_{1k}, ..., w_{Lk}) \sim \text{Dir}(\beta/L, ...\beta/L)$. If we set K and L large, the DMA can give a good approximation in our model.

Comparison with Related Work

Some relevant models using a similar terminology or focusing on a similar problem have been proposed. The Dirichlet enhanced latent semantic analysis (DELSA) model [105] extends LDA by revealing the clustering structure of data. It replace the parametric Dirichlet prior distribution in LDA by a DP. However, it is still parametric when generating topics because it is based on LDA and it requires the topic number to be given. By contrast, in a Bayesian nonparametric topic model, such as the HDP-LDA and our model, the topic number can be inferred.

Paisley et al. [76] developed a nested hierarchical Dirichlet process (nHDP) for hierarchical topic modeling. The model is a generalization of the nested Chinese restaurant process (nCRP) [7], which allows each word to follow its own path to a topic node according to a document-specific distribution on a shared tree. Our model is based on the HDP and NDP, which are different from the nCRP when modeling topics.

More recently, Agrawal et al. [1] proposed an alternative Nested Hierarchical Dirichlet Process (which is also called nHDP). They addressed the problem of modeling documents associated with entities. Their proposed model is an HDP-nesting-HDP model, which allows entities for different documents to be shared. It also utilizes the category information of documents. Documents are clustered by entities. Our model differs from it because ours is based on the LC-HDP. Besides, the hNHDP model identifies local words and local topics, which are never realized by other models.

4.4.5 Inference

We use the Gibbs sampling method to infer the posterior of parameters. The inference proceeds through the following steps.

Sampling the cluster indicators c_j.

As we use the DMA approximation for ω in (4.12), the probability of cluster assignments conditioned on other variables can be calculated as

$$P(c_j = k|c_{-j}, ...) \propto \frac{m_k - 1 + \gamma/K}{M - 1 + \gamma} * \\ \prod_{x_{ji}} \sum_{l=1}^{L} \left(\epsilon_k w_{l0} P(x_{ji}|\theta_{l0}) + (1 - \epsilon_k) w_{lk} P(x_{ji}|\theta_{lk})\right),$$

where $p(x_{ji}|\theta_{lk}) = \theta_{lk}^v, v = x_{ji}$. M is the number of documents, and m_k is the number of documents assigned to cluster k.

Since the global words and local words are disjoint, we can re-write the upper equation as

$$P(c_j = k|c_{-j}, ...) \propto \frac{m_k - 1 + \gamma/K}{M - 1 + \gamma} * \\ \prod_{x_{ji} \in A_0} \sum_{l=1}^{L} (\epsilon_k w_{l0} P(x_{ji}|\theta_{l0})) * \\ \prod_{x_{ji} \in A_1} \sum_{l=1}^{L} ((1 - \epsilon_k) w_{lk} P(x_{ji}|\theta_{lk})),$$

where $A_0 := \{v|q_v^0 = 1\}$ and $A_1 := \{v|q_v^1 = 1\}$ are the sets of global words and local words.

Sampling topic assignment z_{ji} for each word x_{ji}.

$$\begin{aligned} P(z_{ji} = t_{lk}|c_j = k, ...) &\propto (1 - \epsilon_k) * w_{lk} P(x_{ji}|\theta_{lk}) \\ P(z_{ji} = t_{l0}|c_j = k, ...) &\propto \epsilon_k * w_{l0} P(x_{ji}|\theta_{l0}), \end{aligned} \tag{4.14}$$

where t_{lk} and t_{l0} are topic indices.

Sampling the weights $\{w_{l0}\}$ and $\{w_{lk}\}$ for G_0 and G_1.

$$(w_{1k}, ..., w_{Lk}) \sim \mathrm{Dir}(\beta/L + n_{1k}, ..., \beta/L + n_{Lk}) \tag{4.15}$$

$$(w_{10}, ..., w_{L0}) \sim \mathrm{Dir}(\alpha/L + n_{10}, ..., \alpha/L + n_{L0}), \tag{4.16}$$

where n_{lk} is the number of words assigned to topic t_{lk}.

Sampling θ_{lk} and θ_{l0}

$$(\theta_{lk}|...) \sim \mathrm{Dir}(\eta q_1^1 + n_{lk}^1, ..., \eta q_V^1 + n_{lk}^V) \tag{4.17}$$

$$(\theta_{l0}|...) \sim \mathrm{Dir}(\eta q_1^0 + n_{l0}^1, ..., \eta q_V^0 + n_{l0}^V), \tag{4.18}$$

n_{lk}^v is the count when the word v assigned to topic lk.

Sampling ϵ_k

$$(\epsilon_k|...) \sim \mathrm{Beta}(\alpha + \sum_{l=1}^{L} n_{l0}, \beta + \sum_{l=1}^{L} n_{lk}) \tag{4.19}$$

Sampling $\mathbf{q}$. For the word selection variable $\mathbf{q}$ (including q^0 and q^1)[2], we use the Metropolis-Hastings algorithm. In each step, we randomly select a word v and invert its q_v value. When q changes, the associated θ (i.e. the collection of θ_{lk} and θ_{l0}) should also be changed. As it is difficult to integrate out θ for the posterior distribution of q, we update q with θ together. The new candidates $\mathbf{q}^*$ and θ^* are accepted with probability

$$\min\left\{1, \frac{P(\mathbf{q}^{0*}, \theta^*|\mathbf{c}, \mathbf{X}, ...)P(\mathbf{q}^0, \theta|\mathbf{q}^{0*}, \theta^*)}{P(\mathbf{q}^0, \theta|\mathbf{c}, \mathbf{X}, ...)P(\mathbf{q}^{0*}, \theta^*|\mathbf{q}^0, \theta)}\right\} \tag{4.20}$$

This is equal to

$$\min\left\{1, \frac{P(\mathbf{X}|\mathbf{q}^{0*}, \theta^*, ...)P(q_v^{0*})}{P(\mathbf{X}|\mathbf{q}^0, \theta, , ...)P(q_v^0)}\right\} \tag{4.21}$$

where X is the collection of all the documents.

[2] $\mathbf{q}^1$ is dependent on $\mathbf{q}^0$ via the equation $\mathbf{q}^1 = 1 - \mathbf{q}^0$, so we only need to consider $\mathbf{q}^0$ here.

Sampling π

$$(\pi|...) \sim \text{Beta}(\alpha + N_0, \beta + N_1). \tag{4.22}$$

N_0 and N_1 are the numbers of unique words identified as global and local ones respectively. Notice that $N_1 \neq \sum_{l,k} n_{l,k}$, because N_1 counts each word only once.

4.4.6 Experiments

Simulation Study

We designed a simulation study to show two aspects of our model: (a) the effectiveness of finding relevant clusters and (b) the ability to find cluster-specific words and topics. We generated toy datasets with the following steps:

(1) Set the cluster number K' and vocabulary size V. Choose some words as general words and the other as local words. (2) Generate L_1 global topics for all clusters and L_2 local topics for each cluster. The global topics are defined as Dirichlet distributions over all global words, while the local topics are Dirichlet distributions over all local words. Then define each cluster as a mixture of global topics and local topics belonging to it. (3) Generate M documents. For each document, we first select a cluster label for it and then sample D words according to the corresponding cluster distribution.

First, we set $K' = 4$, $V = 16$, $L_1 = L_2 = 2$, $D = 100$ and $M = 500$. As we wanted to show the discriminative words, we used a small vocabulary here. Figures 4.3 and 4.4 illustrate the clustering process and the word differentiation results of our model. The clustering result perfectly matches the real assignments, while the local words we extracted are close to the original setting. The diference may be caused by the small number of samples but the extracted discriminative words are enough for accurate clustering.

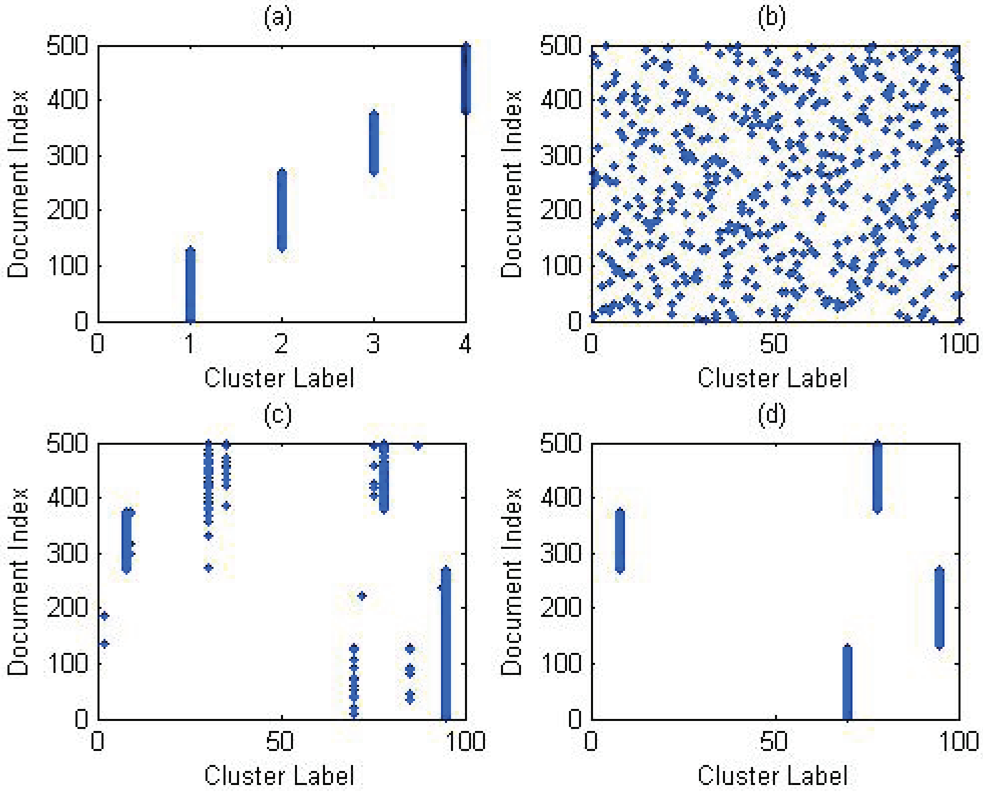


Figure 4.3: Clustering results on toy dataset1. (a)-(d) show the clustering assignments at different iterations: (a) real cluster assignments, (b)initial random assignments, (c) assignments after one iteration, and (d) final assignments (after 30 iterations).

Next, we illustrate the robustness of our model when the proportion of discriminative words is changed. We set $K' = 4$, $V = 200$, $L_1 = L_2 = 5$, $D = 100$, $M = 600$, and varied the proportion of discriminative words from 20% to 80% for 20 trials. For each trial, we sampled for 1000 iterations and discard the first 500. Our model got perfect clustering results in all trials, and on average the accuracy of word differentiation was 83%.

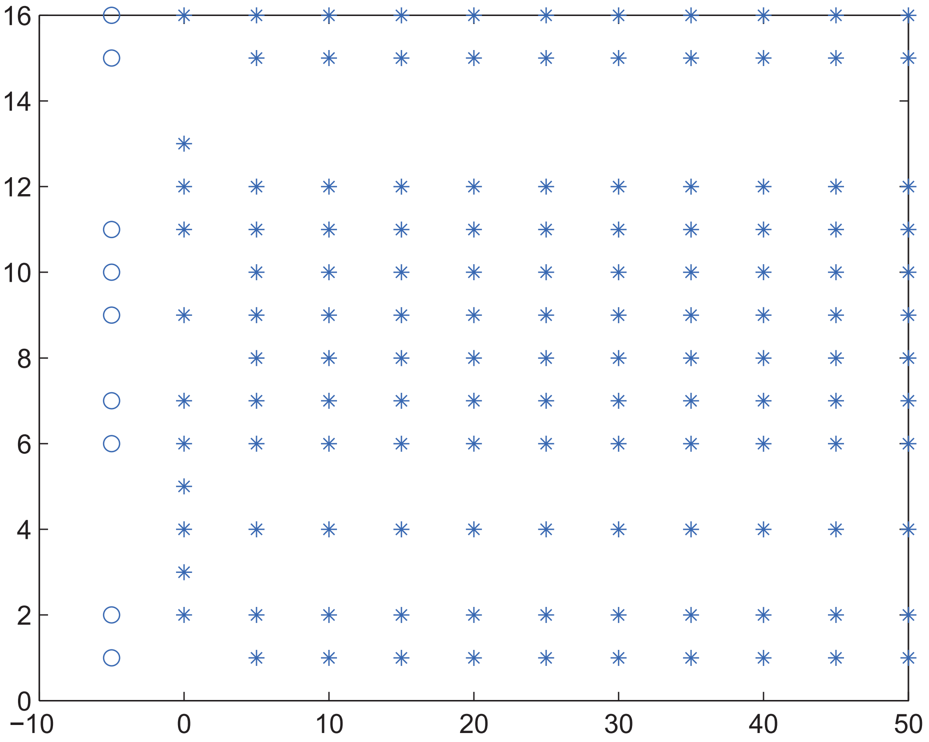


Figure 4.4: Word differentiation on toy dataset1.

Document Modeling on Real Data

We implemented the proposed hNHDP model on two real-world text datasets. The first one is the *NIPS* data, which is used in [92][3]. This version of *NIPS* data collects NIPS articles from 1988–1999 and unifies the section labels in different years. It contains 13649 unique words and 1575 articles separated into nine sections: algorithms and architectures, applications, cognitive science, control and navigation, implementations, learning theory, neuroscience, signal processing, and vision sciences. The other dataset is ”6 conference abstracts (*6conf*)”, which contains abstracts from six international conferences (IJCAI, SIGIR, ICML, KDD, CVPR, and WWW) collected

[3] `http://www.stats.ox.ac.uk/~teh/research/data/nips0_12.mat`

by [19]. It has 11,456 documents and 4083 unique words. We preprocesses the data by removing the stop words and stemming.

We compared our model with other models on training sets of various sizes as in [49]. Each dataset was randomly separated into two disjoint sets, one for training and the other for testing. We generated 6 pairs of training/testing datasets for NIPS data and 5 pairs for 6conf data. For all the real datasets experiments, we used the same setting of parameter settings for our model. We gave the hyperparameter γ a vague value $Gamma(0.1, 1)$ and set $\eta = 0.5$ for H_0 and H_1. The component numbers in DMA approximation are set as $K = 100$, $L = 30$. The other parameters were $\alpha = \beta = 1$. For each training set, we ran 1000 iterations and treated the first 500 as burn-in. In the initialization step, we used a simple feature selection method which ranks words by term variance quality [20]. We selected a random proportion of highest ranked words as discriminative words, while the others were set as global words. This allowed us to accelerate the convergence in the sampling process.

The models used for comparison were the following:

- `HDP-LDA` [92]. We used the HDP mixture model which does not consider the category information of documents. Articles from different sections were not treated differently. We followed the parameter setting procedure given in [92]. The concentration parameters for the two levels were given as: $\gamma \sim Gamma(5, 0.1)$, $\alpha \sim Gamma(0.1, 0.1)$. The base measure of the bottom level was a symmetric Dirichlet distribution over all words with parameters of 0.5.

- `NDP`. This model is based on the nested Dirichlet process. In its settings, G_0 did not exist and $F_k = G_k$. Since the NDP model does not share topics between clusters, it does not distinguish either local topics or local words.

- `hNHDP-nosel`. For this model, we used the same structure as for the proposed hNHDP model. The difference was that here we set $H_0 = H_1 \sim Dir(\eta)$. The

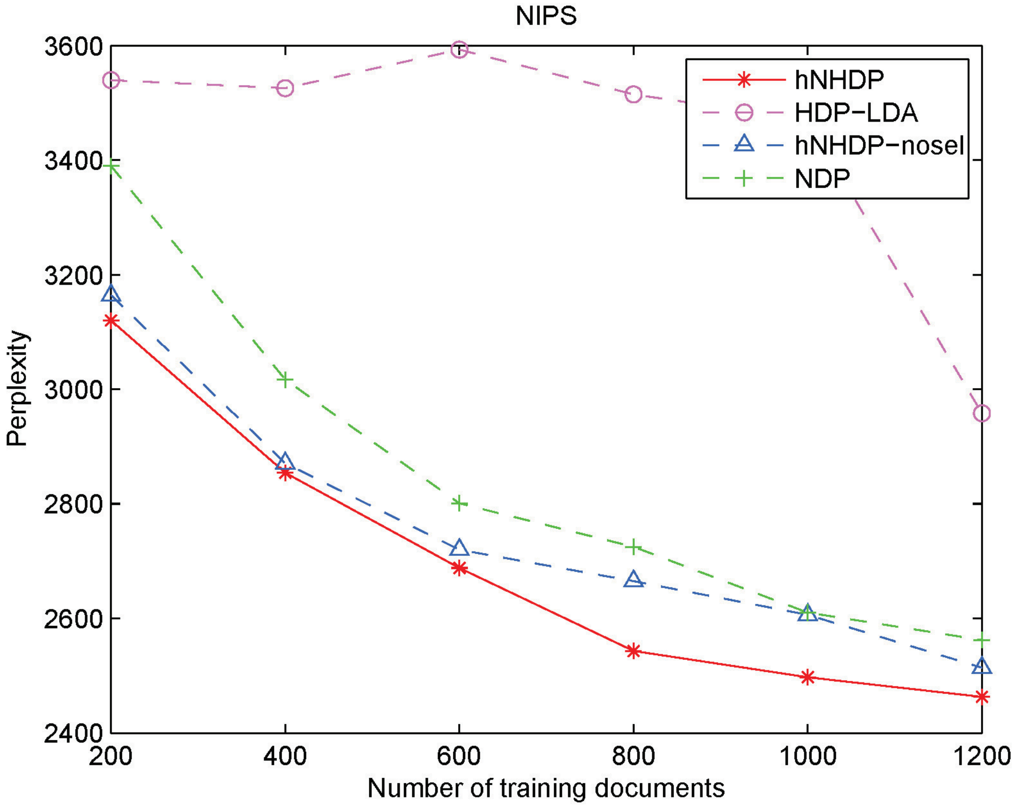


Figure 4.5: Results of document modeling on NIPS data.

base measures H_0 and H_1 were symmetric Dirichlet distributions over all words. In other words, this model does not differentiate words between global and local, while it remains to distinct global topics from local topics.

We evaluated all the models with the test-set perplexity, a standard metric for document modeling which measures how well the models generalize to new data. The perplexity is defined as follows.

$$perplexity(D_{test}) = \exp(-\frac{\sum_{d \in D_{test}} \log p(\mathbf{x}_d | D_{train})}{\sum_{d \in D_{test}} N_d}).$$

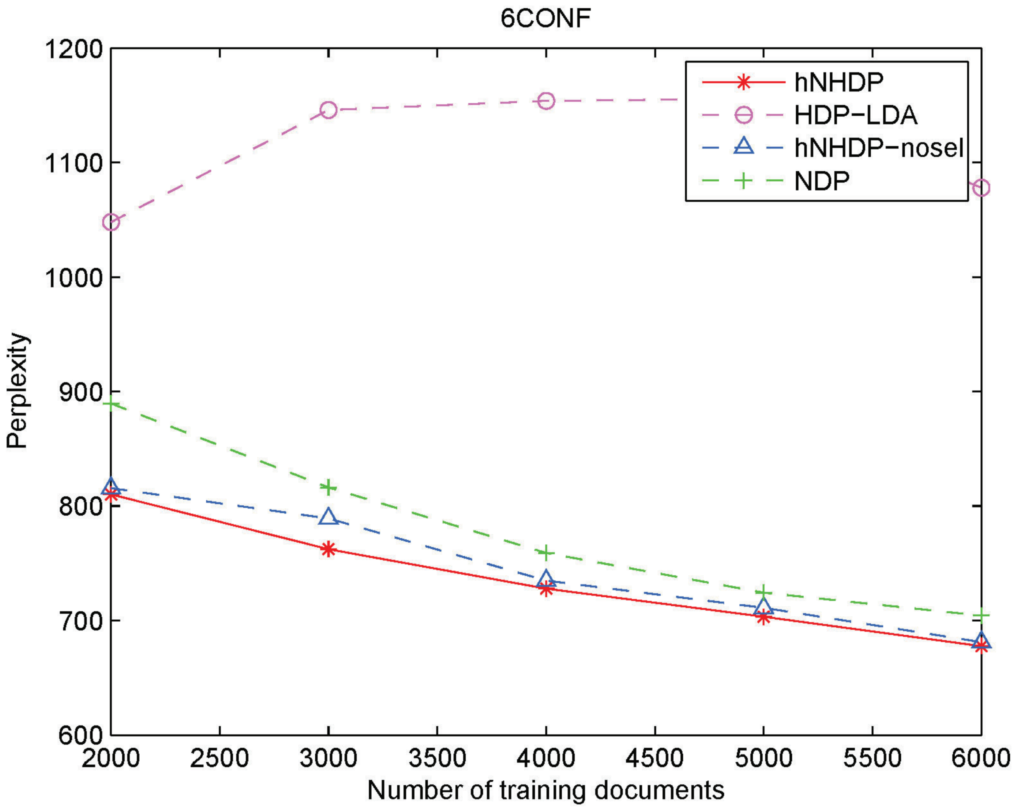


Figure 4.6: Results of document modeling on 6conf data.

Figures 4.5 and 4.6 compare the perplexities on the two datasets. Our proposed model, hNHDP, achieved the best perplexities (lower is better) in all runs. Especially, it exceeded HDP-LDA by a large amount.

In addition, with the same setting for topic-word distributions (a symmetric Dirichlet distribution over all words in vocabulary), the hNHDP-nosel performed better than the NDP and HDP. The former demonstrates the advantage of distinguishing local topics, while the latter indicates the effectiveness of taking advantage of the clustering structure of documents. We also noticed that the performance of the hHNDP was obviously better than that of the hHNDP-nosel only for some training sizes, while the two models got comparable results in other runs. This is reasonable

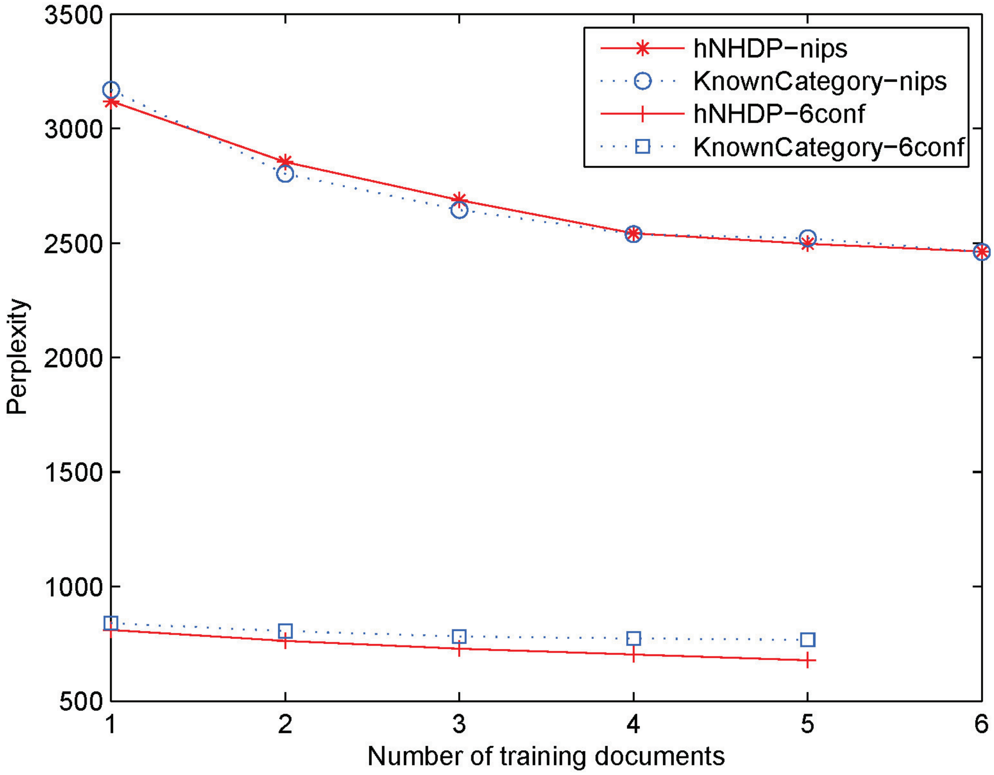


Figure 4.7: Comparison of hNHDP and hNHDP-knowncategory.

because the vocabulary size was so large that the tail words of each topic may have contributed little. Thus, we may suppose that the global words in local topics and the local words in global topics contributed little to predictive performance, resulting in a result similar to the hNHDP's. Nevertheless, the hNHDP still achieved our aim by greatly reducing the model complexity without any performance decrease (in fact the performance increased a little). Its ability to extract local topics and local words is also very useful.

In addition, we wanted to show how well the latent categories found by hNHDP improves document modeling. We developed another model, `hNHDP-knowncategory`, which assumes that the category labels of documents are known. It assigns real labels to documents in the hNHDP model and does not change them during iterations. Comparing the two models (Figure 4.7), we found that the performance of hNHDP was comparable to `hNHDP-knowncategory` on NIPS data, while on 6conf data it was even better. These results indicate the efficiency of hNHDP for document modeling.

Clustering and Visualization

Our last experiment was designed to show the clustering performance of the hNHDP. We first present the clustering results on *6conf* data in Figure 4.8. Although the number of clusters inferred by our model is a little larger than the real one, each conference has its specific clusters, which we can easily differentiate in the figure. Moreover, we could also find the connection between the conferences in Figure 4.8. CVPR is separate from the others, with only a little connection with ICML and IJCAI. ICML and KDD have a large overlap, but ICML has an additional cluster component. SIGIR and WWW also own the same major clusters although the cluster densities may differ. IJCAI is a comprehensive conference, so it includes several clusters shared by other conferences as well as a specific cluster.

We then extracted the typical topics in each cluster and matched them with corresponding conferences. The topics are shown in Table 4.2. The global topics/words and the local topics/words are easily distinguished in the table. The local topics conform to the features of different conferences. From Figure 4.8 and Table 4.2, we can see that both the clusters and the topics can be well explained and that they reveal the structure and features of the data.

We also make a quantitative evaluation of the clustering results, although clustering can be regarded as only a by-product of hNHDP. The evaluation metric used here was the normalized mutual information (*NMI*). It is a clustering accuracy measure that is tolerant to mismatches between the number of clusters and the number of reference classes. Following the definition in [107], *NMI* was estimated as follows:

Table 4.2: Some example topics extracted from *6conf* data. Each column is a topic; for each topic the top 15 words are shown. The numbers in brackets are the typical cluster numbers of each conference shown in Figure 4.8.

Global Topics	Local Topics					
	CVPR(23)	ICML(16)	IJCAI(15)	KDD(52)	SIGIR(31)	WWW(28)
task	imag	learn	system	data	retriev	web
predict	model	algorithm	base	cluster	queri	search
experiment	recognit	problem	knowledg	mine	inform	base
work	object	method	model	model	model	document
describ	base	search	gener	base	document	user
fast	track	reinforc	program	algorithm	base	inform
solut	segment	gener	languag	structur	system	queri
requir	motion	optim	plan	time	relev	retriev
specif	shape	base	learn	graph	languag	content
context	visual	model	semant	pattern	term	index
express	detect	function	process	learn	data	text
categor	estim	plan	problem	detect	effect	approach
parallel	surfac	approach	natur	network	search	page
onlin	vision	constraint	comput	distribut	method	system
properti	match	structur	domain	method	text	model

Table 4.3: Description of datasets for clustering.

Datasets	*s*im3g	*d*if3g	*n*ews4g
Number of documents	1749	1670	2382
Number of clusters	3	3	4
Vocabulary size	15,103	15,491	18,143

$$NMI = \frac{\sum_{h,l} n_{h,l} \log \frac{n \cdot n_{h,l}}{n_h n_l}}{\sqrt{(\sum_h n_h \log \frac{n_h}{n})(\sum_l n_l \log \frac{n_l}{n})}}$$

where n is the number of all documents, n_h is the number of documents in class h, n_l is the number of documents in cluster l, and $n_{h,l}$ is the number of documents in both class$h$ and cluster l. The *NMI* range is [0,1], where a value of 1 denotes a perfect match between clusters and reference classes.

Besides the *6conf* data described above, we experimented on three new datasets generated from the standard 20-newsgroups data[4]: *sim3g*, *dif3g*, and *news4g*. The *sim3g* consists of 1749 documents from 3 newsgroups on similar topics (comp.graphics,

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

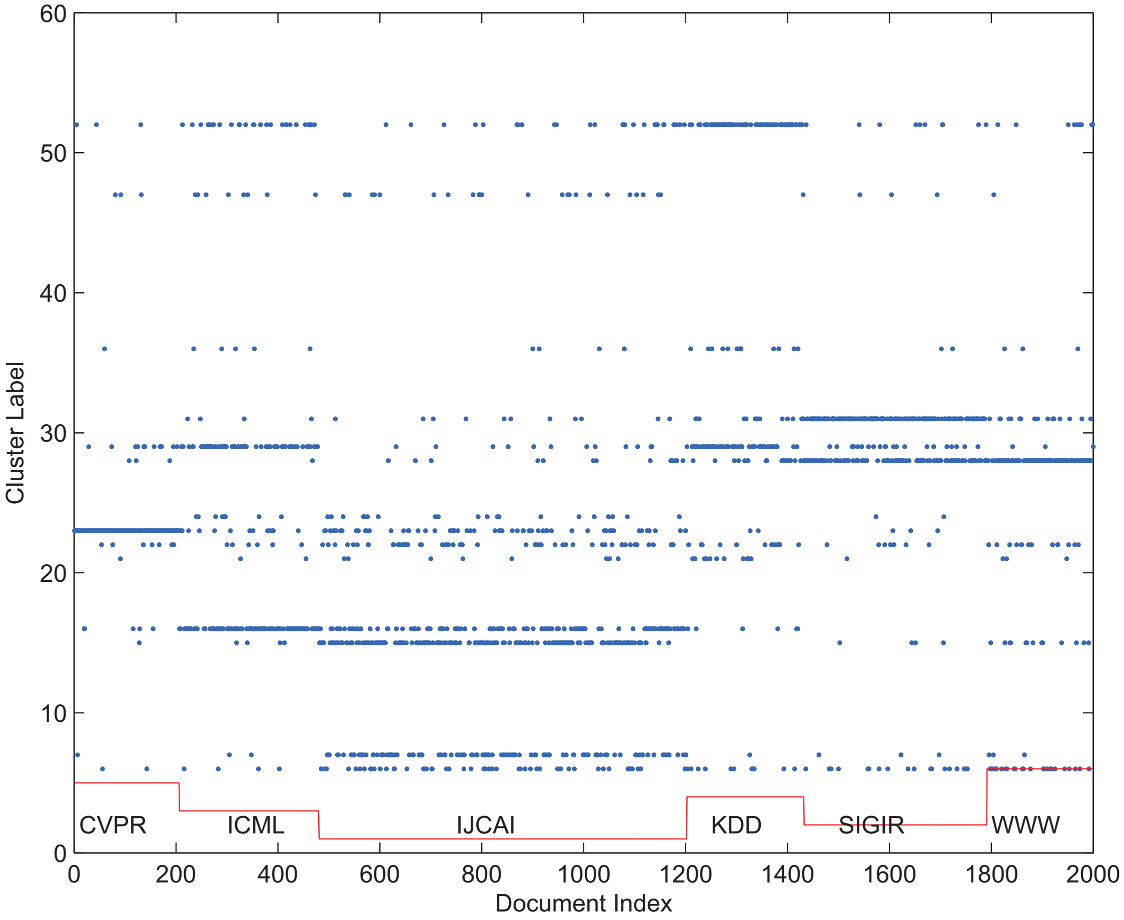


Figure 4.8: Visualization of the clustering results on *6conf* data. The red lines indicate the real labels, while the blue points indicate the clustering assignments.

comp.os.ms-windows, comp.windows.x); *dif3g* contains 3 newsgroups on different topics; and *news4g* has 4 newsgroups involving both similar and different topics (rec.autos, rec.motorcycles, rec.sport.baseball, and sci.med). These three new datasets are summarized in Table 4.3.

To demonstrate the advantage of our model, we compared it with the NDP (see section 4.4.6). The *NMI* values of the clustering are shown in Figure 4.9. In the figure, hNHDP performed consistently better than NDP on all datasets. Similar to general feature selection techniques in clustering, the hierarchical extension and

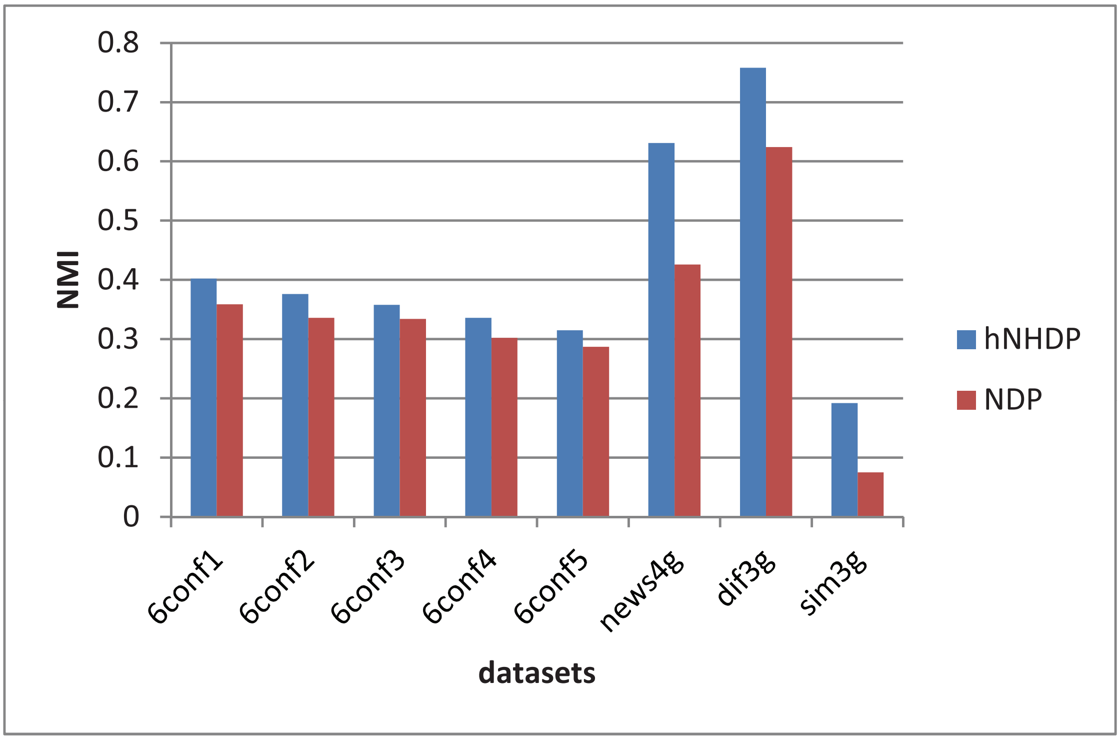


Figure 4.9: Clustering comparison of hNHDP and NDP.

word differentiation, which discriminates clusters only by local topics and local words, improved the clustering quality.

4.4.7 Conclusions

We proposed an extension to the HDP model for modeling groups of data by taking advantage of the latent category information of groups. The hNHDP model clusters the groups and also allows the clusters to share mixture components. The application of the hNHDP to topic modeling illustrates the power of the new prior and provides a way to summarize the document structures. We identify both local topics and local words in the model and discover the implicit document and topic structures. This work can be seen as a potential alternative to document summarization, as well as an improvement to Baysian nonparametric topic models which can be used as a preprocessing step in document summarization.

In addition to document modeling, the hNHDP can also be used for other applications, such as multi-level clustering of patients and hospitals. Moreover, the global exponents can be replaced by some context information, leading to some context-based models. Important future work includes enhancing the computation efficiency.

Chapter 5

Bayesian Nonparametric Summarization and Summary Length Determination

5.1 Background

In previous chapters, we introduced our global optimization approach to multi-document summarization. All the summaries are generated under some length limitation (e.g. 100 words). In fact, in most of the existing summarization systems, people need to first define a constant length to restrict all the output summaries. However, in many cases it is improper to require all summaries are of the same length. Take the multi-document summarization as an example, generating the summaries of the same length for a 5-document cluster and a 50-document cluster is intuitively improper. More specifically, consider two different clusters of documents: one cluster contains very similar articles which all focus on the same event at the same time; the other contains different steps of the event but each step has its own topics. The former cluster may need only one or two sentences to explain its information, while the latter needs to include more.

Research on summary length dates back in the late 90s. Goldstein et al. [29] studied the characteristics of a good summary (single-document summarization for news) and showed an empirical distribution of summary length over document size. However, the length problem has been gradually ignored later, since researchers need to fix the length so as to estimate different summarization models conveniently. A typical instance is the Document Understanding Conferences (DUC)[1], which provide authoritative evaluation for summarization systems. The DUC conferences collect news aritcles as the input data and define various summarization tasks, such as generic multi-document summarization, query-focused summarization and update summarization. In all the DUC tasks, the output is restricted within a length. Then human-generated summaries are provided to evaluate the results of different summarization systems. Limiting the length of summaries contributed a lot to the development of summarization techniques, but as we discussed before, in many cases keeping the summaries of the same size is not a good choice.

Moreover, even in constant-length summarization, how to define a proper size of summaries for the summarization tasks is quite a problem. Why does DUC2007 main task require 250 words while Update task require 100 words? Is it reasonable? A short summary may sacrifice the coverage, while a long summary may cause redundance. Automatically determining the best size of summaries according to the input documents is valuable, and it may deepen our understanding of summarization.

In this chapter, we aim to find the proper length for document summarization automatically and generate varying-length summaries based on the document itself. The varying-length summarization is more robust for unbalanced clusters. It can also provide a recommended size as the predefined summary length for general constant-

[1] After 2007, the DUC tasks are incorporated into the Text Analysis Conference (TAC).

length summarization systems. We advance a Bayesian nonparametric model of extractive multi-document summarization to achieve this goal [57]. As far as we are concerned, it is the first model that can learn appropriate lengths of summaries.

Bayesian nonparametric (BNP) methods are powerful tools to determine the size of latent variables [27]. They let the data ”speak for itself” and allow the dimension of latent variables to grow with the data. In order to integrate the BNP methods into document summarization, we follow the assumption that the original documents should be recovered from the reconstruction of summaries [59, 35]. We use the Beta process as a prior to generate binary vectors for selecting active sentences that reconstruct the original documents. Then we construct a Bayesian framework for summarization and use the variational approximation for inference. Experimental results on DUC2004 dataset demonstrate the effectiveness of our model. Besides, we reorganize the original documents to generate some new datasets, and examine how the summary length changes on the new data. The results prove that our summary length determination is rational and necessary on unbalanced data.

5.2 Related Work

5.2.1 Research on Summary Length

Summary length is an important aspect for generating and evaluating summaries. Early research on summary length [29] focused on discovering the properties of human-generated summaries and analyzing the effect of compression ratio. It demonstrated that an evaluation of summarization systems must take into account both the compression ratios and the characteristics of the documents. Radev and Fan [89] compared the readability and speedup in reading time of 10% summaries and 20%

summaries[2] for topic sets with different number of documents. Sweeney et al. [90] developed an incremental summary containing additional sentences that provide context. Kaisser et al. [41] studied the impact of query types on summary length of search results. Other than the content of original documents, there are also some other factors affecting summary length especially in specific applications. For example, Sweeney and Crestani [89] studied the relation between screen size and summary length on mobile platforms. The conclusion of their work is the optimal summary size always falls into the shorter one regardless of the screen size.

In sum, the previous works on summary length mostly put their attention on the empirical study of the phenomenon, factors and impacts of summary length. None of them automatically find the best length, which is our main task in this chapter. Nevertheless, they demonstrated the importance of summary length in summarization and the reasonability of determining summary length based on content of news documents [29] or search results [41]. As our model is mainly applied for generic summarization of news articles, we do not consider the factor of screen size in mobile applications.

5.2.2 BNP Methods in Document Summarization

Bayesian nonparametric methods provide a Bayesian framework for model selection and adaptation using nonparametric models [27]. A BNP model uses an infinite-dimensional parameter space, but invokes only a finite subset of the available parameters on any given finite data set. This subset generally grows with the data set. Thus BNP models address the problem of choosing the number of mixture components or latent factors. For example, the hierarchical Dirichlet process (HDP) can be used to infer the number of topics in topic models or the number of states in the infinite Hidden Markov model [92].

[2] 10% and 20% are the compression rates, and the documents are from search results in information retrieval systems.

Recently, some BNP models are also involved in document summarization approaches [11, 17]. BNP priors such as the nested Chinese restaurant process (nCRP) are associated with topic analysis in these models. Then the topic distributions are used to get the sentence scores and rank sentences. BNP here only impacts the number and the structure of the latent topics, but the summarization framework is still constant-length. Our BNP summarization model differs from the previous models. Besides using the HDP for topic analysis, our approach further integrates the beta process into sentence selection. The BNP method in our model are directly used to determine the number of summary sentences but not latent topics.

5.3 BNP Summarization

In section 4.3.5, we introduced the Beta processes which are often used for latent factor/feature analysis. Here we integrate the BP processes into our model, the BNP summarization.

5.3.1 Framework of BNP Summarization

Most existing approaches for generic extractive summarization are based on sentence ranking. However, these methods suffer from a severe problem that they cannot make a good trade-off between the coverage and minimum redundancy [35]. Some global optimization algorithms are developed, instead of greedy search, to select the best overall summaries [70]. One approach to global optimization of summarization is to regard the summarization as a reconstruction process [59, 35] . Considering a good summary must catch most of the important information in original documents, the original documents are assumed able to be recovered from summaries with some information loss. Then the summarization problem is turned into finding the sentences that cause the least reconstruction error (or information loss). In this paper, we follow the assumption and formulate summarization as a Bayesian framework.

First we review the models of Chapter 3 and [35]. Given a cluster of M documents $D = d_1, d_2, ..., d_M$ and the sentence set contained in the documents as $X = [x_1, x_2, ..., x_N]$, we denote all corresponding summary sentences as $S = [s_1, ..., s_n]$, where n is the number of summary sentences and N is the number of all sentences in the cluster. A document d and a sentence s or x here are all represented by weighted term frequency vectors in the space $\mathbb{R}^V$, where V is the number of total terms (words).

Following the reconstruction assumption, a candidate sentence x_i can be approximated by the linear combination of summary sentences: $x_i \simeq \sum_{j=1}^{n} w'_j s_j$, where w'_j is the weight for summary sentence s_j. Thus the document can also be approximately represented by a linear combination of summary sentences because it is the sum of the sentences.

$$d_i \simeq \sum_{j=1}^{n} w_j s_j. \tag{5.1}$$

Then the work in [35] aims to find the summary sentence set that can minimize the reconstruction error $\sum_{i=1}^{N} ||x_i - \sum_{j=1}^{n} w'_j s_j||^2$; while the linear representation model in Chapter 3 defines the problem as finding the sentences that minimize the distortion between documents and its reconstruction $dis(d_i, \sum_{j=1}^{n} w_j s_j)$ where this distortion function can also be a squared error function.

Now we consider the reconstruction for each document, if we see the document d as an dependent variable, and the summary sentence set S as the independent variable, the problem to minimize the reconstruction error can be seen as a linear regression model. The model can be easily changed to a Bayesian regression model by adding a zero-mean Gaussian noise ϵ [5], as follows.

$$d_i = \sum_{j=1}^{n} w_j s_j + \epsilon_i \tag{5.2}$$

where the weights w_j are also assigned a Gaussian prior.

The next step is sentence selection. As our system is an extractive summarization model, all the summary sentences are from the original document cluster. So we can use a binary vector $z_i =< z_{i1}, ..., z_{iN} >^T$ to choose the active sentences S (i.e. summary sentences) from the original sentence set X. The Equation (5.2) is turned into $d_i = \sum_{j=1}^{N} \phi_{ij} * z_{ij} x_j + \epsilon_i$. Using a beta process as a prior for the binary vector z_i, we can automatically infer the number of active component associated with z_i. As to the weights of the sentences, we use a random vector ϕ_i which has the multivariate normal distribution because of the conjugacy. $\phi_i \in \mathbb{R}^N$ is an extension to the weights $\{w_1, ...w_n\}$ in (5.2).

Integrating the *linear reconstruction* (5.2) and the *beta process*[3] (4.7), we get the complete process of summary sentence selection as follows.

$$\begin{aligned} d_i &= X(\phi_i \circ z_i) + \epsilon_i \\ X &= [x_1, x_2, ..., x_N] \\ z_{ij} &\sim Bernoulli(\pi_j) \\ \pi_j &\sim Beta(\frac{\alpha\gamma}{N}, \alpha(1 - \frac{\gamma}{N})) \\ \phi_i &\sim \mathcal{N}(0, \sigma_\phi^2 I) \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_\epsilon^2 I) \end{aligned} \tag{5.3}$$

where N is the number of sentences in the whole document cluster. The symbol $\circ$ represents the elementwise multiplication of two vectors.

One problem of the reconstruction model is that the word vector representation of the sentences are sparse, which dramatically increase the reconstruction error. So we bring in topic models to reduce the dimension of the data. We use a HDP-

[3] We use the finite approximation because the number of sentences is large but finite

LDA [92] to get topic distributions for each sentence, and we represent the sentences and documents as the topic weight vectors instead of word weight vectors. Finally d_i is a K-dimensional vector and X is a $K*N$ matrix, where K is the number of topics in topic models.

5.4 Variational Inference

In this section, we derive a variational Bayesian algorithm for fast inference of our sentence selection model. Variational inference [5] is a framework for approximating the true posterior with the best from a set of distributions Q : $q* = \arg\min_{q \in Q} KL(q(Z)|p(Z|D))$. Suppose $q(Z)$ can be partitioned into disjoint groups denoted by Z_j, and the q distribution factorizes with respect to these groups: $q(Z) = \prod_{j=1}^{M} q(Z_j)$. We can obtain a general expression for the optimal solution $q_j^*(Z_j)$ given by

$$\ln q_j^*(Z_j) = \mathbb{E}_{i \neq j}[\ln p(D, Z)] + const. \tag{5.4}$$

where $\mathbb{E}_{i \neq j}[\ln p(D, Z)]$ is the expectation of the logarithm of the joint probability of the data and latent variables, taken over all variables not in the partition. We will therefore seek a consistent solution by first initializing all of the factors $q_j(Z_j)$ appropriately and then cycling through the factors and replacing each in turn with a revised estimate given by (5.4) evaluated using the current estimates for all of the other factors.

Update for Z

$$p(z_{ij}|\pi_j, d_i, X, \phi_i) \propto p(d_i|z_{ij}, x_j, \phi_i)p(z_{ij}|\pi_j)$$

We use $q(z_{ij})$ to approximate the posterior:

$$\begin{aligned}
&q(z_{ij})\\
&\propto \exp\{\mathrm{E}[\ln(p(d_i|z_{ij}, z_i^{-j}, X, \phi_i)) + \ln(p(z_{ij}|\pi))]\}\\
&\propto \exp\{\mathrm{E}[\ln(\pi_j)]\}*\\
&\exp\{\mathrm{E}[-\frac{1}{2\sigma_\epsilon^2}\left(d_i^{-j} - x_j z_{ij}\phi_{ij}\right)^T\left(d_i^{-j} - x_j z_{ij}\phi_{ij}\right)]\}\\
&\propto \exp\{\overline{\ln(\pi_j)}\}*\\
&\exp\{-\frac{\left(\overline{\phi_{ij}^2} * \overline{z_{ij}^2} * \overline{x_j^T x_j} - 2\overline{\phi_{ij}} * \overline{z_{ij}} * \overline{x_j}^T * \overline{d_i^{-j}}\right)}{2\sigma_\epsilon^2}\}
\end{aligned} \tag{5.5}$$

where $d_i^{-j} = d_i - X^{-j}(\phi_i^{-j} \circ z_i^{-j})$, and the symbol $^-$ indicates the expectation value. The $\overline{\phi_{ij}^2}$ can be extended to this form:

$$\overline{\phi_{ij}^2} = \overline{\phi_{ij}}^2 + \Delta_i^j \tag{5.6}$$

where Δ_i^j means the j^{th} diagonal element of Δ_i which is defined by Equation 5.11.

As z_i is a binary vector, we only calculate the probability of $z_{ij} = 1$ and $z_{ij} = 0$.

$$\begin{aligned}
&q(z_{ij} = 1) \propto \exp\{\overline{\ln(\pi_j)}\} *\\
&\exp\{-\frac{1}{2\sigma_\epsilon^2}\left(\overline{\phi_{ij}^2} * \overline{x_j^T x_j} - 2\overline{\phi_{ij}} * \overline{x_j}^T * \overline{d_i^{-j}}\right)\}\\
&q(z_{ij} = 0) \propto \exp\{\overline{\ln(1 - \pi_j)}\}
\end{aligned} \tag{5.7}$$

The expectations can be calculated as

$$\overline{\ln(\pi_j)} = \varphi(\frac{\alpha\gamma}{N} + \overline{n_j}) - \varphi(\alpha + M) \tag{5.8}$$

$$\overline{\ln(1 - \pi_j)} = \varphi(\alpha(1 - \frac{\gamma}{N}) + M - \overline{n_j}) - \varphi(\alpha + M) \tag{5.9}$$

where $\overline{n_j} = \sum_{i=1}^{M} z_{ij}$.

Update for π

$$p(\pi_j|Z) \propto p(\pi_j|\alpha, \gamma, N)p(Z|\pi_j)$$

Because of the conjugacy of the beta to Bernoulli distribution, the posterior of π is still a beta distribution:

$$\pi_j \sim Beta(\frac{\alpha\gamma}{N} + \overline{n_j}, \alpha(1 - \frac{\gamma}{N}) + M - \overline{n_j}) \tag{5.10}$$

Update for Φ

$$p(\phi_i|d_i, Z, X) \propto p(d_i|\phi_i, Z, X)p(\phi_i|\sigma_\phi^2)$$

The posterior is also a normal distribution with mean μ_i and covariance Δ_i.

$$\Delta_i = \left(\frac{1}{\sigma_\epsilon^2} \overline{\tilde{X}_i^T \tilde{X}_i} + \frac{1}{\sigma_\phi^2} I \right)^{-1} \tag{5.11}$$

$$\mu_i = \Delta_i \left(\frac{1}{\sigma_\epsilon^2} \overline{\tilde{X}_i}^T d_i \right) \tag{5.12}$$

Here $\tilde{X}_i \equiv X \circ \tilde{z}_i$ and $\tilde{z}_i \equiv [z_i, ..., z_i]^T$ is a $K \times N$ matrix with the vector z_i repeated K(the number of the latent topics) times.

$$\overline{\tilde{X}_i} = X * \overline{\tilde{z}_i} \tag{5.13}$$

$$\overline{\tilde{X}_i^T \tilde{X}_i} = (X^T X) \circ (\overline{z_i} * \overline{z_i}^T + Bcov_i) \tag{5.14}$$

$$Bcov_i = \text{diag}[\overline{z_{i1}}(1 - \overline{z_{i1}}), ..., \overline{z_{iN}}(1 - \overline{z_{iN}})] \tag{5.15}$$

Update for σ_ϵ^2

$$p(\sigma_\epsilon^2|\Phi, D, Z, X) \propto p(D|\Phi, Z, X, \sigma_\epsilon^2)p(\sigma_\epsilon^2)$$

By using a conjugate prior, inverse gamma prior $InvGamma(u, v)$, the posterior can be calculated as a new inverse gamma distribution with parameters

$$\begin{aligned} u' &= u + MK/2 \\ v' &= v + \frac{1}{2}\sum_{i=1}^{M}(||d_i - X(\overline{z_i} \circ \overline{\phi_i})|| + \xi_i) \end{aligned} \tag{5.16}$$

where

$$\begin{gathered} \xi_i = \textstyle\sum_{j=1}^{N}(\overline{z_{ij}^2} * \overline{\phi_{ij}^2} * x_j^T x_j - \overline{z_{ij}}^2 * \overline{\phi_{ij}}^2 * x_j^T x_j) \\ + \textstyle\sum_{j \neq l} \overline{z_{ij}} * \overline{z_{il}} * \Delta_{i,jl} * x_j^T x_l \end{gathered}$$

Update for σ_ϕ^2

$$p(\sigma_\phi^2|\Phi) \propto p(\Phi|\sigma_\phi^2)p(\sigma_\phi^2)$$

By using a conjugate prior, inverse gamma prior $InvGamma(e, f)$, the posterior can be calculated as a new inverse gamma distribution with parameters

$$\begin{aligned} e' &= e + MN/2 \\ f' &= f + \frac{1}{2}\sum_{i=1}^{M}\left((\overline{\Phi})^T\overline{\Phi} + trace(\Delta_i')\right) \end{aligned} \tag{5.17}$$

5.5 Experiments

To test the capability of our BNP summarization systems, we design a series of experiments. The aim of the experiments mainly includes three aspects:

1. To demonstrate the summaries extracted by our model have good qualities and the summary length determined by the model is reasonable.

2. To give examples where varying summary length is necessary.

3. To observe the distribution of summary length.

We evaluate the performance on the dataset of DUC2004 task2. The data contains 50 document clusters, with 10 news articles in each cluster. Besides, we construct three new datasets from the DUC2004 dataset to further prove the advantage of variable-length summarization. We separate each cluster in the original dataset into two parts where each has 5 documents, hence getting the *Separate* Dataset; Then we randomly combine two original clusters in the DUC2004 dataset, and get two datasets called *Combined1* and *Combined2*. Thus each of the clusters in the combined datasets include 20 documents with two different themes.

5.5.1 Evaluation of Summary Qualities

First, we implement our BNP summarization model on the DUC2004 dataset, with summary length not limited. At the topic analysis step, we use the HDP model and follow the inference in [92]. For the sentence selection step, we use the variational inference described in Section 5.4, where the parameters in the beta process (5.3) are set as $\gamma = 1, \alpha = 1$. The summaries that we finally generate have an average length of 164 words. We design several popular unsupervised summarization systems and compare them with our model.

- The *Random* model selects sentences randomly for each document cluster.

- The *MMR* [10] strives to reduce redundancy while maintaining relevance. For generic summarization, we replace the query relevance with the relevance to documents.

- The *Lexrank* model [22] is a graph-based method which choose sentences based on the concept of eigenvector centrality.

- The *Linear* Representation model [59] has the same assumption as ours and it can be seen as an approximation of the constant-length version of our model.

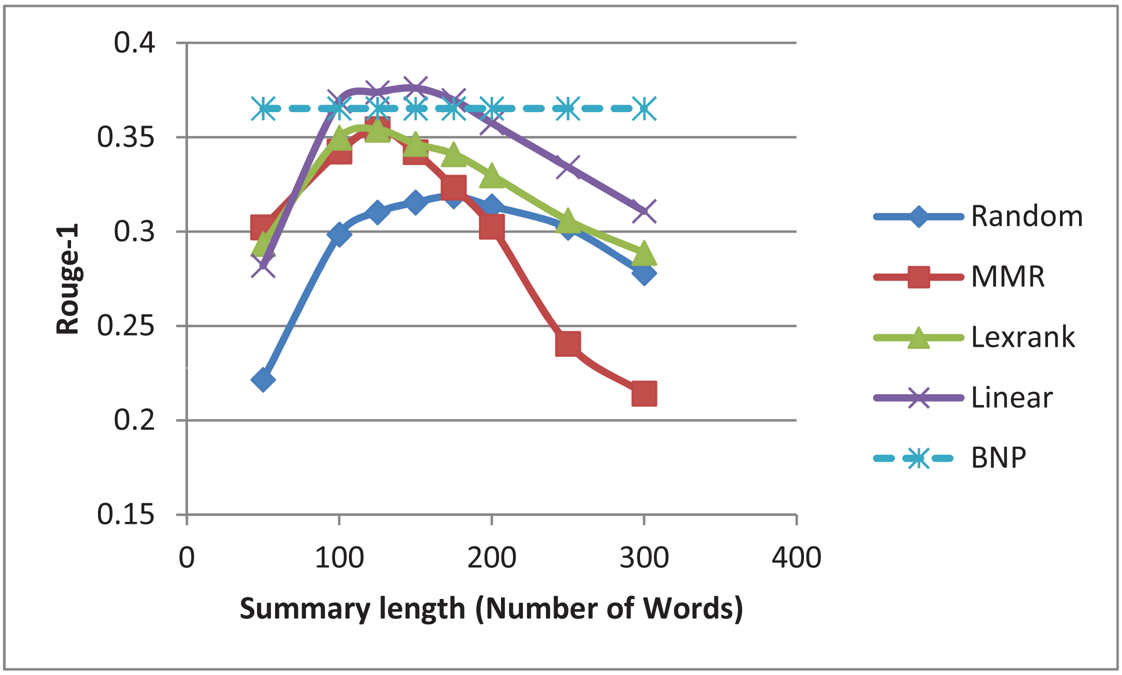


Figure 5.1: Rouge-1 values on DUC2004 dataset.

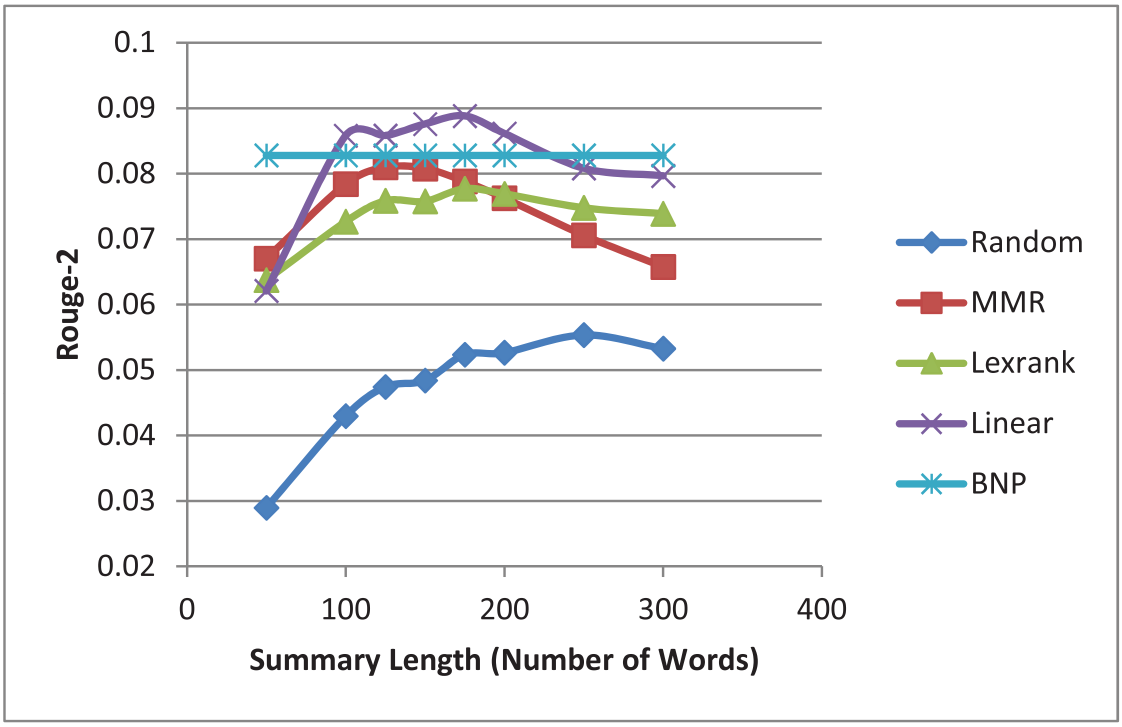


Figure 5.2: Rouge-2 values on DUC2004 dataset.

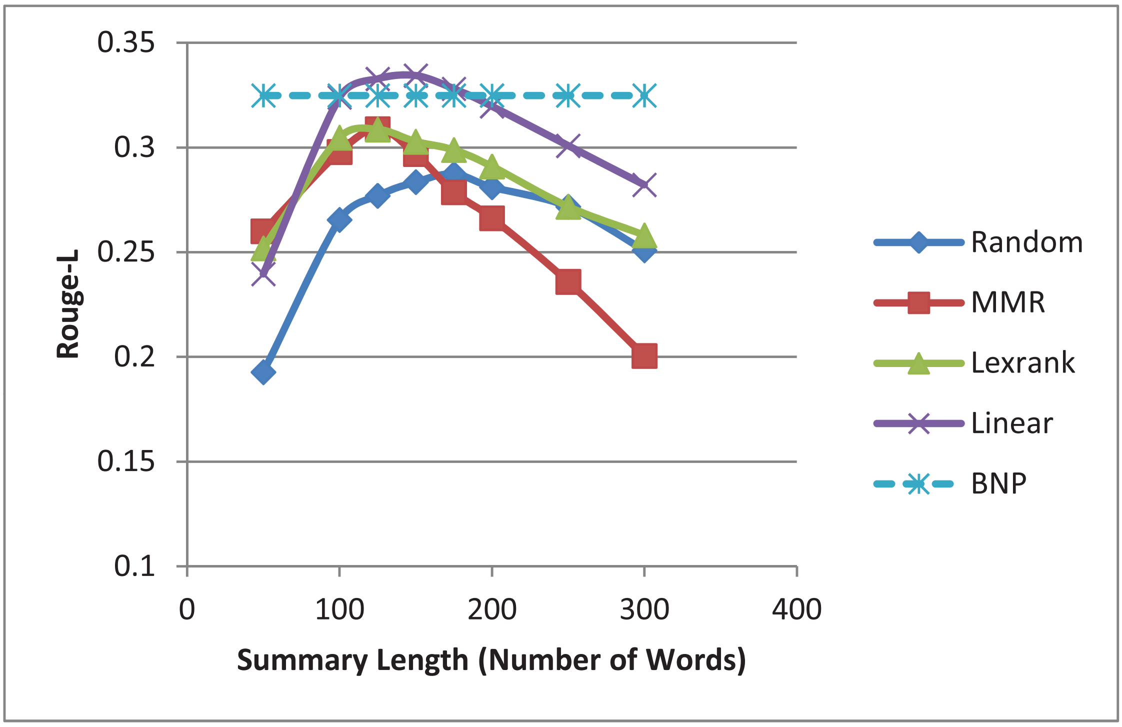


Figure 5.3: Rouge-L values on DUC2004 dataset.

All the compared systems are implemented at different predefined lengths from 50 to 300 words. Then we evaluate the summaries with ROUGE[4] tools [48] in terms of the f-measure scores of Rouge-1 Rouge-2, and Rouge-L. The metric of Rouge f-measure takes into consideration the summary length in evaluation, so it is proper for our experiments. From Fig.5.1, Fig.5.2 and Fig.5.3, we can see that the result of BNP summarization (the dashed line) gets the second best value among all systems. It is only defeated by the *Linear* model but the result is comparable to the best in Fig.5.1 and Fig.5.3; while it exceeds other systems at all lengths. This proves the good qualities of our BNP summaries. The reason that the *Linear* system gets a little better result may be its weights for linear combination of summary sentences are guaranteed nonnegative while in our model the weights are zero-mean Gaussian

[4] we use ROUGE1.5.5 in this work.

variables. This may lead to less redundance in sentence selection for the *Linear* Representation model.

Turn to the length determination. We take advantage of the *Linear* Representation model to approximate the constant-length version of our model. Comparing the summaries generated at different predefined lengths, Fig.5.4 shows the the model gets the best performance (Rouge values) at the length around 164 words, the length learned by our BNP model. This result partly demonstrates our length determination is rational and it can be used as the recommended length for some constant-length summarization systems, such as the *Linear* .

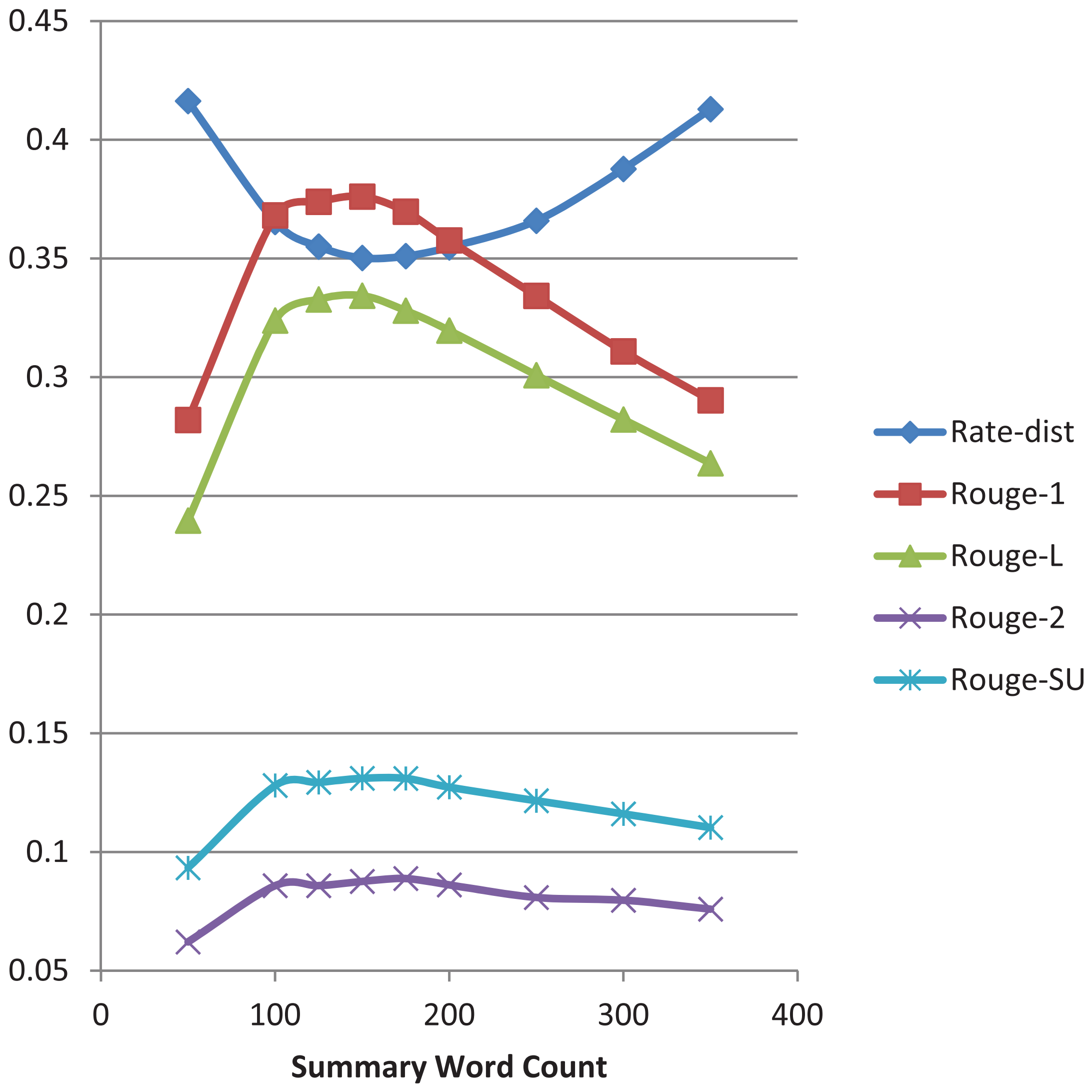


Figure 5.4: Rate-dist value V.S. summary word length.

5.5.2 A New Evaluation Metric

The Rouge evaluation requires golden standard summaries as the base. However, in many cases we cannot get the reference summaries. For example, when we implement experiments on our expanded datasets (the separate and combined clusters of documents), we do not have exact reference summaries. Louis and Nenkova [55] advanced an automatic summary evaluation without human models. They used the

Jensen-Shannon divergence(JSD) between the input documents and the summaries as a feature, and got high correlation with human evaluations and the rouge metric. Unfortunately, it was designed for comparison at a constant-length, which cannot meet our needs. To extend the JSD evaluation to compare varying-length summaries, we propose a new measure based on information theory, the rate-distortion [15].

Rate-Distortion: The distortion function $d(x, \hat{x})$ is a measure of the cost of representing the symbol x to a new symbol $\hat{x}$; and the rate can indicate how much compression can be achieved. The problem of finding the minimum rate can be solved by minimizing the functional

$$\mathcal{F}[p(\hat{x}|x)] = I(X; \hat{X}) + \beta \mathbb{E}(d(x, \hat{x})). \tag{5.18}$$

where $I(X; \hat{X})$ denotes the mutual information. The rate-distortion theory is a fundamental theory for lossy data compression. Recently, it has also been successfully employed for text clustering [85] and document summarization [59]. Slonim [85] claims that the mutual information $I(X; \hat{X})$ measures the compactness of the new representation. Thus the rate-distortion function is a trade-off between the compactness of new representation and the expected distortion. Specifically in summarization, the summaries can be seen as the new representation $\hat{X}$ of original documents X. A good summary balances the compression ratio and the information loss, thus minimizing the function (5.18). So we use the function (5.18)(we set $\beta = 1$) to compare which summary is a better compression. The JS-divergence (JSD), which has been proved to have high correlation with manual evaluation [55] for constant-length summary evaluation, is utilized as the distortion in the function. In the following sections, we simply call the values of the function (5.18) *rate-dist*. In fact, the rate-dist values can be seen as the JSD measure with length regularization.

To check the effectiveness of rate-dist measure, we evaluate all summaries generated in Section 5.5.1 with the new measure (the lower the better). Fig. 5.5 shows that the results accord with the ones in Fig. 5.1 and Fig. 5.3. Moreover, in Fig. 5.4, the curve of rate-dist values has a inverse tendency of Rouge measures (Rouge-1, Rouge-2, Rouge-L and Rouge-SU4 are all listed here), and the best performance also occurs around the summary length of 164 words. This even more clearly reveals that the BNP summarization achieves a perfect tradeoff between compactness and informativeness. Due to the accordance with rouge measures, it is promising to be regarded as an alternative to the rouge measures in case we do not have reference summaries.

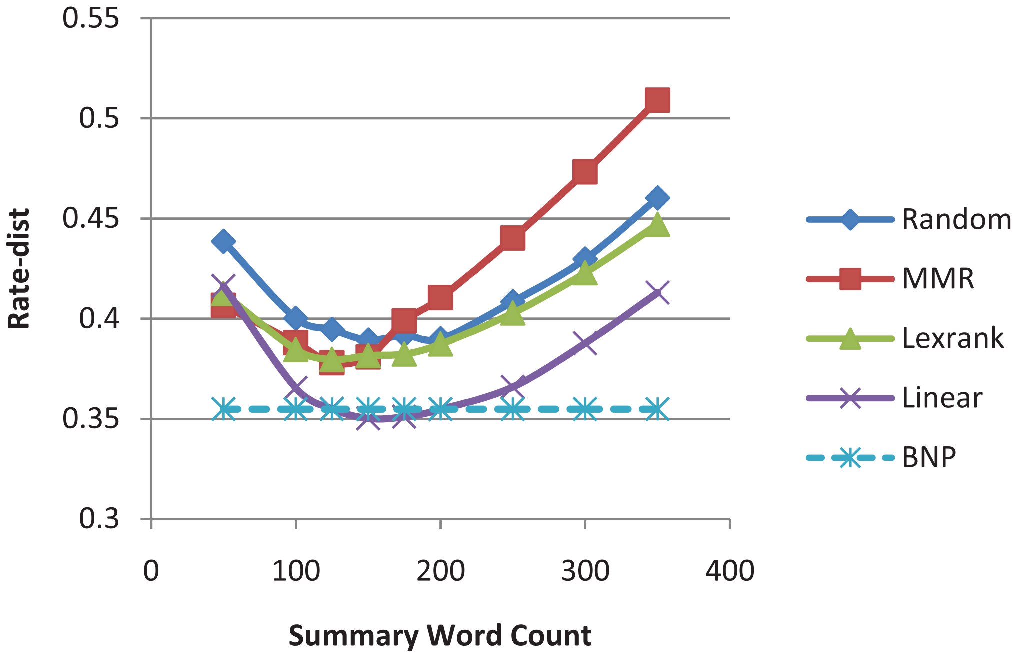


Figure 5.5: Comparison of BNP Summarization with other systems using rate-dist measure.

5.5.3 Necessity of Varying Summary Length

In this section, we discuss the necessity of length determination and how summary length changes according to the input data. As explained before, we generate three new datasets from the original DUC2004 dataset. Now we use them to indicate varying summary length is necessary when the input data varies a lot.

Table 5.1 shows the average summary length of different data sets. The results satisfy the intuitive expectation of summary length change. When we split a 10-document cluster into two 5-document parts, we expect the average summary length of the new clusters to be a little smaller than the original cluster but much larger than half of the original length, because all the documents concentrate on the same themes. When we combine two clusters into one, the summary length should be smaller than the sum of the summary lengths of two original clusters due to some unavoidable common background information but much larger than the summary length of original clusters.

Original	Separate	Combined1	Combined2
164	115	250	231

Table 5.1: Average summary length (number of words) on different datasets

We also run the Linear Representation system at different lengths on the new datasets and evaluate the qualities. As we do not have golden standard for the new datasets, so we only use the rate-dist measure here. Results in Table 5.2,5.3,5.4 show the summaries which do not change the predefined length [5] perform significantly worse than the BNP summarization. All the comparison is statistically significant. So varying summary length is necessary when the input changes a lot, and our model can just give a good match to the new data. This characteristic also can be used to give recommended summary length for extractive summarization systems when given unknown data.

Then we observe the summary length distributions and compression ratios according to document size(the length of the whole documents in a cluster). The average summary length increases (Fig. 5.6), while the compression ratios decreases (Fig. 5.7)

[5] 665 bytes is the DUC2004 requirement and 164 words is the best length on original data

	Predefined	Unchanged	BNP
Length	665 bytes	164 words	115 words
Rate-dist	0.4130	0.4404	0.4007

Table 5.2: Comparison of summary lengths on Separate Dataset.

	Predefined	Unchanged	BNP
Length	665 bytes	164 words	250 words
Rate-dist	0.3768	0.3450	0.3238

Table 5.3: Comparison of summary lengths on Combined1 Dataset.

as document size grows. The rule of the compression ratio here agrees with the rule in [29], although that work is done for single-document summarization.

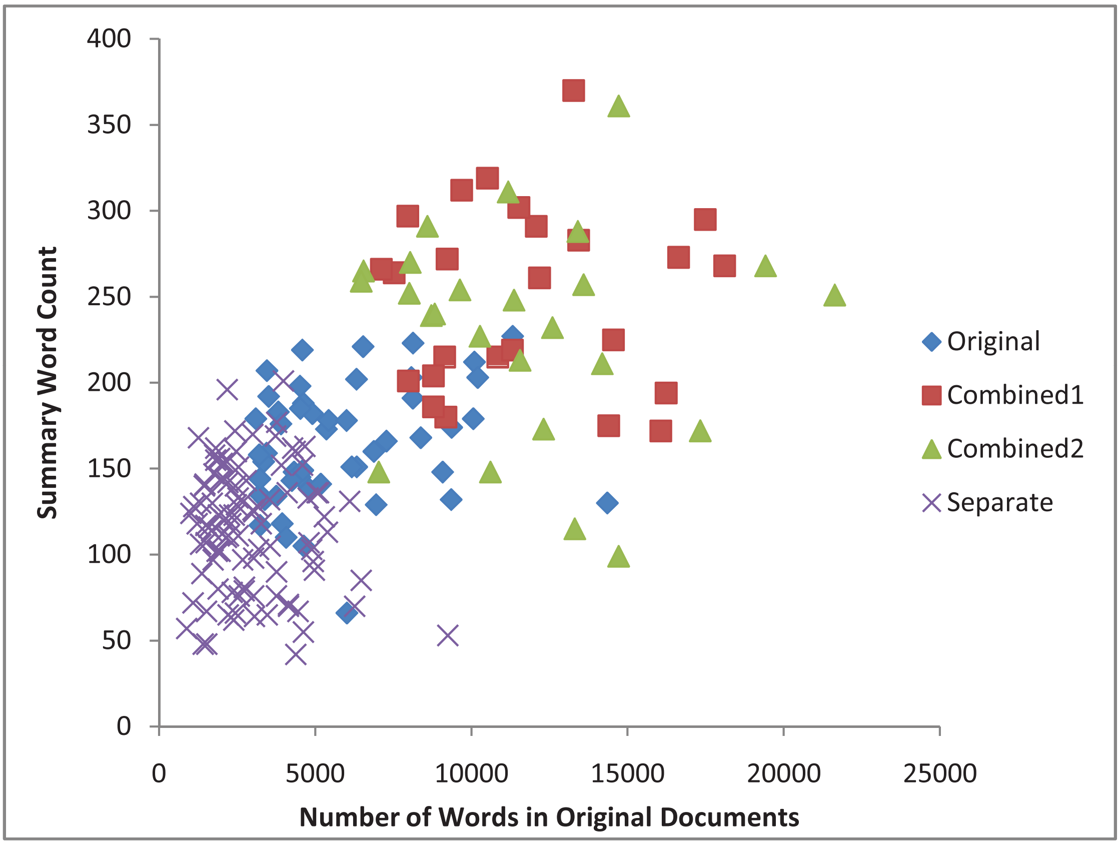


Figure 5.6: The distribution of summary word length.

	Predefined	Unchanged	BNP
Length	665 bytes	164 words	231 words
Rate-dist	0.3739	0.3464	0.3326

Table 5.4: Comparison of summary lengths on Combined2 Dataset.

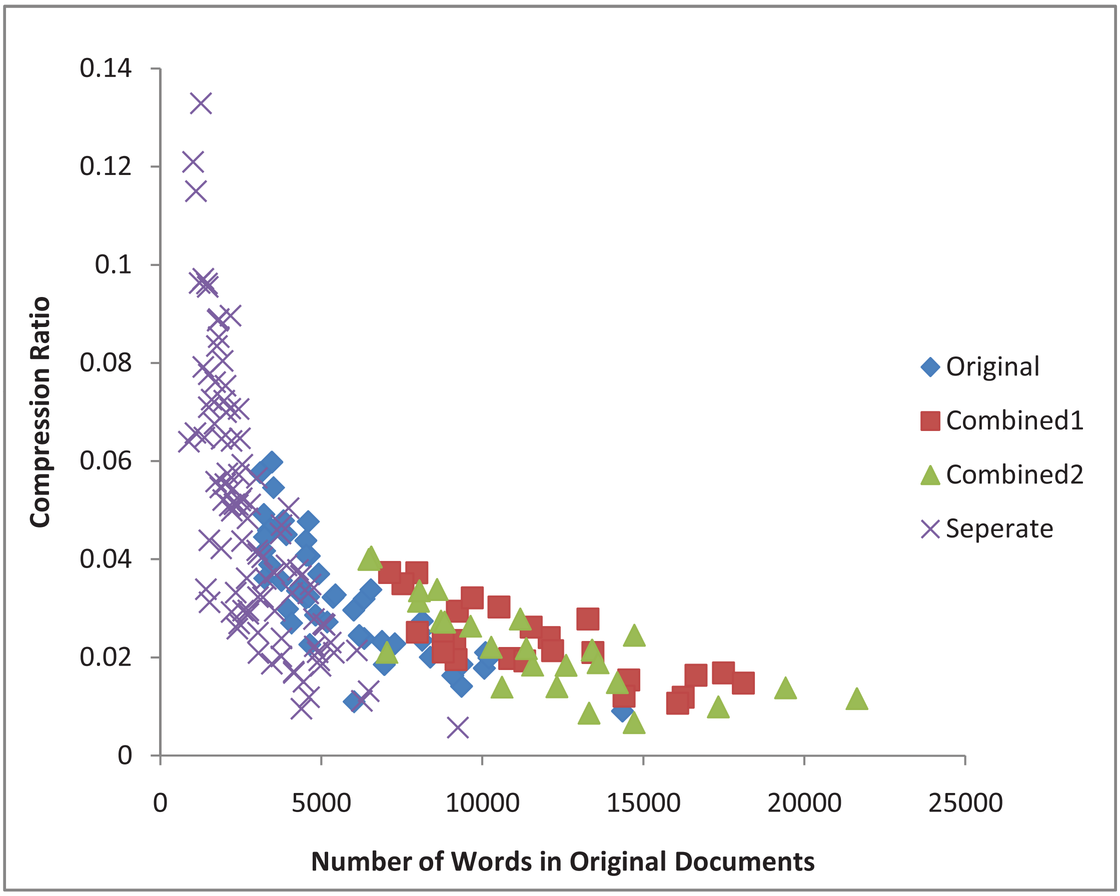


Figure 5.7: Compression ratio versus document word length.

5.6 Conclusion and Future Work

In this paper, we present a new problem of finding a proper summary length for multi-document summarization based on the document content. A Bayesian nonparametric model is proposed to solve this problem. We use the beta process as the prior to construct a Bayesian framework for summary sentence selection. Experimental results are shown on DUC2004 dataset, as well as some expanded datasets. We demonstrate the summaries we extract have good qualities and the length determination of our system is rational.

However, there is still much work to do for variable-length summarization. First, Our system is extractive-base summarization, which cannot achieve the perfect coherence and readability. A system which can determine the best length even for abstractive summarization will be better. Moreover, in this work we only consider the aspect of data compression and evaluate the performance using an information-theoretic measure. In future we may consider more human factors, and prove the summary length determined by our system agrees with human preference. In addition, in the experiments, we only use the imbalanced datasets as the example that intuitively needs varying the summary length. However, the data type is also important to impact the summary length. In future, we may extend the work by studying more cases that need varying summary length.

Chapter 6

Conclusion

In this thesis, we proposed two new algorithms for document summarization: a reconstruction based optimization approach and a Bayesian nonparametric approach. We also improved the current Bayesian nonparametric methods in a related research area, topic modeling, which is an effective preprocessing step of or an alternative to document summarization.

In Chapter 3, we assumed that a good summary should contain most of the important information in the original documents, thus the original documents should be reconstructed by the best summary with the least information loss. We first built optimization systems from the information-theoretic perspective. Different from former sentence-ranking algorithms, our system selected sentences globally by regarding summarization as solving an optimization problem. We designed several reconstruction strategies and defined different distortion measures to evaluate the goodness of reconstruction, deriving finally a flexible and well-performed summarization approach, namely Linear Representation.

The reconstruction-based summarization framework was then extended to a Bayesian nonparametric approach to solving the summary length problem in Chapter 5. We aimed at automatically determining summary length, which was often ignored in traditional summarization systems but formed one of the important factors of summarization. Following the summarization framework of Linear Representation, we integrated a Bayesian nonparametric prior, the Beta process, into the reconstruction from summaries to original documents. We borrowed the power of Bayesian nonparametric in model selection, for summary length determination. The number of summary sentences was automatically determined by posterior inference. The generated variable-length summaries were demonstrated that they had good qualities as well as proper lengths.

Topic representation is an important step in BNP summarization to avoid the word sparseness. Chapter 4 provides the background information for topic modelling as well as Bayesian nonparametrics and their relationships with document summarization. Moreover, In the end of Chapter 4 we proposed a new Bayesian nonparametric topic model, namely, the hybrid nested hierarchial Dirichlet process. The hNHDP model deals with the case that the documents are well organized but we do not know the categories of documents. It could differentiate specific topics and words from common topics and words. It could be used for a robust topic analysis for the original documents in future summarization work. Its visualization of the topic structures and document structures could also be seen as an abstractive form of summarization.

Reconstruction and Bayesian nonparametrics based summarization systems provide new aspects for summarization. They have gained great success in traditional summarization tasks as well as in determining the summary length. Besides better representation of documents, future work contains extension to new types of summarization (e.g. multi-lingual summarization) and improving the computational efficiency. As we introduced in Chapter 2, there have been more and more types of summarization in practice. Our summarization frameworks are very flexible to be modified to adapt to the new types. Meanwhile, when we have huge amounts of data, how to accelerate our algorithms deserves more consideration.

Appendix A

Proof of the Distortion Bound 1

Proof: Many-to-one linear representation has no more distortion than one-to-one representation.

First, we demonstrate the "no more than" relationship. Suppose we have an original sentence i, in p-median model, it can be reconstructed by only the median $\hat{x}_i$; while in linear representation model, it is reconstructed by a linear combination of summary sentences $\sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j$.

we can take $\hat{\lambda}_{ii} = 1; \hat{\lambda}_{ij} = 0 (when j \neq i)$ for $\sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j$, thus we could get

$$\min_{\hat{\lambda}_{ij}} d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) \leq d(x_i, \hat{x}_i)$$

.

Then, we demonstrate that in some constraint, the reconstruction error of linear representation is less than one-to-one representation.

We calculate the partial derivative of $f = d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j)$ with respect to $\hat{\lambda}_{ij}$. Here J-S divergence is taken as an example of the distortion measure.

$$\frac{\partial f}{\hat{\lambda}_{ij}} = \sum_{y} \frac{y_{ij}}{Count} \log \frac{\hat{\lambda}_{ij} y_{ij}}{y_{ij} + \sum_j \hat{\lambda}_{ij} y_{ij}} \leq 0$$

.

Assuming there exists $\hat{\lambda}_{ij}$ such that $\frac{\partial f}{\hat{\lambda}_{ij}} < \frac{\partial f}{\hat{\lambda}_{ii}}$, the new representation $\sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j$ with $\hat{\lambda}_{ii} = 1 - \delta h, \hat{\lambda}_{ij} = \delta h, \hat{\lambda}_{ik} = 0 (k \neq i, j; \delta h > 0)$, have a smaller distortion.

$$d_{Linear} = d_{OneToOne} + (\frac{\partial f}{\hat{\lambda}_{ij}} - \frac{\partial f}{\hat{\lambda}_{ii}})\delta h < d_{OneToOne}$$

Appendix B

Proof of the Distortion Bound 2

Proof: *Distortion between the whole summary and the whole original documents is smaller than the sum of sentence-level distortion.*

For simplicity, let us see the case of K-L divergence first.

$$\begin{aligned}
Dis &= 1/n \sum_{x_i} D_{KL}(p(y,x_i)||p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij}\hat{x}_j)) \\
&= 1/n \sum_{y} \sum_{x_i} p(y,x_i) \log \frac{p(y,x_i)}{p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij}\hat{x}_j)} \\
&\geq 1/n \sum_{y} (\sum_{x_i} p(y,x_i)) \log \frac{p(y,x_i)}{p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij}\hat{x}_j)} \\
&\quad \text{according to the log sum inequality [15]} \\
&= 1/n \sum_{y} (p(y,X)) \log \frac{p(y,x_i)}{p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij}\hat{x}_j)} \\
&= 1/n D_{KL}(p(y,X)||p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij}\hat{x}_j))
\end{aligned}$$

Then it is easy to see the log sum inequality is also correct for J-S divergence, and we can gain the same conclusion when using J-S divergence as the distortion measure following the relationship between K-L divergence and J-S divergence.

$$
\begin{aligned}
Dis &= 1/n \sum_{x_i} D_{JS}\big(p(y,x_i)||p(y,\sum_{\hat{x}_j\in S}\hat{\lambda}_{ij}\hat{x}_j)\big) \\
&= 1/n \sum_{x_i} 1/2 \Bigg\{ D_{KL}\Big(p(y,x_i)||1/2\big(p(y,x_i)+p(y,\sum_{\hat{x}_j\in S}\hat{\lambda}_{ij}\hat{x}_j)\big)\Big) \\
&\quad +D_{KL}\Big(p(y,\sum_{\hat{x}_j\in S}\hat{\lambda}_{ij}\hat{x}_j)||1/2\big(p(y,x_i)+p(y,\sum_{\hat{x}_j\in S}\hat{\lambda}_{ij}\hat{x}_j)\big)\Big)\Bigg\} \\
&\geq 1/n * 1/2 \Bigg\{ D_{KL}\Big(p(y,X)||1/2\big(p(y,X)+p(y,\sum_{\hat{x}_j\in S}\hat{\lambda}_{ij}\hat{x}_j)\big)\Big) \\
&\quad +D_{KL}\Big(p(y,\sum_{\hat{x}_j\in S}\hat{\lambda}_{ij}\hat{x}_j)||1/2\big(p(y,X)+p(y,\sum_{\hat{x}_j\in S}\hat{\lambda}_{ij}\hat{x}_j)\big)\Big)\Bigg\} \\
&= D_{JS}\big(p(y,X)||p(y,\sum_{\hat{x}_j\in S}\hat{\lambda}_{ij}\hat{x}_j)\big)
\end{aligned}
$$

Bibliography

[1] Priyanka Agrawal, Lavanya Sita Tekumalla, and Indrajit Bhattacharya. Nested hierarchical dirichlet process for nonparametric entity-topic analysis. In *Machine Learning and Knowledge Discovery in Databases*, pages 564–579. Springer, 2013.

[2] Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174, 1974.

[3] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004.

[4] Phyllis B Baxendale. Machine-made index for technical literature: an experiment. *IBM Journal of Research and Development*, 2(4):354–361, 1958.

[5] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[6] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[7] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7, 2010.

[8] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[10] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.

[11] Asli Celikyilmaz and Dilek Hakkani-Tur. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 815–824. Association for Computational Linguistics, 2010.

[12] Asli Celikyilmaz and Dilek Hakkani-Tür. Discovery of topically coherent sentences for extractive summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 491–499. Association for Computational Linguistics, 2011.

[13] Yllias Chali and Shafiq R Joty. Improving the performance of the random walk model for answering complex questions. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 9–12. Association for Computational Linguistics, 2008.

[14] John M Conroy and Dianne P O'leary. Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407. ACM, 2001.

[15] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[16] Hoa Trang Dang. Update summarization task and opinion summarization pilot task. In *Proceedings of the First Text Analysis Conference (TAC 2008)*.

[17] William M Darling and Fei Song. Pathsum: A summarization framework based on hierarchical topics. *on Automatic Text Summarization 2011*, page 5, 2011.

[18] Jean-Yves Delort and Enrique Alfonseca. Dualsum: a topic-model based approach for update summarization. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 214–223. Association for Computational Linguistics, 2012.

[19] Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu, and Cindy Xide Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1271–1279. ACM, 2011.

[20] Inderjit Dhillon, Jacob Kogan, and Charles Nicholas. Feature selection and document clustering. In *Survey of text mining*, pages 73–100. Springer, 2004.

[21] Harold P Edmundson. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285, 1969.

[22] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, 22(1):457–479, 2004.

[23] David Kirk Evans, Judith L Klavans, and Kathleen R McKeown. Columbia newsblaster: multilingual news summarization on the web. In *Demonstration Papers at HLT-NAACL 2004*, pages 1–4. Association for Computational Linguistics, 2004.

[24] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005.

[25] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[26] Patrick Flaherty, Guri Giaever, Jochen Kumm, Michael I Jordan, and Adam P Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics*, 21(15):3286–3293, 2005.

[27] Samuel J Gershman and David M Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

[28] Zoubin Ghahramani. Bayesian nonparametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A*, 371(1984), 2013.

[29] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128. ACM, 1999.

[30] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4*, pages 40–48. Association for Computational Linguistics, 2000.

[31] Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. ACM, 2001.

[32] Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *In NIPS*, pages 475–482. MIT Press, 2005.

[33] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies:*

The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 362–370. Association for Computational Linguistics, 2009.

[34] Peter Harremoës and Naftali Tishby. The information bottleneck revisited or how to choose a good distortion measure. In *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*, pages 566–570. IEEE, 2007.

[35] Zhanying He, Chun Chen, Jiajun Bu, Can Wang, Lijun Zhang, Deng Cai, and Xiaofei He. Document summarization based on data reconstruction. In *AAAI*, 2012.

[36] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[37] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*, pages 604–611. Citeseer, 2006.

[38] Houping Jia, Xiaojiang Huang, Tengfei Ma, Xiaojun Wan, and Jianguo Xiao. Pkutm participation at tac 2010 rte and summarization track. *Proc. of TAC*, 2010.

[39] Hongyan Jing and Kathleen R McKeown. Cut and paste based text summarization. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 178–185. Association for Computational Linguistics, 2000.

[40] Karen Sparck Jones and Julia R Galliers. *Evaluating natural language processing systems: An analysis and review*, volume 1083. Springer, 1996.

[41] Michael Kaisser, Marti A Hearst, and John B Lowe. Improving search results quality by customizing summary lengths. In *ACL*, pages 701–709. Citeseer, 2008.

[42] Sinae Kim, Mahlet G Tadesse, and Marina Vannucci. Variable selection in clustering via dirichlet process mixture models. *Biometrika*, 93(4):877–893, 2006.

[43] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM, 1995.

[44] Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, pages 897–904, 2008.

[45] Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha, and Yong Yu. Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of the 18th international conference on World wide web*, pages 71–80. ACM, 2009.

[46] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

[47] Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 463–470. Association for Computational Linguistics, 2006.

[48] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics, 2003.

[49] Dahua Lin and John Fisher. Coupling nonparametric mixtures via latent dirichlet processes. In *Advances in Neural Information Processing Systems 25*, pages 55–63, 2012.

[50] Dahua Lin, Eric Grimson, and John W Fisher III. Construction of dependent dirichlet processes based on poisson processes. 2010.

[51] Tao Liu, Shengping Liu, Zheng Chen, and Wei-Ying Ma. An evaluation on feature selection for text clustering. In *ICML*, volume 3, pages 488–495, 2003.

[52] Elena Lloret and Manuel Palomar. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1):1–41, 2012.

[53] Chong Long, Minlie Huang, Xiaoyan Zhu, and Ming Li. Multi-document summarization by information distance. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 866–871. IEEE, 2009.

[54] Annie Louis and Ani Nenkova. Automatic summary evaluation without human models. In *Notebook Papers and Results, Text Analysis Conference (TAC-2008), Gaithersburg, Maryland (USA)*, 2008.

[55] Annie Louis and Ani Nenkova. Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 306–314. Association for Computational Linguistics, 2009.

[56] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

[57] Tengfei Ma and Hiroshi Nakagawa. Automatically determining a proper length for multi-document summarization: A bayesian nonparametric approach. In *EMNLP*, pages 736–746, 2013.

[58] Tengfei Ma, Issei Sato, and Hiroshi Nakagawa. The hybrid nested/hierarchical dirichlet process and its application to topic modeling with word differentiation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[59] Tengfei Ma and Xiaojun Wan. Multi-document summarization using minimum distortion. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 354–363. IEEE, 2010.

[60] Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. The tipster summac text summarization evaluation. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 77–85. Association for Computational Linguistics, 1999.

[61] Daniel Marcu. From discourse structures to text summaries. In *Proceedings of the ACL*, volume 97, pages 82–88. Citeseer, 1997.

[62] Victoria McCargar. Statistical approaches to automatic text summarization. *Bulletin of the American Society for Information Science and Technology*, 30(4):21–25, 2004.

[63] Ryan McDonald. *A study of global inference algorithms in multi-document summarization*. Springer, 2007.

[64] Kathleen McKeown, Rebecca J Passonneau, David K Elson, Ani Nenkova, and Julia Hirschberg. Do summaries help? In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217. ACM, 2005.

[65] Kathleen McKeown and Dragomir R Radev. Generating summaries of multiple news articles. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82. ACM, 1995.

[66] Rada Mihalcea and Paul Tarau. A language independent algorithm for single and multiple document summarization. 2005.

[67] Andrew H Morris, George M Kasper, and Dennis A Adams. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, 3(1):17–35, 1992.

[68] Peter Müller, Fernando Quintana, and Gary Rosner. A method for combining inference across related nonparametric bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):735–749, 2004.

[69] Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[70] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer, 2012.

[71] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. NAACL-HLT 2004, 2004.

[72] Constantin Orasan, Viktor Pekar, and Laura Hasler. A comparison of summarisation methods based on term specificity estimation. In *LREC*, 2004.

[73] Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. 2010.

[74] You Ouyang, Wenjie Li, Sujian Li, and Qin Lu. Applying regression models to query-focused multi-document summarization. *Information Processing & Management*, 47(2):227–237, 2011.

[75] John Paisley and Lawrence Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 777–784. ACM, 2009.

[76] John Paisley, Chong Wang, David M Blei, and Michael I Jordan. Nested hierarchical dirichlet processes. *arXiv preprint arXiv:1210.6738*, 2012.

[77] John W Paisley, Aimee K Zaas, Christopher W Woods, Geoffrey S Ginsburg, and Lawrence Carin. A stick-breaking construction of the beta process. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 847–854, 2010.

[78] Emily Pitler, Annie Louis, and Ani Nenkova. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 544–554. Association for Computational Linguistics, 2010.

[79] Jim Pitman. Poisson-dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11(5):501–514, 2002.

[80] Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938, 2004.

[81] Dragomir R Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Celebi, Danyu Liu, and Elliott Drabek. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 375–382. Association for Computational Linguistics, 2003.

[82] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

[83] Korbinian Riedhammer, Daniel Gillick, Benoit Favre, and Dilek Hakkani-Tür. Packing the meeting summarization knapsack. In *INTERSPEECH*, pages 2434–2437, 2008.

[84] Abel Rodriguez, David B Dunson, and Alan E Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483), 2008.

[85] Noam Slonim. *The information bottleneck: Theory and applications*. PhD thesis, Hebrew University of Jerusalem, 2002.

[86] Josef Steinberger and Karel Jezek. Sentence compression for the lsa-based summarizer. In *Proceedings of the 7th International conference on information systems implementation and modelling*, pages 141–148, 2006.

[87] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM, 2004.

[88] Krysta Marie Svore, Lucy Vanderwende, and Christopher JC Burges. Enhancing single-document summarization by combining ranknet and third-party sources. In *EMNLP-CoNLL*, pages 448–457. Citeseer, 2007.

[89] Simon Sweeney and Fabio Crestani. Effective search results summary size and device screen size: Is there a relationship? *Information processing & management*, 42(4):1056–1074, 2006.

[90] Simon Sweeney, Fabio Crestani, and David E Losada. show me more: Incremental length summarisation using novelty detection. *Information Processing & Management*, 44(2):663–686, 2008.

[91] Jie Tang, Limin Yao, and Dewei Chen. Multi-topic based query-oriented summarization. In *SDM*, volume 9, pages 1147–1158. SIAM, 2009.

[92] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.

[93] Romain Thibaux and Michael I Jordan. Hierarchical beta processes and the indian buffet process. In *International conference on artificial intelligence and statistics*, pages 564–571, 2007.

[94] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[95] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM, 2008.

[96] Xiaojun Wan. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1546–1555. Association for Computational Linguistics, 2011.

[97] Xiaojun Wan and Jianwu Yang. Improved affinity graph based multi-document summarization. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 181–184. Association for Computational Linguistics, 2006.

[98] Chong Wang and David M Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in neural information processing systems*, pages 1982–1989, 2009.

[99] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. Multi-document summarization using sentence-based topic models. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 297–300. Association for Computational Linguistics, 2009.

[100] Kam-Fai Wong, Mingli Wu, and Wenjie Li. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 985–992. Association for Computational Linguistics, 2008.

[101] Drausin Wulsin, Brian Litt, and Shane T. Jensen. A hierarchical dirichlet process model with multiple levels of clustering for human eeg seizure modeling. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 57–64, 2012.

[102] Shiren Ye, Tat-Seng Chua, Min-Yen Kan, and Long Qiu. Document concept lattice for text understanding and summarization. *Information Processing & Management*, 43(6):1643–1662, 2007.

[103] Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. Multi-document summarization by maximizing informative content-words. In *IJCAI*, volume 2007, page 20th, 2007.

[104] Guan Yu, Ruizhang Huang, and Zhaojun Wang. Document clustering via dirichlet process mixture model with feature selection. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 763–772. ACM, 2010.

[105] Kai Yu, Shipeng Yu, and Volker Tresp. Dirichlet enhanced latent semantic analysis. In *LWA*, pages 221–226, 2004.

[106] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

[107] Shi Zhong and Joydeep Ghosh. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8(3):374–384, 2005.