The Hong Kong Polytechnic University

Department of Computing

An Integrated Summarization Framework with Hierarchical Content Representation

OUYANG You

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

December 2010

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

OUYANG You (Name of Student)

Abstract

With the rapid growth of Internet services, more and more electronic text is accessible on-line. While the abundance of information provides more resources for individuals, it also results in the well-recognized information overload problem -- the excessive amount of information being provided. The technology of automatic text summarization has emerged to deal with this problem.

Automatic text summarization is the process of creating a shortened version of text by computational techniques to help users catch the important content of the original text(s) with affordable time costs. According to the ways of summary composition, there are extractive summarization methods and abstractive summarization methods. Currently, extractive methods are the mainstream, which will be the focus in this dissertation.

The main question to be answered in extractive summarization is how to select a set of sentences from the input documents to form a summary that can best convey the important content of the input documents. Setting off by discovering important words in the input documents to answer the question, we propose several content models for word saliency estimation and word-based sentence ranking and then develop two word-based summarization methods with the content models. Experimental results prove the effectiveness of the proposed methods applied to several authoritative data sets from the Document Understanding Conference (DUC) tasks. Our next target is to incorporate the relations between important words into the summarization process. We propose several methods to identify the latent word relations in the input documents and use them to obtain a hierarchical representation

of the document content. Based on the hierarchical content representation, we propose a novel hierarchical summarization method that follows the general-to-specific style to extract summary sentences. Unsystematically studied in previous researches, hierarchical summarization is characterized by integrating various summarization objectives to simultaneously improve the content and readability of the composed summaries. The experimental results on the DUC data sets prove the advantages of the proposed method over traditional summarization methods. Finally, we conduct several tentative studies to examine the use of more sophisticated content representations beyond single words for improving the hierarchical summarization method. The tentative studies capture several important details in developing good hierarchical summarization methods and shed light on the directions of future work in hierarchical summarization.

Publications Arising from the Thesis

1. **Ouyang, Y.**, Li, S., Li, W.. 2007. Developing learning strategies for topic-based summarization. In Proceedings of the 16th ACM conference on Conference on information and knowledge management (CIKM 2007), pages 79-86.

2. **Ouyang, Y.**, Li, W., Wei, F., Lu, Q.. 2009. Learning Similarity Functions in Graph-Based Document Summarization. In Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy (ICCPOL 2009), pages 189-200.

3 **Ouyang, Y.**, Li, W., Lu, Q.. 2009. An Integrated Multi-document Summarization Approach based on Word Hierarchical Representation. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009), poster session, pages 113-116.

4. **Ouyang, Y.**, Li, W., Zhang, R., Lu, Q.. 2010a. A Study on Position Information in Document Summarization. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pages 919-927.

5. **Ouyang, Y.**, Li, W., Li, S., Lu, Q.. 2010. Inter-topic Information Mining for Query-based Summarization. In Journal of the American Society for Information Science and Technology (JASIST), 61(5), pages 1062-1072.

6. **Ouyang, Y.**, Li, W., Zhang, R., Lu, Q.. 2010b. Applying regression models to query-focused multi-document summarization. In Information Processing & Management (IPM), Article in Press, Corrected Proof.

7. Zhang, J., **Ouyang, Y.**, Wei, F., Hou, Y.. 2009. A Novel Composite Kernel Approach to Chinese Entity Relation Extraction. In Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy (ICCPOL 2009), pages 236-247.

8. Li, W., Wei, F., **Ouyang, Y.**, Lu, Q.. 2008. Exploiting the Role of Named Entities in Query-Oriented Document Summarization. In Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence (PRICAI 2008), pages 740-749.

9. Cai, X., Li, W., **Ouyang, Y.**. 2010. Simultaneous Ranking and Clustering of Sentences: An Reinforcement Approach to Multi-Document Summarization. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pages 134-142.

10. Li, W., **Ouyang, Y.**, Hu, Y., Wei, F.. 2008. PolyU at TAC 2008. In the Text Analysis Conference (TAC) 2008, the *summarization* track.

11. **Ouyang, Y.**, Li, W.. 2009. PolyU at TAC 2009. In the Text Analysis Conference (TAC) 2009, the *summarization* track.

12. **Ouyang, Y.**, Li, W., Zhang, R.. 2010. Keyphrase Extraction Based on Core Word for Identification and Word Expansion. In the Evaluation Exercises on Semantic Evaluation 2010, the *automatic keyphrase extraction from scientific articles* track.

Acknowledgements

I wish to express my gratitude to everyone who contributed to making the dissertation a reality.

First and foremost, I would like to express my deepest thanks to my supervisor Prof. Wenjie Li. During the period of my studies, she not only gave her full effort to this research and supported me during the years to bring it to fruition, but also taught me how to overcome the difficulties in research, which will absolutely be a great wealth in my future career. And, I would like to thank all of the teammates in Prof. Li' group for their continuous support and kind help.

Second, I would like to thank Prof. Qin Lu, my co-supervisor, for her valuable advice and consistent support in the past three years. It is also my great pleasure to thank Prof. Korris Chung, Dr. Yan Liu and Dr. Dacheng Tao, for their insightful comments and helpful advices in my confirmation and guided studies. I am again grateful to all my friends for their encouragement and help.

Finally, I would like to express my deepest appreciation to my family for their constant support and patience during the years to make my dream come true.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background of text summarization

With the rapid growth of the World Wide Web and Internet services, more and more electronic text is accessible on-line. While the abundance of information provides more resources for individuals, it also results in the well-recognized information overload problem -- the excessive amount of information being provided. Lacking time to read everything, users usually expect to have the most important information. This need is duly addressed by the emerging technology of automatic text summarization, a process of creating a shortened version of text by computational techniques to help users catch the most important information in the original text(s) with affordable time costs. The existing researches have investigated the applications of automatic summarization techniques to a variety of challenging problems in the telecommunications industry, data mining systems of text databases, word processing tools, web-based information retrieval tools, on-line information organization systems, etc. As the information overload problem grows, new requirements and applications of summarization keep springing up.

After Luhn's (1958) initial study on automatically generating short abstracts for science articles, many text summarization tasks have been studied. As a matter of fact, the methods for different summarization tasks may be quite diverse. To categorize the summarization tasks and methods, Jones (2007) defined a set of influencing summarization factors that are used to locate the tasks, including input factors, purpose factors and output factors, which included nearly all the

summarization factors explored in the past decade. As it is indeed unnecessary for us to introduce all the factors in this dissertation, we only discuss those related to the studies in this dissertation.

Input factors

Language and Genre

Text summarization can be applied to various types of textual data. The summarization methods for a specific type of text should consider the specific characteristics, such as the languages and genres of the text. In this dissertation, we mainly consider the task of summarizing English newswire documents, which are widely used as a benchmark for summarization methods.

Units

According to the number of source documents, single-document and multi-document summarization tasks can be differentiated. In practice, the multiple inputs in the multi-document tasks usually make the summarization process more complicated than in the single-document tasks. The main cause is the cross-document issues in the multi-document summarization process, such as the redundancy caused by similar sentences in different documents, the organization of summary sentences from different documents, etc.

Purpose factors

Use

Early researches in summarization usually aim at summarizing the input documents without other requirements, which are usually referred to generic summarization. With the development of web-based information retrieval, query-focused summarization, which requires summarizing from a set of documents in response to a given query, attracts more recent attention. More recently, tasks with

even more refined and user-oriented purposes are proposed and studied, such as update summarization, opinion summarization, aspect-based summarization, etc. In this dissertation, we consider both the generic and query-focused summarization tasks.

Output factors

Informative or Indicative

This factor is about the function of the summary. The purpose of an indicative summary is just to alert its readers in relation to the content of the input documents(s). Key phrases, word hierarchies and other forms of text can all be regarded as indicative summaries. In contrast, an informative summary is viewed as a substitute for the input documents, which is usually presented as a new document of a much shorter length.

Extractive or abstractive

The informative summary can be further divided into two categories according to the writing style of the output summary: abstracts and extracts. Abstractive summarization produces a concise abstract that involves re-writing actions on the input documents to more compactly and accurately present the content of the documents, which is closer to what human summarizers do. However, due to the limitation of current natural language generation techniques, automatic abstractive summarization methods are rather underdeveloped in theory and practice. In contrast, extractive summarization, which selects a number of indicative text fragments from the input documents to form an extract, are better studied and more practicable. This is why the vast majority of the existing studies are focused on extractive summarization and so is our study in this dissertation.

Length

Longer summaries can convey more information in the input documents but also take more time to read. Therefore, the compression rate (or summary length) is also an important factor in document summarization. In practical tasks, the requirements of the summary length are usually given by either (1) reducing the documents with a given compression rate, or (2) generating a summary with a given maximum number of words or sentences.

Quality

For different summarization tasks, the criteria for judging the quality of the summaries may also differ. In this dissertation, we mainly consider three common quality criteria, including saliency, coverage and fluency.

Saliency: a summary should cover the most salient concepts of the input documents. Therefore, a good summarization method should be able to discover these concepts, which is usually cast as a saliency estimating problem or a ranking problem.

Coverage: a summary should cover as many salient concepts as possible within the length limit. For this objective, how to reduce the repeating concepts in a summary is the main issue to be considered.

Fluency: the sentences in a summary should be logically organized for easy reading and understanding. This requires the summarization methods to re-order the extracted sentences in the summary.

1.2 Motivation

Since automatic summarization tasks are usually complicated tasks involving many factors, we have to apply multiple techniques to generate a summary from input documents. In a typical extractive summarization framework, the summarization task is cast as a multi-stage process involving three kinds of summarization techniques: sentence ranking, redundancy control and sentence ordering.

Summaries are composed of sentences. So the sentence ranking method is actually the core component in most extractive summarization methods. By estimating the saliency scores of the sentences in the input documents, the most salient sentences can be identified and selected to compose a summary with good saliency. Besides sentence ranking, the redundancy control technique is also very important in the sentence selection process. As real data usually contains many repeating concepts, it is unnecessary to repeat one concept in the summary. Repeating concepts cannot bring any more information to the summary but only make it less compact. The redundancy control technique is conducive to improving the conciseness of a summary by excluding the sentences that contain many repeating concepts. Thus the summary is able to cover more salient concepts provided with the fixed length constraint. Finally, after all the summary sentences are extracted, the sentence ordering technique is used to organize the sentences in a logical order and make the summary more readable. There are various sentence ordering methods, including the majority methods which order two sentences based on their orders in the input documents, the chronological methods which are based on temporal information of the sentences, the clustering-based methods which place

sentences on the same topic together, etc.

In a typical extractive summarization scheme, the three kinds of techniques are sequentially applied to the input documents in order to achieve the objectives sequentially. We call this scheme "sequential summarization" in our study. Figure 1 below illustrates the process of sequential summarization.

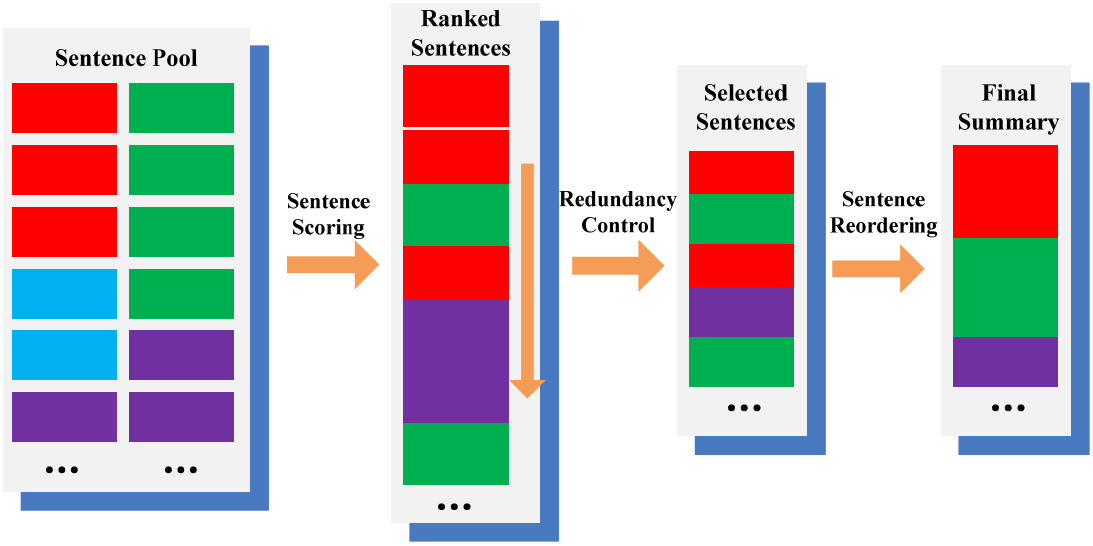

Figure 1. A typical sequential summarization process

Motivated by simultaneously achieving various summarization objectives, there are also methods that integrate the multiple stages in the sequential summarization process into one single process. Since the coverage and fluency of a summary are actually summary-level objectives, they cannot be well handled by optimizing the independent sentences only. Therefore, an intuitive way of integrated summarization is to apply the optimizing process to the whole summary in order to address all the objectives. However, since optimization-based summarization methods attempt to find the best summary globally, they have to face the exponential number of possible sentence combinations, which renders them intractable.

Regarding the summarization process illustrated in Figure 1, the reason why we call it "**sequential**" summarization is that it treats the target summary as an object of

sequentially-distributed sentence slots and fills the slots one after another with the extracted sentences. In this dissertation, we introduce another type of summarization methods, called "**hierarchical**" summarization. As illustrated in Figure 2, a hierarchical summary is defined as a summary that uses the parent-subsidiary relationship to organizes the summary sentences as a hierarchy. In composing a hierarchical summary, relevance and redundancy can be naturally ensured by the inclusion of new and related information when inserting a summary sentence into the hierarchy. Meanwhile, the summary sentences can be organized by the parent-subsidiary relationship in the hierarchy to improve the fluency of the summary. The sentence relationship is actually used to ensure the different summarization objectives and thus hierarchical summarization can be considered as a type of integrated summarization. Compared to sequential summarization, hierarchical summarization has the advantage of using the sentence relationship, which is crucial in the construction of the sentence hierarchy and for the improvement of the summary quality.

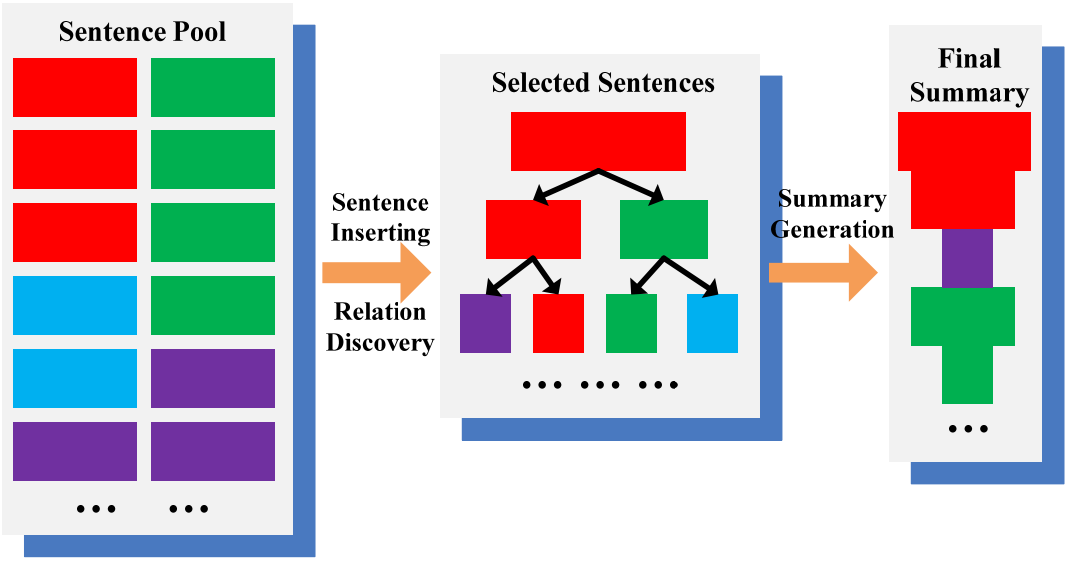

Figure 2. A hierarchical summarization process

In this dissertation, we develop a hierarchical summarization framework. The motivation of the framework comes from the general-to-specific human summarization process. As a matter of fact, there are many different writing styles used by human summarizers, such as storyline, logical order, etc. Nevertheless, among the various styles, the general-to-specific style is a common style suitable for many summarization tasks, especially for multi-document summarization tasks.

When summarizing a set of documents, a human summarizer may first draw an overall picture of the document content in his mind to catch the main topics, the supporting details as well as the parent-subsidiary relations among them. In the summarization process, he/she may start with a couple of sentences to cover the main topics. Then he/she may consider providing supporting details, such as specific examples, reasons and statistics etc., to prove or explain the main topics. He/she may even want to refer to more specific subsidiary details if the summary length allows. Motivated by this, we also follow the general-to-specific summarization style in the automatic hierarchical summarization framework. First, the general sentences are selected into the sentence hierarchy as the top-level nodes in the sentence hierarchy. The selection of the general sentences is similar to sentence selection in sequential summarization. Then, we continue to select the sentences that can support the ideas embedded in the general sentences according to the parent-subsidiary relationship between them. This is achieved by measuring the recommendation degree of the new sentences by the general sentences. Moreover, we may further select the sentences that are recommended by any selected sentence. Through this recommending process, a sentence hierarchy can be constructed along with the recommendation information between sentences. Finally, the hierarchical summary can be generated by extracting the summary sentences from the sentence hierarchy.

An important feature of the hierarchical summarization framework is that it achieves multiple summarization objectives in one sentence selection process. In a word, we choose to "select the sentences to construct a concise and well-organized summary" in the framework instead of "making the already-selected sentences more concise and well-organized" in the traditional multi-stage summarization framework. Therefore, it can be viewed as an integrated summarization framework that can simultaneously solve the problems of sentence ranking, redundancy control, and sentence ordering. In addition, the quality of the summary in coverage or fluency is less limited by the saliency-driven sentence ranking results. Different from the optimization-based methods that are also targeted to integrated summarization, we achieve the aim of objective integration by utilizing sentence relationship instead of optimizing the whole summary. The main advantage is that the resulting sentence selection process still follows the greedy algorithm, not subject to the exponential number of possible sentence combinations. We believe that it is much more efficient.

In the proposed hierarchical summarization framework, we mainly rely on the sentence relationship beyond limit of independent sentences and are thus able to deal with summary-level objectives. Without doubt, the identification of sentence relationship is the core problem in developing the framework. In our study, the parent-subsidiary relationship between two sentences is identified by the related concepts in them. The basic idea is: a supporting concept is more likely to be covered in the summary when it can be related to a general concept that is already covered. For example, once we cover a general word "school" by a sentence of the summary, we may like to continuously cover "student" or "teacher" in the following sentences. An example sentence pair is provided below to illustrate the idea.

*Sentence **A**: the **schools** that have vigorous music programs tend to have higher*

academic performance.

*Sentence **B**: among the lower-income **students** without music involvement, only 15.5 percent achieved high **math scores**.*

Assume that the sentence ***A*** is already selected. we will then move on to select the sentence ***B*** considering that the new concept "student" in the sentence ***B*** is related to the existing concept "school" in the sentence ***A***, and similarly the new concept "math score" is related to the existing concept "academic performance". We define a conditional saliency measure for the subsidiary sentence ***B*** to the general sentence ***A***, which is determined by this kind of "new and related" concepts. Note that the conditional saliency measure can also be regarded as an asymmetric relationship between the two sentences, which indicates the recommendation degree of a sentence by another.

1.3 Methodologies and Contributions

In the dissertation, we conduct a series of studies to solve various problems in developing the hierarchical summarization framework.

(1) First of all, we investigate the word-based summarization methods as the starting point. In the study, we consider the problem of how to discover the important content of the input documents, which is necessary for almost every extractive summarization method. We first propose a learning-based method, in which machine learning models are used to learn the scoring functions from a set of pre-defined features for estimating the saliency scores of words in the input documents. Based on the word saliency estimation, a word-based sentence ranking model is proposed for sentence extraction and summary generation. The purpose of this initial study is to

understand the influencing factors in the sentence ranking problem. Based on the experimental results of the learning-based method, we propose another content model that accounts for the word frequency feature only for word saliency estimation. Although the model is much more light-weighted, its performance is comparable to the learning-based method.

From these studies, we develop an effective and efficient summarization method that captures some essence in sentence retrieval, including the principles for word saliency estimation, and redundancy control, etc. This method can be used as a good starting point for developing more sophisticated summarization methods. In this dissertation, it is used as the foundation for the hierarchical summarization methods to be developed.

(2) In the second study, a hierarchical summarization framework is developed with word relations incorporated into the summarization process in addition to the information within the independent words. To establish the framework, we first develop several methods to automatically identify word relations that are not explicit in the input documents. The relations are used to organize the words in the documents as a hierarchical text graph, which simulates humans' overall understanding of the documents. Different from most existing studies on word hierarchy whose target is just the hierarchy itself, the word hierarchy in our method is developed with the following summarization process in sight.

With the identified word relations, we define the subsumption relationship between sentences, which measures the degree that a sentence will recommend another sentence to be included in the summary. With the subsumption relationship, we finally introduce the hierarchical summarization method that uses sentence relationship to extract summary sentences. When compared to most existing

summarization methods that independently select every summary sentence, the hierarchical method has the advantage of using sentence relationship in the summarization process, which enables it to generate informative, coherent and fluent summaries. Experiments are conducted on various data sets to demonstrate the advantages of the hierarchical method over the traditional sequential methods. We also show that the novel summarization framework can be well applied to different summarization tasks by explaining the general-to-specific sentence selection process in the specific tasks.

The contribution of this study is the novel hierarchical summarization framework, which is also the main contribution of this dissertation. Based on the analysis of the methodology and the experimental results, we suggest that hierarchical summarization has great potentials and should be regarded as a new direction for future summarization endeavours.

(3) In the final study, we aim to further improve the hierarchical summarization framework with more sophisticated representations of the content of the input documents. We explore the use of three different content representations in the study, including key phrases, WordNet synsets, and latent topics generated by hLDA (a complicated probabilistic model). The corresponding methods based on the representations are proposed and then compared to the word-based method in order to test the effectiveness of these sophisticated content representations. The contribution of this study is that it illustrates many important details about hierarchical summarization, such as the definitions of vertices and edges in the hierarchical content representation of the documents, which are very helpful in developing sophisticated hierarchical summarization methods.

1.4 Organization

The remainder of this dissertation is organized as follows: Chapter 2 gives the details of background information and related works, including sentence ranking methods, redundancy control methods, sentence ordering methods, optimization-based methods and other summarization methods. In Chapter 3, we investigate the content models for word saliency estimation and word-based sentence ranking. In Chapter 4, we introduce the hierarchical summarization framework and hierarchical summarization methods. Chapter 5 extends the word-based hierarchical summarization method to explore phrase-based methods, synset-based methods and hLDA-based methods. The final chapter concludes all the studies in this dissertation and identifies the directions of future work.

Chapter 2 Literature Review

This dissertation focuses on a novel hierarchical summarization framework that integrates multiple summarization methods. Therefore, we first review the previous researches on each category of methods respectively, including sentence ranking methods, redundancy control methods and sentence ordering methods. Then, optimization-based summarization methods are introduced as the previous research on integrated summarization. Finally, we also briefly review some other indicative summarization methods.

2.1 Sentence Ranking Methods

Sentence ranking or similar methods, which determine the important parts of input documents, are one of the core components of extractive summarization. Therefore, the sentence ranking problem has been given high priority in the area. In the literature of summarization, various sentence ranking methods are continuously proposed through the years and the mainstream methods are introduced below.

2.1.1 Feature-based sentence ranking methods

Feature-based sentence ranking methods measure the saliency of sentences with a set of features, which character the sentence saliency from different aspects. In the ranking methods, different features are combined together to obtain a composite function for sentence saliency estimation. Various features have been proposed and examined in previous researches. In the early studies such as (Luhn, 1958;

Baxendale 1958; Edmensun, 1969), surface features were mainly considered, such as word frequency, position, title word, cue word, etc. From the 1990s, the development of text summarization techniques moved into a new era. Many natural language processing techniques were applied to the area and yielded many effective features.

Among the features proposed in previous researches, the most common type of features is the frequency-based features. In the very initial study by Luhn (1958), it was already pointed out that words appearing more frequently in a document are usually more indicative for the document. In Luhn's method, word frequency was used as an indicator of keywords and the summary was composed to cover more keywords. This idea is very basic that it nearly works in every practical summarization task. In fact, frequency information is so important that most existing summarization methods considered it in their summarization processes. Nenkova & Vanderwende (2005) have exclusively exploited the role of frequency information in document summarization when developing the Sumbasic system. In their study, it was found that the most frequent words in the input documents stood much larger chances to appear in human summaries. For example, four human summaries were able to cover 94.66% of the top 5 words and 85.25% of the top 12 words in the input documents. Based on this observation, they used the bag-of-words model to derive a sentence scoring method, in which the saliency score of a word was about proportional to its frequency and the sentence score was estimated by average word score. They reported that Sumbasic outperformed most systems in the Document Understanding Conference 2004 (DUC) competition though it considered only frequency information.

Frequency-based features were widely-adopted in many successful summarization methods. Besides the word frequency (i.e., total number of word

appearances), other forms of frequency information were also studied in previous researches. In (Radev et al., 2000), the classical IDF (inversed document frequency) was also used in word saliency estimation. The resulting system MEAD performed very well in DUC 2003 (Radev et al., 2003). In the Fastsum system from (Schilder & Kondadadi, 2008), the document frequency of a word was used to reflect its ability in covering the whole document set instead of word frequency. They reported that document frequency was more effective than word frequency from the experimental results on the DUC 2007 data set. Since extractive summarization can also be viewed as the process of retrieving important sentences, a sentence-level frequency, the inversed sentence frequency (ISF), was also considered as an indicator of important words (Neto et al., 2000).

The use of frequency information in different summarization tasks varies. For documents with titles, the word frequency in the titles was regarded as an effective feature since title words are normally more indicative than general words in main texts (Edmundson, 1969). In query-focused summarization, the frequency of a word in the query is obviously very important in measuring its saliency (Ouyang et al., 2007). Moreover, the frequency of a word in the query-related context is also a good feature for measuring its relevance to the query (Ouyang et al, 2010).

Besides the Uni-gram-based features mentioned above, various *N*-gram-based features were also concerned in discovering the important content of input documents. However, *N*-grams (*N* larger than 1) are not as effective as Uni-grams in practice. This is mainly due to the data sparse problems commonly existing in summarization tasks, which may only involve a small set of documents. In (Martins & Smith, 2009), they first proposed a basic summarization method with Uni-gram features and then incorporated Bi-gram features into the method. However, the

experiments showed that the additional Bi-grams did not bring any improvement to the sentence ranking results. In actual data, the proportion of unimportant *N*-grams to all *N*-grams quickly becomes very large when *N* increases. This makes it much harder to discover important *N*-grams for large *N*s when there are more noises. One solution to this problem is to consider the significant *N*-grams only, such as the *N*-grams appearing more frequently than some fixed thresholds (Celikyilmaz & Hakkani-Tur, 2010). Nevertheless, previous researches still suggested that Uni-grams were much more effective than the other *N*-grams.

Besides simply counting appearances, language models were also used to model more kinds of frequency information. For example, Gupta et al. (2007) compared different content models for frequency-based saliency estimation and found that the log-likelihood of a word in input documents was better than the word frequency. Lin & Hovy (2000) regarded the indicative degree of a word to a given document set as the difference between the distribution of the word in the given set and the distribution in another document set consisting of randomly-collected documents. They used the chi-square hypothesis test to judge whether the word is significantly more relevant to the given document set. The resulting measure is used to find the indicative words, which was called “term signature” in their study. The term signature is a good feature for word saliency estimation that a summarization method with only two features (term signature and query frequency) performed as well as the top systems in DUC 2005 and 2006 (Conroy et al., 2006). Later, implicit semantic information in the input documents was also explored for sentence ranking. In (Gong & Liu, 2001), the Latent Semantic Analysis (LSA) model was used to analyze the semantic concepts in the input documents. Then, important concepts were identified according to the singular vectors of the LSA results. However, it was showed that

LSA-based summarization methods could not outperform traditional frequency-based methods regardless of its complexity. The feasibility of semantic analysis models for sentence ranking was not proved. Wang et al., (2008) further applied the symmetric matrix factorization method for a sentence-level semantic analysis on the input documents. They showed that the symmetric matrix factorization method was better than LSA. However, their method still did not perform state-of-the-art. Daumé & Marcu (2006) proposed a Bayesian summarization method that followed the Latent Dirichlet Allocation (LDA) model to discover the latent topics in the given documents. The resulting system performed very well in the Multilingual Summarization Evaluation (MSE) 2005 and DUC 2006. In (Haghighi & Vanderwende, 2009), the hierarchical version of LDA (hLDA) was used to model the documents and rank the sentences. In their method, only the most general topic in the hLDA result was used to select summary sentences. They also proposed other sentence ranking methods for comparison, including several frequency-based methods and a method based on the non-hierarchical LDA model. Experimentally, the hLDA-based method performed better than all of the other methods.

In short, frequency information is very basic in sentence ranking and it is almost adopted in every practical summarization system. Previous researches also showed that the summarization methods with only frequency information can already perform quite well in various summarization tasks.

Position-based features are another widely-used kind of features in document summarization, which are especially effective in generic summarization. As early as in (Edmundson, 1969), position information was embedded in a location-based summarization method, which assigned different weights to the sentences in a

document according to their ordinal positions. The hypothesis behind the method is that the importance of a sentence is likely to decrease with its distance from the beginning of the document. Position information has since been used in many summarization methods, usually in the form of sentence position features. For example, in the MEAD system (Radev et al., 2000), a position feature was included together with frequency-based features. It was defined as a descending function of the sentence position. In the conclusive overview by Nenkova (2005) on the DUC 2001-2004 competitions, it was reported that position information was very effective for generic summarization. In generic single-document summarization, a lead-based baseline system that simply took leading sentences as the summary outperformed most systems submitted in the DUC 2001 and 2002 competitions. In generic multi-document summarization, the lead-based baseline was still competitive in composing short summaries though it appeared not so good in composing longer summaries. Later, position information was applied to more summarization tasks. In query-focused summarization, sentence position features were widely used in learning-based summarization methods as one of the features for calculating the composite sentence scores (Ouyang et al, 2007; Toutanova et al, 2007). Schilder and Kondadadi (2008) analyzed the effectiveness of the features used in their learning-based sentence ranking model. By comparing the ROUGE results of each individual feature, they argued that position features were less effective than frequency features in query-based summarization. In (Gillick et al., 2009), the role of position information in update summarization was studied. By using ROUGE to measure the density of valuable words at each sentence position, it was observed that the first sentence in a newswire document was especially important for update summaries. They thus defined a binary sentence position feature based on this

observation and as expected this feature did improve the performances in practical update summarization tasks. In (Ouyang et al., 2010a), a word position feature was defined by word positions instead of sentence positions. It was compared to traditional sentence position features on various summarization data sets and the advantages of word position features were proved by the experimental results. Moreover, they also examined the effectiveness of position information in different summarization tasks and the results further proved the conclusions observed in previous studies.

Besides frequency-based features and position-based features, there are also many other features proposed in previous researches. In (Edmendson, 1969), cue words such as adverbs of conclusions were considered as good indicators of summary sentences. In contrast, cue words such as ordinals or cardinals were regarded as negative indicators. Negative indicators were also commonly used in practical systems, usually in the form of filtering strategies. For example, in many summarization systems submitted to the DUC competitions, the sentences containing the pattern “somebody says” were regarded as less important since quotation sentences are usually less indicative than declarative sentences. Besides cue phrases, more syntactic-based and semantic-based text units were also used as indicators of summary sentences. In (D'Avanzo & Magnini, 2005), key phrases were first identified from the input documents and then sentences were ranked according to the number of key phrases in them. Named entities or noun phrases were also considered to derive linguistic features beyond *N*-grams (Goldstein et al., 1999). Barzilay & Elhadad (1997) adopted the lexical chains (i.e., the sequences of related words in the documents) to improve the lexical coherency of the composed summary. Leskovec et al. (2004) conducted a deep semantic analysis on the input documents to extract the

"subject–predicate–object" logical form triples, which were used in the subsequent sentence ranking process. Event-based summarization methods, which summarize a document set as a series of events, is another kind of methods using complex text units for summarization. Usually, the atomic events in event-based methods are defined by the relationship between named entities. For example, in (Li et al., 2006), an atomic event was defined as a local structure consisting of a verb or an action noun along with two related named entities. Generally, these complex text units were expected to be able to more sophisticatedly model the document content and thus better summarization methods could be achieved.

A common characteristic of the above features is that they mainly consider the internal information inside the input documents. Besides internal information, external resources were also considered for solving the ranking problem in previous researches. For example, Toutanova et al. (2007) used synonym and hyponymy dictionaries to expand the words. Different words with the same sense were combined together to better represent the concepts. The advantage of the method was that the data sparse problem was alleviated by merging the synonyms. The famous lexical database WordNet (Fellbaum, 1998) was another resource widely used in previous researches. It can either be used as a concept dictionary for word expansion (Schiffman et al., 2002), or be used to calculate the semantic similarity between sentences (Li et al., 2005). Besides semantic dictionaries, World Wide Web, the largest textual resource in the world, was also considered for extending the input document set. In (Jagarlamudi et al., 2006), the Yahoo search engine was used to retrieve extra documents according to the given query for query-focused summarization. By expanding the input document set, more relevant concepts were obtained and thus a summary with richer content could be composed. The online

encyclopedia Wiki was also considered as a good knowledge resource. For example, Svore et al. (2007) used the Wikipedia entities to define extra semantic features besides traditional features for sentence ranking.

The advantage of feature-based sentence ranking methods is that they are able to flexibly combine different aspects of sentence saliency by representing aspects as features and combining features in certain ways. For the feature combination problem, the most common solution is a linear composite function that sums all the features together, in which the feature weights may either be manually assigned or experimentally tuned. However, it is not easy to obtain the optimum weights through these ways. Since the definitions of features can be quite diverse, the differences of their characteristics and scales are unpredictable. Therefore, machine learning models were widely used as a better solution for the feature combination problem. Besides, there were also studies on unsupervised solutions. For example, Wei et al. (2009) proposed an unsupervised co-ranking model which combined two features through a mutual reinforcement process. Nevertheless, learning-based methods are one of the most common and reliable methods for feature combination. We will detail the researches on learning-based methods in the next section.

2.1.2 Learning-based sentence ranking methods

The application of machine learning in document summarization has a long history. Kupiec et al. (1995) first proposed a trainable summarization method that combined multiple word-based features to rank sentences. They used a naïve Bayesian classifier to learn the feature weights according to a set of paired documents and summaries. It was reported that the learning-based method that combined all the features was better than any other method using one single feature.

Most early studies followed this idea and extended Kupiec's work by examining more extensive features and/or classification models (Mani & Bloedorn, 1998; Chuang & Yang, 2000; Neto et al., 2002). Hirao et al. (2002) used a set of documents in which key sentences are manually-annotated to train a Support Vector Machine (SVM) model, which was then used to differentiate summary sentences from non-summary sentences in new document sets. They reported that SVM outperformed many other machine learning models on the Japanese Text Summarization Challenge (TSC) data set, such as decision tree and neutral network. Zhou and Hovy (2002) proposed a Hidden Markov Model (HMM) based sentence ranking method and trained the parameters of HMM on the labelled data generated from the Yahoo Full Coverage Collection. The resulting system was comparable to the best system in the DUC 2001 competition. Zhao et al. (2005) applied the Conditional Maximum Entropy model to the DUC 2005 query-focused summarization task yet just achieved mediocre performances. Shen et al. (2007) presented a Conditional Random Fields (CRF) based method for generic summarization and reported that CRF performed better than most previously-used models, such as HMM, SVM, etc.

A common feature of the above methods is that they all cast the sentence ranking problem as classification problems. More recently, learning-to-rank models have been examined for the problem. Amini et al. (2005) investigated the use of learning-to-rank models for single-document summarization and compared the proposed learning-to-rank model to a logistic classifier. It was reported that the learning-to-ranking model outperformed the classification model. Fisher and Roark (2006) considered a perceptron-based ranking method which was learned from automatically-constructed training data. The resulting system ranked the 8th among

the 34 submitted systems in DUC 2006. In DUC 2007, Toutanova et al. (2007) proposed the PYTHY system in which a log-linear ranking function was learned to combine more than 20 features. The system performed very well in the DUC 2007 competition, ranking the second among the 30 participating systems. Learning-to-rank models were also applied to webpage summarization and outperformed classification models again (Wang et al, 2007; Metzler & Kanungo, 2008). Amini and Usunier (2009) presented a transductive learning model that was able to learn the ranking function with fewer labelled instances. This model also outperformed all the classification models investigated in their study. Compared to classification models, the main advantage of learning-to-rank models for sentence ranking is easy acquisition of the training data. It is usually easier to judge the relative preference between two sentences than to exactly tell whether a particular sentence is a summary sentence or not.

More recently, regression models were also considered for the sentence ranking problem. In our previous work, we (Ouyang et al., 2007) used the Support Vector Regression (SVR) model to learn a sentence scoring function. Since sentence scores are indeed continuous, regression models are intuitively more suitable for the problem than classification models and learning-to-rank models. By comparing different kinds of Support Vector Machines, we reported that regression models were indeed able to achieve better performances in practical data (Ouyang et al., 2010b). The SVR-based learning strategy was followed the work presented in (Schilder & Kondadadi, 2008; Metzler & Kanungo, 2008), which further proved the effectiveness of regression models.

The success of learning-based methods largely relies on the sufficient training data. Although there were some previous studies reporting the use of manually

annotated data (Hirao et al., 2002), the available training data for the sentence ranking problem is actually very rare in practice. Since document summarization tasks involve many factors, it usually takes a lot of time and effort to manually annotate the training data. To reduce the expenses of manual annotation, semi-automatic strategies which used other resources to generate the training data were adopted in most existing learning-based methods. In practical data sets, a common manually-generated resource is the standard summaries that are primarily created by human summarizers for automatic evaluation. As a solution for lack of training data, learning-based sentence ranking methods can make a reference to the human summaries to construct pseudo training data. For existing classification models, human summaries were used to determine positive sentences and negative sentences (Kupiec et al., 1995; Chuang & Yang, 2000). They were also used to judge the preferences between sentences for learning-to-rank models (Fisher & Roark, 2006; Toutanova et al., 2007). Usually, the training data were constructed by calculating the similarity of a sentence to the human summaries as a reliable approximation of the real sentence importance. The approximated saliency was then used to label the class of a sentence or judge the preference between two sentences. Certainly, they could also be directly used to train regression models (Ouyang et al., 2007).

The reason why human summaries can be used to construct pseudo training data is that the concepts included in these manually-written summaries are normally the most important concepts exhibited in the input documents. Experiments done by Conroy et al. (2006) supported this idea. In their study, they defined an “Oracle” sentence score which was calculated from the probability distribution of the Uni-grams in human summaries. Then, they used the “Oracle” scores to rank the

sentences and compose the summary. They found that the summaries generated by directly using the "Oracle" scores to extract summary sentences are even comparable to the best human summaries on the DUC 2006 data set under the ROUGE evaluation. This showed that human summaries were effective in measuring the importance degree of sentences and thus they were suitable for constructing training data. In most existing methods, the scoring functions for training data construction were very similar to the criteria of the automatic evaluation method ROUGE, which was mainly based on *N*-gram-based matching schemes. Recently, (Celikyilmaz & Hakkani-Tur, 2010) used the hLDA model to discover the semantic topics in the input documents and the human summaries. They defined a sophisticated estimation of the true sentence saliency based on the discovered topics and reported that the hLDA-based saliency scores could train better sentence ranking models than the *N*-gram-based scores.

Besides being used to train the scoring functions for feature-based methods, machine learning models were also used to ensure the conditions for other summarization methods. For example, it was used to estimate word saliency instead of sentence saliency in (Yih et al., 2007). The estimated word saliency was then fed to a word-based summarization process. In the work described in (Leskovec et al., 2004), the logic form triples were identified through a typical learning-based framework and were used to rank the sentences. In (Ouyang et al., 2009), various machine learning models were used to estimate the relevance between two sentences, which was subsequently used for a graph-based sentence ranking method.

2.1.3 Graph-based sentence ranking methods

There was also a trend to use graph-based ranking models for the sentence

ranking problem in recent years, motivated by considering the relationship between different text units. TextRank (Mihalcea & Tarau, 2004) and LexRank (Erkan & Radev, 2004), both following the famous PageRank algorithm, were good examples. Erkan and Radev (2004) used weighted undirected graphs to represent the documents, in which vertices were the input sentences and edges were established by the cosine similarity between sentences. A PageRank-style iterative ranking algorithm was applied to the document graph to compute the importance scores of the sentences. Following the idea of the HITS algorithm, Zha (2002) considered a bipartite graph to capture the interactions between words and sentences and to simultaneously rank the sentences and the words. The idea behind the method was a mutual reinforcement principle that a word which appears in many salient sentences should have a high saliency score and so does a sentence which contains many salient words. Mihalcea and Tarau (2005) presented a comparative study on different graph-based ranking algorithms, including HITS, PageRank and Positional Power Function. They found that when the same similarity function was used to construct the graphs, the performances of different algorithms were very close. Later, more kinds of graphs were used to model the input documents. Ye et al. (2007) defined directed graphs called “document concept lattice” in their study to represent the relationship between concepts. The concept lattice of a document set was used to measure the importance of concepts and sentences. Wei et al. (2008) proposed a three-level graph-based ranking model, which considered the relationship between documents, sentences and words. The model integrated the intra-unit relationship considered in PageRank-style methods and the inter-unit relationship considered in HITS-style methods. The resulting system performed very well on the DUC 2005 data set. In (Wang et al., 2009a), the hypergraph was used to model the relationship

between more than two sentences, which might not be well modeled by traditional graphs. The strong representation ability of hypergraph did lead to better sentence ranking results. Besides explicit text units such as documents, sentences and words, implicit text units were also considered in graph-based methods. In (Wan & Yang, 2008), the sentences were first grouped into sentence clusters, Then, the sentence clusters were regarded as meaningful topics and used as extra nodes in the text graphs. PageRank and HITS algorithms were applied to the new graphs. It was concluded that the use of extra nodes was able to improve the ranking result.

In specific summarization tasks, graph-based ranking methods should also be specifically developed. For example, for query-focused summarization, it has been well acknowledged that the effect of the given query should be considered in the sentence ranking process. OtterBacher et al. (2005) proposed a query-biased version of TextRank that used the query to modify the random walking process. The query-biased method performed better than the baseline method that simply ignored the query. Their idea was followed by Wan et al. (2006) who further differentiated between inter-document and intra-document walking processes. Addressing the issue of similarity measure with respect to a specified context, Tombros and Rijsbergen (2004) pioneered the development of query-sensitive similarity functions. The idea was followed by (Wei et al., 2008), who applied the query-sensitive similarity to a query-focused summarization task. In (Ouyang et al., 2009), the similarity between two sentences was estimated by multiple similarity measures, which were combined by machine learning models. It was reported that the composite similarity function was better than any single similarity measure for the graph-based ranking algorithm.

2.1.4 Other sentence ranking methods

In the above reviews, we introduced three mainstream sentence ranking methods. Of course, there were many other sentence ranking methods that cannot be exactly classified into one of the categories, such as the clustering-based ranking methods (Hatzivassiloglou, 2001; Nomoto & Matsumoto, 2001). In clustering-based ranking methods, sentences are first grouped into clusters and the summary is generated by picking a sentence from each cluster one after another. Therefore, the main problems to be considered in clustering-based sentence ranking methods are the sentence clustering problem, the cluster ranking problem and the within-cluster sentence ranking problem. In (Wang et al., 2009b), the Latent Dirichlet Allocation (LDA) model was used to discover a set of latent topics in the input documents, which in turn were used to select summary sentences instead of clusters. In (Cai et al., 2010), a reinforcement process between clustering and ranking was considered to simultaneously improve clustering and ranking results. As a matter of fact, there are still many other kinds of sentence ranking methods. Considering the relevance to the dissertation and the limitation of pages, we are not able to include all of them here.

2.2 Redundancy Control Methods

With the sentence ranking results, summary sentences are usually selected following the descending order of rank to cover more salient concepts. Certainly, better sentence ranking methods will lead to summaries with better saliency. On the other side, the sentence selection strategy is also very important in composing better summaries. One of the most important issues in sentence selection is redundancy

control. Carbonell and Goldstein (1998) proposed the famous Maximal Marginal Relevance (MMR) method which for the first time introduced the redundancy control technique to the summarization area. With the MMR method, in each round of the sentence selection process, the saliency score of every unselected sentence was reduced according to its similarity to the newly-selected summary sentence. The unselected sentences were re-ranked for the next round of sentence selection. With this score reduction scheme, two similar sentences were unlikely to be selected into the summary at the same time. The MMR method was able to reduce the redundancy in the summary and thus to cover more salient concepts. A parameter was used in the MMR method to control the penalty to the repeating concepts. It could be viewed as a compromise between saliency and coverage. Redundancy removal was rapidly recognized as a necessary step in document summarization. MMR and its variations were widely used in many successful summarization systems, such as (Radev et al., 2000; Jagarlamudi et al., 2006; Ouyang et al., 2000). The variations of MMR followed similar ideas but differed in detail. In (Li et al., 2005), a simpler and more efficient strategy was used to reduce redundant sentences. Instead of re-ranking the remaining sentences in each round of sentence selection, they simply ignored the sentences that are too similar to any existing summary sentence. Close performance was achieved yet with better efficiency, since the time for the re-ranking processes was saved. Lacatusu et al. (2005) adopted a semantic parser to discover the predicate-argument structures in the sentences and measured the redundancy between two sentences with overlapping predicate-argument structures instead of overlapping words. Wan et al. (2007) devised a greedy algorithm to impose the diversity penalty based on both the similarity scores between sentences and the ranking scores of the selected summary sentences. Xie and Liu (2008) defined

several corpus-based or knowledge-based similarity measures instead of the cosine similarity metric used in the initial MMR method.

Later, other methods were also considered to handle the redundancy issue. For example, Conroy et al. (2006) used a method from numerical linear algebra, i.e., the pivoted QR decomposition, to solve the problem. Given a set of candidate sentences, summary sentences were selected to cover as many unique words in the candidate sentences as possible. They reported that pivoted QR outperformed MMR on the DUC 2005 and 2006 data sets.

Naturally, redundancy can be eliminated by forcing the concepts in each new summary sentence to be different to the previous summary sentences. However, it is interesting that currently no practical summarization method has chosen this way. Instead, they just assigned a relatively small penalty to the repeating concepts instead of full-penalty. This was because most summarization methods actually relied on only a small set of core concepts to ensure the saliency of most summary sentences. As pointed by Katragadda and Varma (2009), most query-focused summarization methods mainly relied on query terms to achieve good ROUGE performances. They found that more than 75% of the sentences picked by automatic summarization systems were “query-biased” (containing at least one query term). As for human summaries, the proportion of query-bias sentences was only about 50%. This showed that automatic summarization systems relied more on the same extent of core terms than human summarizers. Therefore, it was not a good choice for an automatic summarization system to ignore the mentioned words in the sentence selection process as those words were still necessary to judge the saliency of the next summary sentence to be selected.

Besides being augmented after sentence ranking, redundancy control can also

be incorporated with sentence ranking, such as in optimization-based summarization methods. The optimization-based methods are regarded as a type of integrated summarization methods and they will be introduced in Section 2.4.

2.3 Sentence Ordering Methods

In the saliency-driven extractive summarization process, summary sentences are selected in terms of the descending order of saliency. In this process, two successively selected sentences are likely unrelated. Therefore, we need to re-order all the extracted sentences to make the summary more fluent. In different summarization tasks, the sentence re-ordering problems differ too. In the initial work by (Luhn, 1958), the extracted sentences from a single document were just listed by their original order in the document. It was pointed out in (Jing, 1998) that the best order of the selected sentences does not always follow the original order. Nevertheless, the original ordering is indeed quite effective for single-document summarization and the readability of the output summary is acceptable in most cases. As for multi-document summarization tasks, the ordering problem becomes much more complicated because the summary sentences come from different documents.

A method naturally extending the original ordering method is major ordering, which chooses the most common order of two sentences in all the input documents as their order in the output summary. In the major ordering method proposed by (Barzilay et al., 2002), two sentences to be ordered were first mapped into different documents to find their orders in each document and then the final order was determined by the major one among these orders. However, it is hard or even impossible to accurately match a sentence into a document that does not contain it.

To better model the fuzzy relationship between two sentences, probabilistic models were considered (Lapta, 2002; Barzilay & Lee, 2004). In these studies, pair-wise connections between the atomic features of sentences were synthesized to obtain a conditional probability of one sentence to another and the conditional probability was used to determine the next sentence for each extracted sentence.

Another common sentence ordering method is the chronological ordering method, which orders the sentences according to their positions on the time line of the context. There are many previous studies that were devoted to identify and normalize the temporal information in order to align the sentences on a unique time line (Wiebe et al., 1998; Mckeown et al., 1999; Filatova & Hovy, 2001). However, the absolute time information of a sentence is actually very hard to obtain. As a matter of fact, not all the sentences have explicit temporal information. Therefore, chronological ordering methods were usually applied to the data that contained explicit temporal information. For example, for the newswire data sets in which each document is associated with a publishing date, a common method was to measure the time of a sentence by the publishing date of the document (Lin & Hovy, 2001; Li et al., 2005). However, the effectiveness of such kind of methods is not satisfactory on most practical data sets. Sometimes, it even could not perform significantly better than random ordering. For chronological ordering methods, temporal information identification is the most difficult problem to be solved in future.

Besides the major ordering method and the chronological ordering method, Barzilay et al., (2002) also introduced a clustering-based ordering method to group similar sentences together in the summary. The basic idea was that two related sentences should be put in near positions since they may probably talk about the same topic. The idea was followed by Ji and Nie (2008). They first used a

content-based clustering scheme to partition summary sentences into clusters. Then the ordering problem of all the summary sentences was decomposed into cluster ordering and sentence ordering within each cluster. In (Bollegala et al., 2005; Bollegala et al., 2006), different sentence ordering methods were combined by machine learning models to obtain a hybrid ordering method. They reported that the hybrid ordering method outperformed any single ordering method.

Sentence ordering is a very difficult problem. It needs the semantic relationship between summary sentences. In fact, the quality of sentence organization in automatic summaries is one of the most unsolved issues when compared to human summaries. There is still a long way to go to obtain a satisfactory solution to the sentence ordering problem.

2.4 Optimization-based Summarization Methods

In a typical extractive summarization framework, the sentence ranking, redundancy control and sentence ordering methods are sequentially applied on the input documents to obtain the output summary. The methods serve the roles of ensuring the saliency, coverage and fluency of the summary respectively. This ranking-selecting-ordering framework was widely adopted in existing summarization methods. Motivated by the idea of simultaneously ensuring multiple objectives, optimization-based summarization methods were studied in recent years. Filatova and Hatzivassiloglou (2004) first formulated the problem of sentence selection as a maximum set coverage problem. Similarly in (Yih et al., 2007), sentence selection was cast as a Knapsack problem, i.e., selecting the sentences to maximize the informative content-words in the input documents. Instead of greedily selecting

summary sentences by the descending order of rank, they used the dynamic programming algorithms to solve the Knapsack problem for discovering the best subset of sentences. In the optimizing process, saliency and coverage of the summary were simultaneously ensured. In (Takamura & Okumura, 2009), more algorithms were examined to better solve the set coverage problem. In (Haghighi & Vanderwende, 2009), the optimization target was modified to minimizing the KL-divergence between the input documents and the output summary instead of the maximum set coverage problem. Better performances were achieved by defining the new target.

In the above methods, the two integrated objectives were saliency and coverage. Li et al. (2009) considered a framework that was able to integrate more objectives. The diversity, coverage and balance of the summary were all represented by summary-level constraints and optimized by structured-SVM, i.e., the version of Support Vector Machine for predicting multivariate or structured outputs. The resulting summarization method performed very well on the DUC 2001 single-document data set.

The benefit of optimization-based methods is that it is able to simultaneously achieve different summarization objectives. In the methods following the greedy algorithm (such as MMR), the optimization is conducted on independent sentences. Therefore, they cannot ensure the global optimization of the whole summary. Moreover, global objectives such as fluency or coherency cannot be well handled. In contrast, optimization-based methods are advanced in these problems and they target at optimizing the whole summary. However, there are also some problems that limit the use of optimization-based methods, especially the efficiency issue. Since optimization-based methods aim at finding the best subset of sentences, they have to

face the exponential number of possible sentence combinations. The optimization problems considered in the methods, such as the Knapsack problem, are typical *NP*-complete problems. Therefore, efficiency becomes a critical problem when the number of candidate sentences increases. In previous researches, optimization methods were either applied to smaller data sets, such as in single-document summarization tasks (Li et al., 2009), or had to consider approximation algorithms as effective solutions of the *NP*-hard problems (Takamura & Okumura, 2009).

2.5 Other Summarization Methods

Besides the ranking-selecting-ordering methods and the optimization-based methods introduced above, there are also summarization methods following other styles. For example, extractive summaries were composed by paragraphs instead of sentences (Salton et al., 1997). In this section, we briefly review some indicative summarization methods.

RST-based methods are an important kind of summarization methods which were popular in the early years. These methods tried to parse every input document into a rhetorical structure tree (RST), which explored the discourse relations among the sentences. The summary was composed based on the discourse structure of the documents, as in (Marcu, 1999). RST can help improve the discourse-level quality of the summary. However, RST constructing is much harder than sentence ranking and that is why RST-based methods were rarely followed afterwards.

Abstractive summarization methods are another category of methods that are continuously studied in the literature, which involve rewritings on the source document. However, due to the difficulty of natural language generation, the

development of abstractive methods is quite limited. In early abstractive systems, template-based methods were mainly adopted since it was much easier to fill templates than to write whole sentences. FRUMP (Dejong, 1982) was an early documented non-extractive system, followed by STREAK/PLANDOC (McKeown et al., 1995), SUMMON (McKeown & Radev, 1995), SUMMARIST (Hovy & Lin, 1999), SumUM (Saggion & Lapalme, 2002), etc. However, template-based methods were usually confined in specific domains. Later, modifications on the original sentence were also considered as sort of abstractive summarization, which were easier than rewriting new sentences. Sentence compression (or sentence simplification), which removes unnecessary words and phrases from a sentence to make it more concise, was viewed as an abstractive technique (Knight & Marcu, 2000; Zajic et al., 2007; Nomoto, 2007; Liu & Liu, 2009). However, even shallow reductions may break the grammar correctness of the original sentence when core components of the sentence are wrongly removed. Jing and McKeown (2000) proposed the "cut and paste" algorithm that had greater power in rewriting sentences than sentence compression. In the "cut and paste" algorithm, sentences were first reduced and then multiple reduced sentences with similar ideas were pasted together to generate a compact summary sentence. Information fusion was another sophisticated abstractive method (Barzily et al., 1999). In the information fusion method, similar elements that indicated the same concept were first extracted from the input documents and then the related elements were combined to form a single sentence based on an automatic sentence generator. As a matter of fact, the more rewriting processes are involved in the summarization process, the harder it is to ensure the readability of the composed summary. The application of abstractive summarization in practical tasks still relies on the advance of natural language

generation techniques in future.

There were also researches on generating indicative summaries instead of informative summaries. In some previous researches, the keywords of a document were regarded as sort of summaries. Beyond simply listing the keywords, Lawrie et al. (1999) proposed a word hierarchy construction method that automatically organized the words in the input documents by a hierarchical tree. The word hierarchy was viewed as an indicative summary. Witbrock and Mittal (1999) proposed a statistical framework for title generation. The targets of summarization in their study were the brief titles which are not necessary to be whole sentences. In the studies of this dissertation, we mainly consider informative summaries – aggregations of unbroken sentences with correct grammars.

2.6 Chapter Summary

In this chapter, we have briefly overviewed the literature of document summarization. Different techniques in document summarization were presented and analyzed. First of all, we introduce several mainstream sentence ranking methods, including feature-based, learning-based and graph-based methods. In the next section of this dissertation, we will propose various word-based summarization methods that explore word-level features with machine learning models. We then describe the methods for redundancy control and sentence ordering, which are also important in document summarization. Moreover, optimization-based methods are introduced as typical integrated summarization methods that simultaneously achieve different summarization objectives. In our study, several hierarchical summarization methods have been developed and will be explained in Chapter 4 and Chapter 5. These

mentions are regarded as a new type of integrated summarization methods. At the end of the chapter, some other representative indicative summarization methods that are less related to the dissertation are also briefly introduced.

Chapter 3 Word-based Summarization Methods

In this chapter, we introduce the word-based summarization methods that follow the idea of trying to cover more salient words in summaries. We first propose a learning-based method for word saliency estimation in which multiple features are combined by machine learning models. The estimated word saliency is then used in a word-based sentence ranking method. With the initial experimental results, we re-study the content model for word saliency estimation and develop a frequency-based sentenced ranking method, which is both more effective and efficient than the learning-based method.

3.1 Chapter Overview

Given a set of input documents, a good summary should be able to cover the most salient concepts in the input documents. To achieve this objective, we need to identify the salient concepts and develop a summarization process which is able to cover as many salient concepts as possible. In this chapter, we use words to represent concepts for simplicity. Therefore, two main issues to be considered are (1) measuring the saliency of words to determine which words are more important than others and thus should be included by the summary; (2) designing a sentence selection method that is able to cover more salient words.

In previous researches, various content models are proposed to address the two

issues, including feature-based methods, learning-based methods, language models, etc. As a starting point of our study, we consider a learning-based framework to measure the word saliency. Firstly, a total of ten features are designed to depict different aspects of word saliency. Then, the features are combined by machine learning models to obtain the composite score that estimates the word saliency. In our study, we consider three kinds of machine learning models for feature combination, namely classification models, regression models and learning-to-rank models. The effectiveness of different models for sentence ranking is compared both theoretically and empirically.

Based on the estimated saliency, we develop a word-based summarization method, which follows a greedy algorithm to maximize the total saliency score of the words in the summary. Similar to most extractive summarization methods, we first rank the sentences and then select summary sentences following the descending order of rank. The MMR method is applied in the sentence selection process for redundancy control.

To examine the effectiveness of the proposed methods, we conduct a series of experiments on several data sets from the DUC competitions. Inspired by the observation on the initial results, we propose another word-based summarization method that uses a simpler content model for word saliency estimation. It postulates a log-linear hypothesis on the relationship between true word saliency and word frequency. In the resulting summarization method, the hypothesis is used in both word saliency estimation and redundancy control. This makes the method more integrated and straightforward in achieving saliency and coverage. Experiments are again conducted on the DUC data sets and demonstrate the effectiveness of the new method.

3.2 Word Saliency Estimation

In this section, we introduce the learning-based framework for word saliency estimation, including the feature design and the learning process.

3.2.1 Task, data set and evaluation metrics

Word saliency is estimated by features. So features play an important role in the learning-based framework. In different summarization tasks, the influencing factors to word saliency differ. Therefore, particular feature sets are necessary for particular summarization tasks. In the study of this chapter, we take the query-focused multi-document summarization task defined by the National Institute of Standards and Technology (NIST) in DUC 2005 as the example task to introduce the learning-based framework. The task requires creating from a set of relevant documents (usually 25-50 documents) a brief, well-organized and fluent summary to the information seeking need indicated in a given topic description. NIST specifies the task as the main evaluation task for 3 years since 2005 and thus it provides a good benchmark for researchers to exchange their ideas and experiences. An example topic from the DUC 2006 data set is provided below.

<topic>

<num> D0601A </num>

<title> Native American Reservation System - pros and cons </title>

<narr> Discuss conditions on American Indian reservations or among Native American communities. Include the benefits and drawbacks of the reservation system. Include legal privileges and problems. </narr>

</topic>

As illustrated above, the topic description contains both a brief title and a detailed narrative description. It clearly specifies the summarization target. In each year, NIST assessors develop a total of about 50 DUC topics, with each topic consisting of a topic description and a relevant document set. A DUC 2005 topic contains 25-50 related documents selected from Los Angeles Times and Financial Times of London, while a DUC 2006 and a DUC 2007 topic contains exactly 25 documents from Associated Press, New York Times and Xinhua Newswire.

System-generated summaries are evaluated by both manual and automatic evaluation metrics in DUCs (Dang, 2005). In this paper, we use one of the automatic evaluation metrics, the Recall Oriented Understudy for Gisting Evaluation (ROUGE)[1], to evaluate our summarization methods. ROUGE (Lin & Hovy, 2002) is a state-of-the-art automatic summarization evaluation method, which mainly considers *N*-gram comparisons between system summaries and human summaries. For example, ROUGE-2 evaluates a system summary by matching the Bi-grams in it against the Bi-grams in human summaries, i.e.,

$$R_2(S) = \frac{\sum_{j=1}^{h} \sum_{t_i \in S} Count(t_i \mid S, H_j)}{\sum_{j=1}^{h} \sum_{t_i \in S} Count(t_i \mid H_j)}$$

where S is the summary to be evaluated, H_j (j=1, 2, …, h) is the j^{th} human summary. t_i indicates a Bi-gram in the summary S, $Count(t_i \mid H_j)$ is the number of times that t_i occurs in the j^{th} human summary H_j and $Count(t_i \mid S, H_j)$ is the number of times that t_i occurs in both S and H_j.

Besides ROUGE-2, we also report ROUGE-1 and ROUGE-SU4 in the experiments. ROUGE-1 is very similar to ROUGE-2. But it matches Uni-grams

[1] The parameters for running ROUGE is "-n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 –d"

instead of Bi-grams. ROUGE-SU4 matches both Uni-grams and skip-Bi-grams[2]. A more detailed description of ROUGE can be found in (Lin & Hovy, 2002). Although ROUGE just uses simple *N*-gram-based statistics, it works quite well. For example, in DUC 2005, ROUGE-2 has a Spearman correlation of 0.95 and a Pearson correlation of 0.97 compared with human evaluation. ROUGE has been adopted for evaluation in most current summarization researches.

3.2.2 Feature design

For the DUC query-focused multi-document summarization task, we define a total of ten features to measure word saliency. Notice that the learning-based framework is not confined to this task. It can be also applied to other summarization tasks by re-designing the features.

Frequency-based features

Frequency is fundamental in measuring word saliency. In our study, we include several frequency-based features to character different frequency information.

First of all, the count of appearances of a word in the input documents is considered as a basic feature. It directly reflects the dominative degree of the word in the input documents. Denote the frequency of a word w in the document set D as $freq(w|D)$, the word frequency feature (**TF**) is defined as

$$f_{TF}(w) = freq(w \mid D)$$

Besides the word frequency in all the documents, we also consider the maximum word frequency in each document to emphasize the words that dominate

[2] A skip-Bi-gram is a pair of words in their sentence order, allowing for gaps within a limited size (the size is 4 here).

one particular document. Denote the frequency of w in a document d as $freq(w|d)$, the maximum word frequency feature (**MTF**) is defined as

$$f_{MTF}(w) = \text{Max}_{d_i \in D} freq(w \mid d_i)$$

In addition, document frequency is also considered as a good indicator of word saliency. A document frequency feature (**DF**) is defined as

$$f_{DF}(w) = |\{d_i \mid freq(w \mid d_i) > 0\}|$$

In fact, document frequency is about linear to word frequency because the words that appear more frequently are also likely to appear in more documents. The main difference between the two features is that document frequency does not bias to the words that appear very frequently in only a few documents., it has been reported in (Schilder & Kondadadi, 2009) that document frequency was more effective than word frequency for sentence ranking in their work.

Notice that we do not include any sentence-level frequency feature in our study. This is because sentence frequency is almost equal to word frequency in practice. Unlike documents, it is very rare for a sentence to contain the same word more than once. Therefore, the sentence frequency and the word frequency of a word are usually very close and thus we believe that sentence frequency is unnecessary when word frequency features are already included.

Besides the information inside the input documents, the information in external resources is also considered for word saliency estimation. For example, a set of randomly-collected documents are regarded as irrelevant documents and the words that appear frequently in the irrelevant documents are deemed as general words, which are probably not important to the given document set. The inversed document frequency (**IDF**) measure is a common measure that follows the idea. It is usually

calculated by the logarithm of inversed document frequency, i.e.,

$$f_{IDF}(w) = \mathrm{Log}(N / N_w),$$

where N is the total number of documents and N_w is the number of documents that contain w. The difference of the IDF-based feature to the document frequency f_{DF} is that IDF is calculated from an external document collection. In our experiments, we take the IDF scores from the open-source summarization system MEAD[3].

The entropy of a word over all the document sets is used to measure how likely the word belongs to a specific document set. The entropy-based feature (**EN**) is calculated from the whole document set collection $C=\{D_1, D_2, \ldots, D_N\}$ as

$$f_{EN}(w) = -\sum_i [freq(w|D_i)/freq(w|C) \cdot \log(freq(w|D_i)/freq(w|C))]$$

where $freq(w|\cdot)$ indicates the word frequency of w in each document set.

The features derived from language models are also considered in the study. Here we use a log-likelihood statistic which was initially proposed by (Dunning, 1993) and first introduced to summarization by (Hovy & Lin, 2000). This measure uses the chi-square hypothesis test to test whether the distributions of a word are significantly different in relevant documents (i.e., the input documents) and irrelevant documents (i.e., other documents in the corpus). Denote the frequency of a word w occurring in the relevant set and the irrelevant set as O_{11} and O_{12} respectively, the frequency of all the other words occurring in the relevant set and the irrelevant set as O_{21} and O_{12} respectively, the criteria for determining the significant words (**SIG**) is calculated by the log-likelihood ratio between the probabilities of w under the two sets, i.e.,

---

[3] Available at http://www.summarization.com/mead/;

$$f_{SIG}(w) = -2\log\frac{b(O_{11};O_{11}+O_{12},p)\cdot b(O_{21};O_{21}+O_{22},p)}{b(O_{11};O_{11}+O_{12},p_1)\cdot b(O_{21};O_{21}+O_{22},p_2)},$$

where $b(k;n,x) = \binom{n}{k}x^k(1-x)^{(n-k)}$ is the binomial distribution. The parameters p, p_1 and p_2 are estimated by the maximum likelihood estimation method, i.e., $(O_{11}+O_{21})/(O_{11}+O_{12}+O_{21}+O_{22})$, $O_{11}/(O_{11}+O_{12})$, $O_{21}/(O_{21}+O_{22})$ respectively.

In query-focused summarization, the content anticipated in the summary should be related to a given query and thus the words in the query are especially important. A query-based word frequency feature (**QF**) is defined as the frequency of the word in the query (denoted as Q), i.e.,

$$f_{QF}(w) = freq(w|Q)$$

Since queries are quite short compared against documents, they are not allowed to include all the relevant words in the input documents. In our study, we consider another query-based feature based on the relevance of the words to the query. Denote all the appearances of w in the input documents as $A(w)$, the feature (**QTF**) is defined as

$$f_{QTF}(w) = |\{a \mid (a \in A(w)) \wedge (\exists w_q \in Q, w_q \in S_a)\}|,$$

where w_q is a non-stopword in Q and S_a is the sentence containing a.

The feature f_{QTF} is very similar to the query-independent frequency feature f_{TF}. Their difference is that f_{TF} counts all the appearances of a word while f_{QTF} only counts those co-occurring with the query. In fact, not all the appearances of the word are relevant to the query. Therefore, f_{QTF} is expected to be better than f_{TF} for word saliency estimation in query-focused summarization tasks.

Position-based features

Position features is another common kind of features in text summarization. Usually, position features are defined by the distance of sentences to the beginning of the document. For the word saliency estimation problem, we follow the method proposed in (Yih et al., 2007) to transfer sentence-level position information to word-level position information. The position feature of a word is defined by the positions of all the sentences that contain it.

We first define the position feature of an appearance (denoted as *a*) by the position of the corresponding sentence (denoted as *s*). The inverse proportion function is adopted in the computation, i.e., $f(a) = 1/i$, where *i* is the ordinal position of *s*. Then, we define two position features (**AP** and **MP**) which are calculated by the average and maximum position features of all its appearances respectively. For a word *w*, the features are defined as

$$f_{AP}(w) = \sum_{a_i \in A(w)} f(a_i) / |A(w)| \quad \text{and} \quad f_{MP}(w) = MAX_{a_i \in A(w)} f(a_i)$$

3.2.3 Feature combination

With the designed features, we need to combine them to calculate the overall saliency scores of words, denoted as *score*(*w*). We consider two common composite functions for feature combination: the linear sum function and the exponential product function, i.e.,

Sum $score(w) = \sum_i \lambda_i f_i(w)$

Product $score(w) = \prod_i f_i(w)^{\lambda_i}$

λ_i are the weights of the features, which are usually manually-assigned or experimentally-tuned.

Besides the weighted functions, we also consider machine learning models as a more sophisticated way for feature combination. With labelled training samples, machine learning models are supposed to be able to find the optimum composite function for the fixed feature set. In our study, we adopt the Support Vector Machine model (SVM; Vapnik, 1995) and formulate the problem as below.

Based on the features, a word w can be represented by a feature vector V_w that consists of the feature values of w. The objective is then the estimation of saliency score *score*(w) from the feature vector V_w. Therefore, the target is indeed a mapping function f: $V_w \rightarrow$ *score*(w). In SVM, the optimum function is selected from a candidate function pool through an optimizing process on the training samples. For example, the linear SVM considers a linear function pool, which consists of weighted linear functions for feature combination, i.e.,

$$\{f(V_w) = W \cdot V_w + b \mid W \in R^n, b \in R\},$$

where n is the dimension of feature vectors, W and b are parameters to be optimized. The target is to find the best parameters W_0 and b_0 based on a training data D that consists of labelled samples $\{(w_i, d_i) \mid i = 1,...,l\}$ (d_i is the label of word w_i). In SVM, the optimum function is found by minimizing the structure risk, which is calculated by the formula below:

$$\Phi(W, b) = \frac{1}{2} \| W \|^2 + C \cdot \sum_{i=1}^{l} L(W, b, d_i)$$

The first part of the risk is the normalization factor that is used in margin-based machine learning models such as SVMs. The second part is the total loss of the candidate function $f(x) = Wx + b$ on all the training samples, in which L is a loss function that determines the penalty when the predicted label $f(V_{w_i})$ and the real

label d_i are not equal. C is a weight balancing the two parts. In different types of SVMs, training samples and loss functions are different. More details are introduced in the next section.

3.2.4 A theoretical comparison of the learning models for word saliency estimation

Three kinds of machine learning models are considered in our study, i.e., classification models, regression models and learning-to-rank models. To compare the effectiveness of different models for feature combination, we use the SVM models as the representatives for them. The representative models are Support Vector Classification (SVC), Support Vector Regression (SVR) and Ranking-SVM (Vapnik, 1995; Gunn, 1998; Joachims, 2002). As a matter of fact, these models are quite similar, all following the above risk formula to find the optimum composite function. The difference of them lies in the second part of the risk formula, which indicates the total loss on training samples.

In classification models, the training data consists of a set of labelled samples $\{(x_i, y_i) \mid i = 1, \ldots, l\}$ in which x_i indicates the input and $y_i \in \{+1, -1\}$ indicates the class label. The basic principle for finding the optimum classification function is to minimize the total classification error $\sum_{i=1}^{l} L(f(x_i), y_i)$. An example loss function L is the indicator function, i.e., if $f(x_i) = y_i$, the loss is 0; otherwise, the loss is 1. In the classification-based word saliency estimation, the input is the feature vector of a word and the output label indicates whether the word is important to the summary or not. Therefore, the training data for classification models should contain a positive word set and a negative word set. Notice that SVC finally classifies the instances by

the sign of the composite scores. Since the target in our study is to estimate word saliency, we just use the score of feature combination as the saliency estimation instead of the predicted label.

In learning-to-rank models, the learning target is a determining function that ranks a set of input instances. The training data of learning-to-rank models is a set of instance pairs with preferences $D = \{(x_1^1, x_1^2, r_1), (x_2^1, x_2^2, r_2), \cdots, (x_l^1, x_l^2, r_l)\}$, in which x_i^1, x_i^2 are two input instances and r_i is the relative preference between them. In Ranking-SVM, the preference between two instances is judged by the composite scores calculated from feature combination. Therefore, similar to SVC, we can use the composite scores for saliency estimation directly. In Ranking-SVM, the error is the total ranking loss on the training samples, i.e., $\sum_{i=1}^{l} L(f(x_i^1, x_i^2), r_i)$. The loss function L is defined by whether the real rank r_i is consistent with the preference judged by the candidate function f.

Different from above mentioned two models, regression models directly learn the mapping function between saliency scores and input feature vectors. In the training data $\{(x_i, y_i) \mid i = 1, \ldots, l\}$, y_i is a continuous real value that stands for the real word saliency. The loss function $\sum_{i=1}^{l} L(f(x_i), y_i)$ measures the total gap between real values and predicted values of the candidate function f. An example loss function is the square loss function $L(a,b) = (a-b)^2$. Compared to the other two models that are learned from discrete training data, regression models are naturally closer to the word saliency estimation problem. Saliency is continuous!

3.2.5 Training data construction

In the learning process, we learn the composite function from training data and then apply the learned function to obtain the saliency scores of words in test data. Still, there is one problem left, i.e., lack of training data. As a matter of fact, there is no ready data for word saliency estimation. Moreover, it is actually impractical to manually annotate training data. The complexity of the summarization task yields a difficult problem even for human summarizers. In fact, it is even not easy to judge the preference between two words, not to mention providing exact categories or accurate saliency scores. To solve the problem, we adopt the common strategy in which "nearly true" saliency scores are semi-automatically assigned to the words with reference to human summaries. An assumption is made here that the words appearing in the human summaries are the important words. The frequency of a word in human summaries is used as the indicator of its real saliency. Moreover, we use document frequency (denoted as $f_{HDF}(w)$) instead of word frequency to avoid large gaps between the scores of different words. The used measure can also be explained as the number of human summarizers who choose the word in their summaries.

Based on the pseudo word saliency, training data sets are constructed by the following strategies: for classification models, the words that appear in at least one human summary are regarded as positive words and the words that do not appear in any human summary are regarded as negative words; for learning-to-rank models, two words with different document frequencies are regarded as an available pair; for regression models, the document frequency is directly used to learn the regression function.

Since the above strategies are semi-automatic, the qualities of some generated

training samples are not guaranteed. As mentioned before, not all important words in the input documents are fully covered by human summaries due to summary length limitation. Also, there are some words in the free-style human summaries that cannot be found in the input documents. Considering these issues, a filtering strategy is applied to refine the quality of the constructed training data. In the strategy, only those words that are more dominative in the input documents are regarded as available sources for training data. We use the features $f_{TF}(w)$ and $f_{EN}(w)$ to filter the words that are too rare or too general, i.e., $f_{TF}(w) < \alpha_1$ or $f_{EN}(w) > \alpha_2$.

3.3 Word-based Summarization Method

In this section, we introduce the word-based summarization method, which extracts the most salient sentences based on the result of word salience estimation.

3.3.1 Word-based sentence ranking model

In the method, we follow an intuitive idea to maximize the total saliency of the summary. Given the word saliency estimation result, denoted as $score(w)$, the saliency score of a sentence s is calculated as the total saliency scores of words in it, i.e., $score(s) = \sum_{w_i \in s} score(w_i)$. Moreover, the saliency of a summary S consisting of a set of sentences $\{s_1, \ldots, s_n\}$ is calculated as the sum of sentence scores, i.e., $score(S) = \sum_{s_j} score(s_j) = \sum_{s_j} \sum_{w_i \in s_j} score(w_{ji})$. According to the task definition, the length of summaries should not exceed a given number of words. Therefore, the sentences should be selected to maximize the average word saliency of the summary,

i.e., $\frac{\sum_{w_i \in S} score(w_i)}{|S|}$. This normalized scoring function is the one used for sentence ranking.

3.3.2 Sentence selection with redundancy control

Once sentences have been ranked, summary sentences are selected following the descending order of rank to maximize the saliency of the output summary. To control redundancy, we apply a MMR-style method here. Summary sentences are selected through an iterative process. Each time when a new summary sentence is selected, the ranking scores of all the remaining candidate sentences are revised according to their similarities to the selected sentences, i.e.,

$$score'(s) = \lambda * score(s) - (1-\lambda) * Max_{s_i \in S_{selected}} Similarity(s, s_i),$$

where $S_{selected}$ is the set of selected sentences. We use the cosine similarity metric to calculate the similarity between two sentences and set λ to 0.7 as in most methods using MMR. After the scores are revised, the remaining sentences are re-ranked by the new scores and then the top-ranked sentence will be selected in the next round. The iterative process continues until the length of the output summary reaches the limit.

3.3.3 An extractive summary example

An example summary is provided below to illustrate the output of the proposed extractive summarization method. The source data set is the DUC topic "D0701A" from the DUC 2007 data set, which talks about "Southern Poverty Law Center and Morris Dees".

The Southern Poverty Law Center tracks hate groups, and Intelligence Report covers right-wing extremists.

In 1987, Dees won a $7 million verdict against a Ku Klux Klan organization over the slaying of a 19-year-old black man in Mobile, Ala., forcing the group to turn over its headquarters building.

The victims are suing the Aryan Nations and founder Richard Butler.

The Portland case is similar to the Keenan lawsuit, in that Dees argued that White Aryan Resistance founders Tom and John Metzger incited the skinheads to commit murder.

Next, the Aryan Nations In an attempt to seize the compound of the Aryan Nations, Morris S. Dees Jr. went to court in Coeur d'Alene, Idaho.

Dees said he could win a multimillion-dollar civil judgment that would put the South Carolina Klan out of business.

His disciples have included some of the most notorious figures in the white supremacist movement, such as Robert Mathews, who founded a neo-Nazi offshoot of the Aryan Nations, and Buford Furrow, who is awaiting trial in Los Angeles on charges of killing an Asian-American postal carrier and shooting up a Jewish day care center last summer.

3.4 Experimental Results

In this section, we introduce the evaluations on the proposed methods, including the learning-based word saliency estimation method and the word-based summarization method.

3.4.1 Experiment set-up

We conduct a series of experiments on the DUC query-focused multi-document summarization task introduced in Section 3.2. In the experiments, we first evaluate the proposed methods on the DUC 2007 data set and then extend to the DUC 2005 and 2006 data sets. Each data set contains about 50 topics to be summarized, with each topic consisting of about 25-50 newswire documents and an additional topic description as the query. Before the summarization process, pre-process is performed to clean the data sets. Stop-words are removed from the sentences and Porter-stemmer (Porter, 1980) is used to stem the remaining words in order to unify different words with the same morphological root. According to the task definition, system-generated summaries are strictly limited to 250 English words.

To obtain candidate sentences, we use the sentence segmentation tool provided by DUC[4] to segment the input documents into sentences. Moreover, we also consider several heuristic rules to remove invalid summary sentences before sentence selection, including newspaper heads indicating the resources of the news, incomplete sentences that are too short, and quotation sentences that usually present subjective idea of individuals. For the machine learning models, SVM^{light} (Joachims, 1999) is used to implement all the three kinds of learning models and the parameters of SVM^{light} are set to default.

[4] Available at http://duc.nist.gov/

3.4.2 A comparison of different feature combination methods

In the first experiment, we use all the features to estimate the word saliency for all the runs of the word-based summarization system. Therefore, the main influential factor of performance is the feature combination method. By this, we directly compare the effectiveness of different combination methods, including the linear function, the product function and the learning models (SVC, SVR and Ranking-SVM). The experiment is conducted on the DUC 2007 data set, which is the latest data set of the query-focused multi-document summarization task. The training data sets for learning-based methods are also constructed from the DUC 2007 data set and a two-fold cross validation scheme is used to achieve open tests. For consistency, we use the DUC 2007 data set to construct the training data in this and all the follow-up experiments unless otherwise stated. Table 1 below presents the average ROUGE-1, ROUGE-2 and ROUGE-SU4 scores and the corresponding 95% confidential intervals of the word-based systems on the DUC 2007 data set. Moreover, we include three systems submitted in the DUC 2007 competition for reference: the **DUC baseline** system, a leading-based summarization system that simply returns all the leading sentences in one document up to the length limit as the summary; the system **15** that is the best submitted system in DUC 2007; and the system **H** that indicates the summaries generated by a human summarizer.

Table 1. Results of the systems with different feature combination methods on the DUC 2007 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Linear	0.4184 (0.4123-0.4242)	0.1072 (0.1031-0.1116)	0.1570 (0.1531-0.1611)
Product	0.3805 (0.3735-0.3883)	0.0835 (0.0790-0.0882)	0.1358 (0.1311-0.1408)
Classification	0.4211 (0.4152-0.4274)	0.1103 (0.1063-0.1144)	0.1628 (0.1588-0.1670)
Learning-to-Rank	0.4286 (0.4223-0.4348)	0.1164 (0.1124-0.1205)	0.1679 (0.1637-0.1723)
Regression	**0.4301 (0.4237-0.4365)**	**0.1175 (0.1134-0.1219)**	**0.1682 (0.1642-0.1725)**
15	0.4409 (0.4332-0.4481)	0.1239 (0.1189-0.1288)	0.1750 (0.1701-0.1897)
H	0.4785 (0.4636-0.4934)	0.1289 (0.1154-0.1422)	0.1840 (0.1737-0.1931)
DUC Baseline	0.3091 (0.3000-0.3185)	0.0599 (0.0561-0.0639)	0.1036 (0.0995-0.1077)

From the results in Table 1, we can observe that the systems with the same feature set and different feature combination methods have different performances. This shows that feature combination is indeed an important factor of the sentence scoring function. Among all the combination methods, the product-based method performed significantly worse than all the others. Compared to the sum-based method, the product-based method is more sensitive to the effectiveness of independent features. An ineffective feature may have serious impacts on the overall product and this causes the ineffectiveness of the method.

In overall, the performances of learning-based methods are better than the sum-based method and the regression-based method performs the best among the learning-based methods. Though the differences are not quite significant, we can still suggest that regression models are probably better for finding the optimum scoring function with the fixed feature set.

When being compared to the reference systems, the proposed systems significantly outperform the leading-based baseline. More important, the best system **Regression** performs comparably to the best submitted system **15** in DUC 2007. This result clearly shows the competiveness of the proposed methods.

To further confirm the results, we extend the experiments to the DUC 2005 and DUC 2006 data sets. Table 2 and Table 3 below provide the average ROUGE-1, ROUGE-2 and ROUGE-SU4 scores and the corresponding 95% confidential intervals of the systems on the DUC 2005 and 2006 data sets. Systems **15** and **24** are the best submitted systems; systems **H** and **A** indicate the summaries generated by one the human summarizers in each year respectively.

Table 2. Results of the systems with different feature combination methods on the DUC 2005 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Linear	0.3655 (0.3550-0.3751)	0.0693 (0.0663-0.0723)	0.1241 (0.1213-0.1269)
Product	0.3534 (0.3478-0.3590)	0.0595 (0.0568-0.0622)	0.1079 (0.1043-0.1115)
Classification	0.3663 (0.3569-0.3757)	0.0701 (0.0677-0.0736)	0.1243 (0.1202-0.1382)
Learning-to-	0.3702	0.0711	0.1299

Rank	(0.3653-0.3764)	(0.0694-0.0753)	(0.1231-0.1308)
Regression	**0.3770 (0.3713-0.3828)**	**0.0761 (0.0727-0.0793)**	**0.1329 (0.1294-0.1363)**
15	0.3767 (0.3716-0.3818)	0.0738 (0.0711-0.0764)	0.1326 (0.1300-0.1354)
H	0.4220 (0.4059-0.4382)	0.0880 (0.0770-0.0998)	0.1471 (0.1366-0.1594)
DUC Baseline	0.2784 (0.2673-0.2895)	0.0416 (0.0386-0.0446)	0.0885 (0.0842-0.0924)

Table 3. Results of the systems with different feature combination methods on the DUC 2006 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Linear	0.3865 (0.3803-0.3927)	0.0867 (0.0838-0.0906)	0.1371 (0.1327-0.1415)
Product	0.3705 (0.3652-0.3757)	0.0769 (0.0731-0.0807)	0.1213 (0.1173-0.1254)
Classification	0.3887 (0.3836-0.3928)	0.0897 (0.0851-0.0943)	0.1407 (0.1366-0.1448)
Learning-to-Rank	0.3977 (0.3923-0.4031)	0.0901 (0.0861-0.0941)	0.1423 (0.1395-0.1451)
Regression	**0.4011 (0.3956-0.4069)**	**0.0929 (0.0884-0.0972)**	**0.1473 (0.1433-0.1512)**
24	0.4073 (0.4009-0.4137)	0.0950 (0.0907-0.0992)	0.1534 (0.1494-0.1574)
A	0.4530 (0.4446-0.4623)	0.1001 (0.0898, 0.1123)	0.1648 (0.1574-0.1734)
DUC Baseline	0.2981 (0.2874-0.3084)	0.0491 (0.0451-0.0534)	0.0962 (0.0918-0.1006)

The results in the above tables again demonstrate the advantage of the regression-based method that performs better in both the DUC 2005 and DUC 2006 data sets than the classification model and the learning-to-rank model.

3.4.3 The results of training from different data sets

In the above experiments, the learning models are all trained from the DUC 2007 data set. In this experiment, we examine the effectiveness of the regression-based method with different training data. The training data construction strategy is applied to the DUC 2005, 2006 and 2007 data sets to generate three different training data sets. Each training data set is used to learn a regression function. Then the learned regression functions are applied to all the three DUC data sets to test their efficiency. When a DUC data set is used for both training data construction and model test, a two-fold cross validation scheme is used. Tables 4-6 below present the 3×3 training-test result matrices.

Table 4. The ROUGE-1 scores on different training sets and test sets

Train \ Test	**2005**	**2006**	**2007**
2005	**0.3794 (0.3734-0.3850)**	**0.4027 (0.3969-0.4089)**	**0.4315 (0.4252-0.4373)**
2006	0.3763 (0.3705-0.3819)	0.3998 (0.3940-0.4059)	0.4299 (0.4235-0.4361)
2007	0.3770 (0.3713-0.3828)	0.4011 (0.3956-0.4069)	0.4301 (0.4237-0.4365)

Table 5. The ROUGE-2 scores on different training sets and test sets

Train \ Test	2005	2006	2007
2005	**0.0779 (0.0747-0.0809)**	**0.0935 (0.0889-0.0980)**	**0.1173 (0.1133-0.1215)**
2006	0.0766 (0.0734-0.0798)	0.0921 (0.0878-0.0964)	0.1170 (0.1130-0.1211)
2007	0.0761 (0.0727-0.0793)	0.0929 (0.0884-0.0972)	0.1175 (0.1134-0.1219)

Table 6. The ROUGE-SU4 scores on different training sets and test sets

Train \ Test	2005	2006	2007
2005	**0.1356 (0.1323-0.1387)**	**0.1482 (0.1442-0.1524)**	**0.1696 (0.1657-0.1737)**
2006	0.1328 (0.1296-0.1362)	0.1461 (0.1422-0.1502)	0.1680 (0.1641-0.1721)
2007	0.1329 (0.1294-0.1363)	0.1473 (0.1433-0.1512)	0.1682 (0.1642-0.1725)

In the results, the regression-based system trained on the DUC 2005 data set performed slightly better than those trained on the DUC 2006 and 2007 data sets. However, the results are actually very close. This showed that the regression-based summarization method is quite stable in the data sets.

3.4.4 A performance analysis for individual features

No matter what method is used for feature combination, the features always have the greatest influences on the composite scoring function. In this experiment,

we intend to test the effectiveness of single features by using each feature for sentence ranking. In the experiment, each feature is used as the word salience score and fed into the word-based summarization method. Table 7 below presents the average ROUGE-1, ROUGE-2 and ROUGE-SU4 scores and the corresponding 95% confidential intervals of the summarization systems with single features on the DUC 2007 data set, along with two systems using the composite functions.

Table 7. Results of the systems with single features on the DUC 2007 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
TF	0.4166 (0.4102-0.4229)	0.1069 (0.1029-0.1107)	0.1593 (0.1552-0.1636)
MTF	0.4061 (0.3985-0.4136)	0.0968 (0.0920-0.1020)	0.1523 (0.1473-0.1574)
DF	**0.4232 (0.4166-0.4230)**	**0.1134 (0.1093-0.1174)**	**0.1652 (0.1611-0.1696)**
SIG	0.4083 (0.4020-0.4148)	0.1010 (0.0968-0.1053)	0.1540 (0.1497-0.1583)
EN	0.4130 (0.4065-0.4193)	0.1068 (0.1020-0.1116)	0.1569 (0.1527-0.1611)
IDF	0.4130 (0.4062-0.4197)	0.1052 (0.1006-0.1101)	0.1562 (0.1517-0.1606)
MP	0.4170 (0.4106-0.4237)	0.1101 (0.1056-0.1149)	0.1588 (0.1545-0.1632)
AP	0.4095 (0.4026-0.4165)	0.1042 (0.0996-0.1091)	0.1549 (0.1506-0.1596)
QF	0.4000 (0.3931-0.4065)	0.0997 (0.0954-0.1036)	0.1509 (0.1466-0.1551)
QTF	0.4181 (0.4118-0.4246)	0.1076 (0.1031-0.1116)	0.1606 (0.1562-0.1649)
Linear	0.4184	0.1072	0.1570

	(0.4123-0.4242)	(0.1031-0.1116)	(0.1531-0.1611)
Regression	**0.4301 (0.4237-0.4365)**	**0.1175 (0.1134-0.1219)**	**0.1682 (0.1642-0.1725)**

The results show that the effectiveness of features varies. Among all the features, the document frequency (DF) performs the best. A surprising result is that the system with only the DF feature even performs better than the system with the linear combination of all the ten features. We attribute the reason to the characteristics of the designed features. Firstly, the true word saliency actually involves many other factors and thus it actually cannot be perfectly measured by these features; secondly, the designed features overlap to some extent, for example, the DF feature and the TF feature. The interrelation between features may be the main cause of the ineffectiveness of linear combination. On the other hand, the regression-based methods are more sophisticated in combining features and thus they can employ more features to obtain better sentence ranking results.

3.5 Re-studying the Content Models

In this section, we further analyze the results on single features and then propose a new content model for word saliency estimation, which is both more efficient and effective than the learning-based framework.

3.5.1 Analyzing the relationship between the frequencies

As illustrated above, the document frequency feature is the most efficient

among the proposed features. It is better than the word frequency features which are expected to be more informative intuitively. Motivated by this, we'd like to further analyze the data to develop better models for word saliency estimation.

In the word-based summarization method, the effectiveness of a feature is mainly determined by its ability in approximating the true word saliency. In this study, we explain the true word saliency as the times that it is expected to appear in the output summary. Here we approximate it by the frequency of the word in human summaries (denoted as **HF**) as in the learning-based methods. Based on this approximation, a comparison between the features and **HF** is conducted to analyze the relationship between the features and the true word saliency. Table 8 below lists the frequencies of the most frequent words in two different document sets. **TF**, **DF**, **QTF** and **HF** indicate the word frequency feature, the document frequency feature, the query-co-occurrence feature and the frequency in human summaries, respectively.

Table 8. The frequency information in two practical data sets

Word	**TF**	**DF**	**QTF**	**HF**	**Word**	**TF**	**DF**	**QTF**	**HF**
Simpson	259	25	259	38	**art**	196	18	196	44
Brown	52	9	42	7	**music**	194	18	194	29
Goldman	47	11	39	5	**school**	181	18	181	26
auction	38	6	25	5	**student**	101	15	60	6
trial	38	11	22	4	**education**	83	13	59	15
lawyer	37	7	25	4	**program**	73	11	64	9
million	33	7	23	6	**children**	64	8	44	3
Nicol	30	5	29	7	**city**	54	10	47	1

From the table, we can observe that the features are actually about linear to HF. To further prove the idea, we use the variance analysis to calculate the significance of the linear relationship between the features and HF. Results show that the linear relationship between QTF/TF/DF and HF are all very significant (> 99.9%). As to the most frequent 100 words in each DUC 2007 document set, the *P*-values of the linear coefficients of QTF/TF/DF to HF are 7.35E-30, 1.95E-33 and 1.21E-49 respectively. Among them, the DF feature is the most significant and this may explain why it obtained the best performance in the experiment on single features.

While a word may appear hundreds of times in the input documents, it does not appear equally frequent in the summaries. Summaries are usually much shorter. Therefore, the dominating words under TF/QTF are likely to be too dominating in the summary. In contrast, the scale of DF is closer to the scale of HF and this is the reason why it does better in approximating the true word saliency. Based on this idea, we postulate an assumption that the logarithm of word frequency is better for word saliency estimation because it can avoid the impact of over-dominating. The idea is detailed in the next section as a log-frequency hypothesis.

3.5.2 The log-frequency hypothesis

In this section, we present another word-based summarization method based on a log-frequency hypothesis. First of all, we give the log-frequency hypothesis as: the real saliency of a word is proportional to the logarithm of word frequency rather than the original word frequency. To prove the rationality of the hypothesis, we examine the linear coefficients between log-QTF/log-TF/log-DF and on the top 100 words from each DUC 2007 document set. The corresponding *P*-values are 5.78E-51,

2.12E-58 and 3.94E-49 respectively. Compare to the *P*-values of the original frequency features (7.35E-30, 1.95E-33 and 1.21E-49), we can see than the *P*-values of the word features QTF/TF are much improved, while the *P*-value of the document feature DF is not much changed. Therefore, we can assume that the log-frequency hypothesis does hold for the word-level frequency.

To further prove the the log-frequency hypothesis, we try to deduce it from the bag-of-words model. Under the bag-of-words model, the probability of a word *w* in a document set *D* is proportional to its frequency, i.e., $p(w) = freq(w) / |D|$, where $freq(w)$ indicates the frequency of *w* in *D* and $|D|$ indicates the total number of words in *D*. The probability of a sentence, denoted as *s*, is then calculated as the product of word probabilities, i.e., $p(s) = \prod_{w_i \in s} p(w)$. Moreover, the probability of a summary consisting of a set of sentences (denoted as *S*) can be calculated by the product of sentence probabilities, i.e., $p(S) = \prod_{s_j \in S} p(s_j)$. To obtain the optimum summary, an intuitive idea is to select the sentences that maximize the overall summary probability $p(S)$, which is equivalent to maximizing

$$\log_\alpha p(S) = \log_\alpha \prod_{s_j \in S} p(s_j) = \log_\alpha \prod_{s_j \in S} \prod_{w_{ji} \in s_j} p(w_{ji}) = \sum_j \sum_i \log_\alpha (p(w_{ji}))$$
$$= \sum_j \sum_i (\log_\alpha freq(w_{ji}) - \log_\alpha |D|) = \sum_j \sum_i (\log_\alpha freq(w_{ji})) - |S| \cdot \log_\alpha |D|$$

where w_{ji} indicates the *i*th word in the *j*th sentence s_j and $|S|$ indicates the total number of words in *S*, α is the constant log-base. Under the condition that the length of the summary reach the length limit, both $|S|$ and $|D|$ are constants. Then, the above optimization target is equivalent to maximizing $\sum_j \sum_i (\log_\alpha freq(w_{ji}))$. Comparing this formula to the one in the proposed word-based sentence ranking method, this

indeed implies $score(w_{ji}) = \log_\alpha freq(w_{ji})$, i.e., the importance of a word is proportional to its log-frequency.

After explaining the log-frequency hypothesis by the bag-of-words model, now we use it for sentence ranking, in particular, we use $\log_\alpha freq(w)$ to estimate the saliency scores of the words. As mentioned previously, the word saliency can also be explained as the number of times that it should be covered in the output summary. According to this idea, we consider a dynamic word scoring scheme for redundancy control. Intuitively, when a word is covered by a newly-selected sentence, the remaining times that it needs to be covered in the future will reduce by one. Thus the current saliency score of the word, which measures the number of times that it should be covered in future, should also reduce by one, i.e., $\log_\alpha freq(w) - 1 = \log_\alpha \frac{freq(w)}{\alpha}$. This yields a natural dynamic word scoring method for redundancy control, i.e., when a new sentence is selected, the frequency feature of the words in it will be damped by multiplying a damping factor 1/α. The advantage of this method over the MMR method is that it is more consistent with the word-based sentence ranking method. In MMR, the formula of the score damping scheme is given as $score'(s) = \lambda * score(s) - (1-\lambda) * Max_{s_i \in S_{selected}} Similarity(s, s_i)$, in which the reduced importance score is usually calculated by the cosine similarity metric. In fact, such similarity measures are not in accordance with the sentence scores estimated by the sentence ranking methods. Therefore, it is not natural to reduce the sentence score by the similarity score as in the MMR method.

On the other side, in the proposed method, word damping and word-based ranking are both based on the log-frequency hypothesis. The initial word saliency

score and the reduced saliency score can both be explained by the number of times that a word should be covered in the summary and thus they are directly comparable. Therefore, the saliency and redundancy objectives are more integrated in our method than in traditional methods which apply sentence ranking methods and redundancy control methods sequentially.

3.5.3 A word-based summarization method based on the log-frequency hypothesis

Based on the log-frequency hypothesis, we develop a frequency-based summarization method. The details of the method are provided below.

Set the initial importance scores of the words, $score(w) = \log_{\alpha} freq(w)$;

While the summary length does not exceed the limit

Rank the sentences by the scoring function $\frac{score(s) = \sum_{w_i \in s} score(w_i)}{|s|}$;

Select the highest ranked sentence s_0;

Re-set the importance of all the words in s_0 by $score(w) = 1/\alpha \cdot score(w)$;

3.5.4 Experimental results

Experiments are again conducted on the DUC data sets to evaluate the new word-based summarization method. We intent to examine the effectiveness of the log-frequency hypothesis in two aspects: the substitution of frequency-based features by log-frequency features for word saliency estimation and the word damping

scheme for redundancy control.

3.5.4.1 A comparison of the original frequencies and the log-frequencies

In the first experiment, we compare the word-based systems with the original features and the corresponding logarithmic features (denotes as **Log-**). Three frequency-based features (**DF**, **TF**, **QTF**) are included in the experiment. Table 9 below presents the average ROUGE-1, ROUGE-2 and ROUGE-SU4 scores and the corresponding 95% confidential intervals of the systems with each feature on the DUC 2007 data set.

Table 9. The results of the systems with the original frequencies and the log-frequencies on the DUC 2007 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
DF	0.4232 (0.4166-0.4300)	0.1134 (0.1093-0.1174)	0.1652 (0.1611-0.1696)
Log-DF	**0.4285 (0.4227-0.4346)**	**0.1155 (0.1114-0.1198)**	**0.1672 (0.1634-0.1713)**
TF	0.4166 (0.4102-0.4229)	0.1069 (0.1029-0.1107)	0.1593 (0.1552-0.1636)
Log-TF	**0.4243 (0.4172-0.4314)**	**0.1137 (0.1091-0.1185)**	**0.1643 (0.1599-0.1689)**
QTF	0.4181 (0.4118-0.4246)	0.1076 (0.1031-0.1116)	0.1606 (0.1562-0.1649)
Log-QTF	**0.4314 (0.4252-0.4372)**	**0.1175 (0.1127-0.1218)**	**0.1681 (0.1639-0.1723)**

The experimental result shows that the systems with the log-frequencies do outperform the systems with the original frequencies. The result clearly demonstrates the advantages of log-frequency in word saliency estimation, especially for word-level frequencies.

3.5.4.2 A comparison of the redundancy control methods

In the second experiment, we examine the effectiveness of the word damping method for redundancy control. We first run the word-based system with a single feature (**QF** or **Log-QF**) and without applying redundancy control method. Then we use the MMR method and the score damping method to handle the redundancy (denoted as **MMR** and **Damping** respectively). Table 10 below presents the average ROUGE-1, ROUGE-2 and ROUGE-SU4 scores and the corresponding 95% confidential intervals of the runs on the DUC 2007 data set.

Table 10. The results of the systems with different redundancy control methods on the DUC 2007 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
QTF	0.4093 (0.4027-0.4155)	0.1049 (0.1009-0.1089)	0.1575 (0.1533-0.1616)
MMR	0.4181 (0.4118-0.4246)	0.1076 (0.1031-0.1116)	0.1606 (0.1562-0.1649)
Damping	**0.4456 (0.4397-0.4514)**	**0.1185 (0.1144-0.1227)**	**0.1711 (0.1671-0.1751)**
Log-QTF	0.4115 (0.4053-0.4176)	0.1076 (0.1036-0.1117)	0.1597 (0.1556-0.1639)
MMR	0.4314 (0.4252-0.4372)	0.1175 (0.1127-0.1218)	0.1681 (0.1639-0.1723)

Damping	**0.4456 (0.4399-0.4517)**	**0.1189 (0.1140-0.1239)**	**0.1711 (0.1671-0.1755)**

It appears that redundancy control is indeed very important in document summarization. The performances are much improved by incorporating redundancy control methods. Moreover, the damping factor method shows its advantages over the MMR method. We attribute the reason to the fact that the damping factor method is more consistent with the word-based sentence ranking method, which makes it better in handling the redundancy for this particular sentence ranking method.

3.5.4.3 Parameter tuning for the damping factor

Next, we investigate the effect of the damping factor α for redundancy control. The performances of the system with different α's are provided in the figures below. In the figures, the horizontal ordinates are plotted by $1/\alpha$ instead of α. Moreover, we also plot the ROUGE results versus the parameter λ in the MMR method $score'(s) = \lambda * score(s) - (1-\lambda) * Max_{s_i \in S_{selected}} Similarity(s, s_i)$. In fact, the role of λ is similar to α, which is also used to control the penalizing degree on repeating concepts. Both λ and α range from 0 to 1, with 0 indicating full penalty and 1 indicating no penalty.

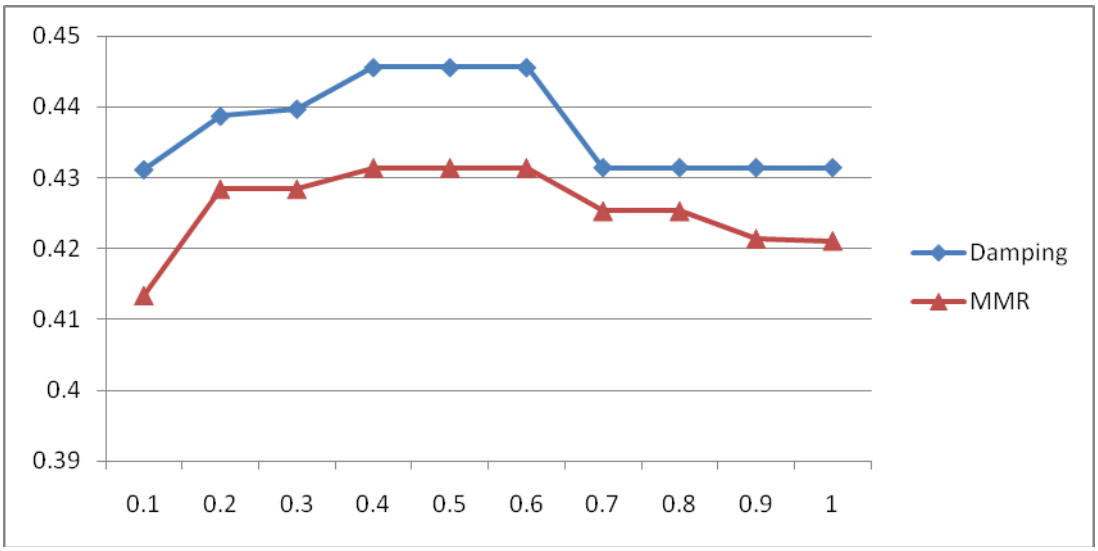

Figure 3. ROUE-1 versus the damping factor

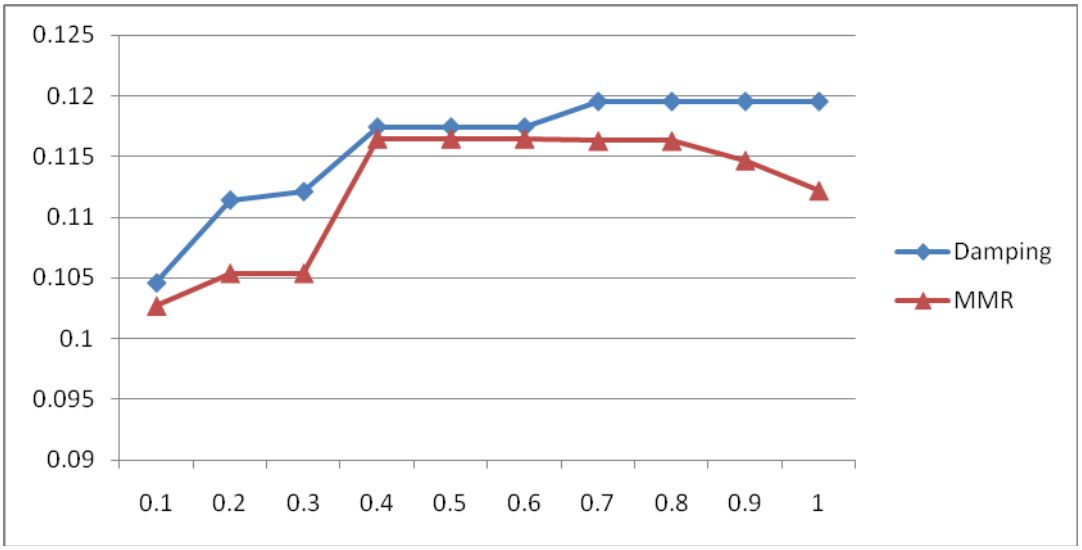

Figure 4. ROUE-2 versus the damping factor

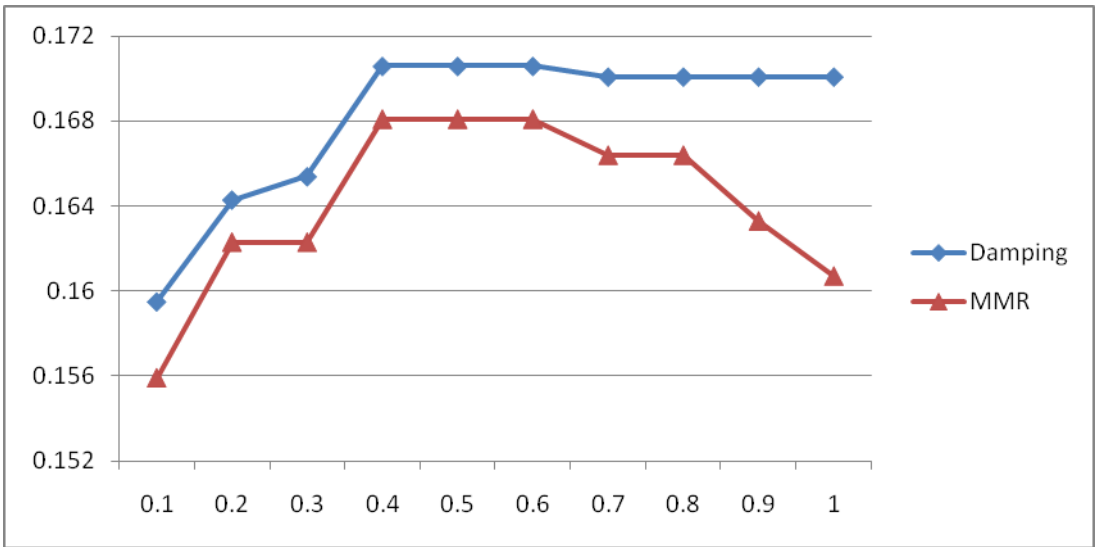

Figure 5. ROUE-SU4 versus the damping factor

We discuss the relationship between the damping factor and the ROUGE scores based on these results. With smaller damping factors, which mean harder penalties on repeating words, the resulting summary tends to include more diverse words and thus it stands a chance to share more words with human summaries, which may lead to higher ROUGE-1 scores. However, at the meanwhile, the average saliency of the words in the summary becomes lower and thus the number of salient words may decrease on the contrary. Consequently, the ROUGE-1 score may also drop when the damping factor is too small. Moreover, the ROUGE-2 score may decrease even more since it requires matching two continuous words, which is obviously harder than matching one single word. This is consistent with the results in Figure 4 that very small damping factor does lead to bad ROUGE-2 results. As to the ROUGE-SU4 result that considers both Uni-grams and Bi-grams, the effect of the damping factor on it can be viewed as a composite effect of those on ROUGE-1 and ROUGE-2.

As in the MMR method, the effect of λ is similar. The best ROUGE scores are obtained with medial λ's. Moreover, we can see that the score damping method performs better than the MMR method in most cases, which again presents the

advantages of the score damping method in redundancy control for the proposed word-based summarization method.

3.5.4.4 Results on generic summarization

Besides query-focused summarization, we are also interested in the effectiveness of the frequency-based summarization method in generic summarization. In this section, we further conduct an experiment on the DUC 2004 generic multi-document summarization data set. This data set includes 45 document sets, with each set consisting of 10 newswire documents. The task requires producing a summary for each document set and the length of summaries is limited to 665 bytes. Table 11 below presents the average ROUGE-1, ROUGE-2 and ROUGE-SU4 scores and the corresponding 95% confidential intervals of the systems on the DUC 2004 data set. The **Original** system is the system without redundancy control. We also include the best submitted system in DUC 2004, denoted as **Best**, for reference.

Table 11. The results of the systems on the DUC 2004 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Original	0.3413 (0.3277-0.3556)	0.00825 (0.0739-0.0920)	0.1200 (0.1113-0.1288)
MMR	0.3702 (0.3586-0.3783)	0.0868 (0.0765-0.0995)	0.1264 (0.1185-0.1327)
Damping	**0.3788 (0.3667-0.3912)**	**0.0929 (0.0849-0.1013)**	**0.1333 (0.1261-0.1404)**
Best	0.3785 (0.3676-0.3894)	0.0916 (0.0827-0.1000)	0.1318 (0.1245-0.1391)

On the DUC 2004 data set, the effectiveness of the new summarization method based on the log-frequency hypothesis is again obvious. The resulting system performs better than the best submitted system in DUC 2004. This shows that the proposed word-based summarization method can well identify the salient content of the input documents for both generic and query-focused summarization although it only depends on the single frequency information.

3.6 Chapter Summary

In this chapter, we investigate the content models for word saliency estimation and sentence selection. We consider a typical learning-based framework in which a total of ten features are used to estimate the word saliency. Three kinds of machine learning models are examined to address the feature combination issue. Experimental results show that the regression model is the best in discovering the optimum sentence scoring function with the same feature set. Moreover, the system based on the proposed methods performs comparably well to the top systems submitted to the DUC competitions.

We then conduct a further study on the content model for word saliency estimation. A log-frequency hypothesis is postulated on the relationship between word frequency and true word saliency. Based on the hypothesis, we propose another summarization method that measures word saliency by the logarithm of word frequency instead of the original frequency. Besides being used for sentence ranking, the hypothesis is also applied to redundancy control, acting as a dynamic damping scheme on the saliency score of words. The resulting method performs quite well in the experiments. It performs comparably well to the best systems submitted to the

DUC competitions. Moreover, it outperforms the more complex learning-based method using all the ten features. This clearly shows the rationality of the log-frequency hypothesis.

In our previous publications (Ouyang et al., 2007; Ouyang et al., 2010b), the learning-based framework were also used to rank sentences and summarize documents. The difference of the previous studies to the study presented in this section is that the features were designed at sentence-level previously instead of word-level. Nevertheless, the advantages of regression models for sentence ranking were proved in all the studies. Regression-based summarization method was recognized as a new type of learning-based summarization method and was followed by Schilder and Kondadadi (2008) and Jin et al. (2010), etc.

From the study, we can conclude that word-based sentence ranking methods are actually quite effective in practice. The results show that the intuitive idea of covering more frequent words can yield very powerful summarization methods. Still, we'd like to incorporate the information beyond frequency into the summarization process to develop better summarization methods. In the next chapter, we will study the use of word relations in hierarchical summarization methods, which is the main target of the dissertation.

Chapter 4 Word-based Summarization with Hierarchical Representation

In this chapter, we consider the relations between words in input documents to improve the frequency-based summarization method. We propose several methods to identify word relations and represent the content of the input documents as a word hierarchy. Then we design a summarization framework based on the hierarchical content representation, which takes advantage of the relationship between summary sentences to simultaneously achieve different summarization objectives.

4.1 Chapter Overview

In the last chapter, we have discussed the content models for word saliency estimation. The content models are used to ensure the saliency of the composed summary. On the other hand, a good summary should also be able to cover as many salient concepts as possible given a fixed length. Moreover, it should be fluent and easy-reading. In most existing methods, these issues are handled by other summarization techniques besides sentence ranking, such as redundancy control and sentence re-ordering.

As introduced in chapter 1, the main focus of this dissertation is the study of a hierarchical summarization framework which is able to integrate different summarization objectives. In the framework, a hierarchical content representation of input documents is constructed by using words as vertices and word relations as

edges. It is used to imitate humans' overall understandings on the document content. Then, a sentence hierarchy is constructed by mimicking the general-to-specific human summarization process, which can also be viewed as a traversing process in the word hierarchy following the general-to-specific order. With the sentence hierarchy, the hierarchical summary can be generated by extracting the top-level sentences from the sentence hierarchy.

The key component of the framework is the subsumption relationship between sentences. It is explained as the recommendation degree of one sentence by another, i.e., when we have already included a sentence ***A*** in the summary, how much we want to further include another sentence ***B*** to support the idea of sentence ***A***. As shown in chapter 1, the relationship between two sentences is determined by the relations between the words in them. Naturally, the first problem to be solved is the identification of word relations in input documents, which are not explicitly provided in texts.

In the following sections, we will introduce the identification of word relations, the definition of subsumption sentence relationship and the hierarchical summarization framework one by one. Both automatic and manual evaluations are conducted to illustrate the advantages of the proposed hierarchical method over traditional sequential methods.

4.2 The Subsumption Sentence Relationship

4.2.1 Word relation identification

Generally, word relations are recognized either by linguistic relation databases

such as WordNet (Banerjee & Pedersen, 2002), or by frequency-based statistics such as co-occurrences. In our study, the target is the subsumption relations between words under the context of input documents. Therefore, it is actually a local problem in which the word senses are confined to the local context. Thus we regard statistic-based methods to be more suitable than database-based methods, considering databases are usually built for general relations.

4.2.1.1 Co-occurrence-based relation identification

Co-occurrence is basic in statistic-based word relation identification. The co-occurrences of two words directly reflect their relevance. Considering the fact that the two words with higher frequencies are also likely to co-occur more frequently, normalizations are usually applied to the count of co-occurrences and different relevance measures can be obtained, such as word coverage, point-wise mutual information, etc. In our study, we consider a method derived from the coverage-based measure proposed by Sanderson and Croft (1999). In (Sanderson and Croft, 1999), the association of two words is defined by two conditions: $P(a|b) \geqq 0.8$ and $P(b|a) < P(a|b)$. In this definition, word a subsumes word b if the documents in which b occurs are a subset, or nearly a subset of the documents in which a appears. We follow their idea to identify the subsumption word relations in our study. Moreover, several additional conditions are also included in our method, which takes full consideration of the characteristics of document summarization and thus is more suitable for the subsequent summarization process. For example, in document summarization, a document set to be summarized may just contain a few documents (for example, 10 documents per set in the DUC 2004 data set). Therefore, document-level co-occurrences of two words may not be significant enough. We

consider sentence-level co-occurrences in our study instead of the document-level co-occurrences which were used in most previous methods on word relation identification (Sanderson & Croft, 1999; Kummamuru et al., 2004).

Let's first introduce some necessary measures before we begin to introduce how to identify word relations. The ***Spanned Sentence Set*** (*SPAN* for short) of a word w in a document set D whose sentence set is denoted by S_D, is defined as the set of sentences containing w, i.e., $SPAN(w)=\{s \mid s \in S_D \wedge w \in s\}$. The *SPAN* of a word reflects its ability in representing the whole document set at sentence-level.

Given an another word w_0, the ***Concept Coverage*** (*COV* for short) of a word w over w_0 is devised to reflect to what extent w brings new information compared against the already-known information provided by w_0. Based on the definition of *SPAN*, $COV(w|w_0)$ is defined as the proportion of sentences in $SPAN(w)$ that are also in $SPAN(w_0)$, i.e., $COV(w|w_0) = |SPAN(w) \cap SPAN(w_0)| / |SPAN(w)|$. The smaller the coverage is, the more likely w will bring new information to w_0.

For a pair of associated words, we refer the general one as the subsuming word and the other one as the subsumed word. In our method, we constrain the more important word in a pair of associated words always being the subsuming word. Here the frequency-based content model proposed in the last chapter is used to measure word importance. With these conditions, all the word relations in the input documents are identified by examining the words following the descending order of importance. The details of the identification process are given below.

Rank the word list $W = \{w_1, \ldots, w_N\}$ by their importance scores;

For i from 1 to N

    For j from i+1 to N

        If ($COV(w_j|w_i) > \lambda$) //$\lambda$is a pre-given threshold

            Judge that w_j is subsumed by w_i;

In the process, a parameter λ is used to control the strength of the condition for judging word associations. A smaller λ indicates a looser condition and thus more relations can be discovered. Meanwhile, unrelated words may also be wrongly associated. Thus the overall accuracy of all the identified relations may drop. Therefore, a proper λ is actually very important for the method.

With word relations, words can be structurally organized and represented as a graph. Since we assume that only more important words can subsume less important words, the graph does not contain any cycle. Therefore, the constructed word graph is a directed acyclic graph (DAG, a directed graph with no directed cycles), which can also be viewed as a hierarchy in its partial order. Figure 6 and Figure 7 provide an example graph and the corresponding hierarchy to illustrate the idea (There are actually more vertices and edges in the real graph, which are not shown in the figures due to the limited spaces). In the DAG, we also introduce a virtual word (denoted as *ROOT-W*) besides the real words in the input documents. It stands for the center vertex of the DAG, also, the root node of the hierarchy. To every word that is not subsumed by any other word, we regard it as a general word and attach it to *ROOT-W*. To achieve this objective, we define *ROOT-W* as a virtual word that spans all the sentences in the input documents. Therefore, $COV(w|ROOT\text{-}W)$ equals 1 for

every real word *w* and thus the relation between *w* and *ROOT-W* absolutely exists. The reason why we introduce such a virtual word can be explained by the role of the virtual root word in the summarization process, which will be detailed in Section 4.3.1 later when introducing the hierarchical summarization method.

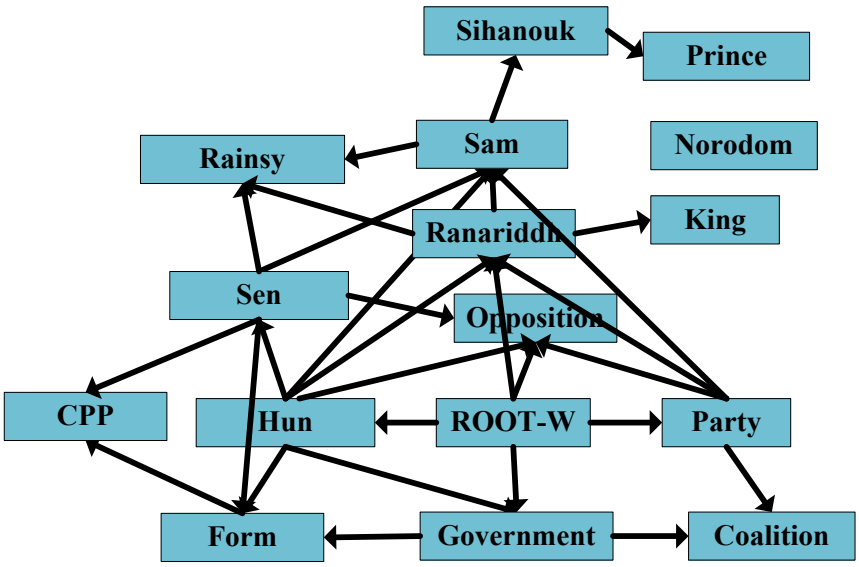

Figure 6. An example of the automatically-constructed word DAG

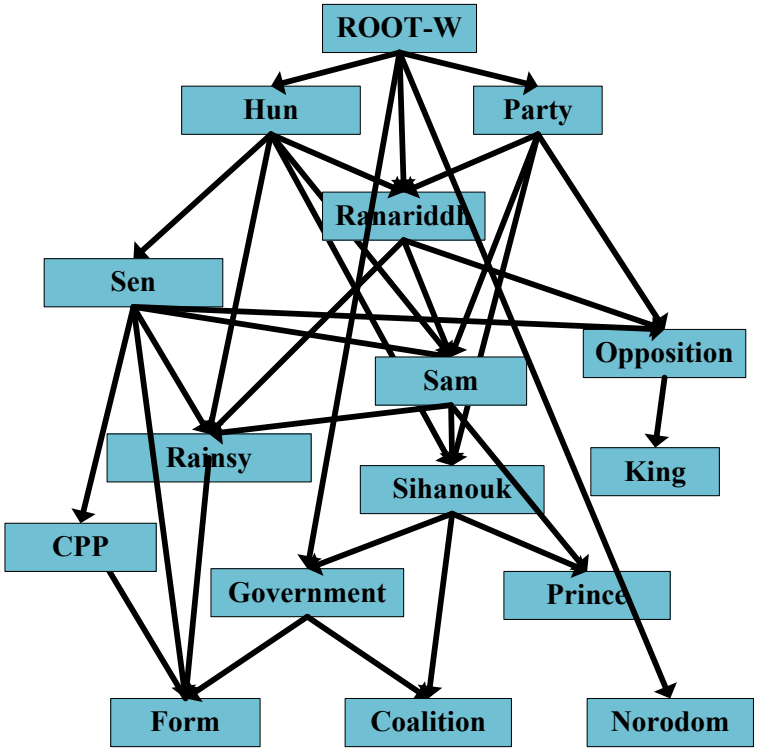

Figure 7. The hierarchical view of the word DAG in its partial order

Transitive Reduction

The above-mentioned subsumption relation between two words reflects the recommendation of one word by the other. Here we also consider a transitive reduction of the relations, i.e., to three words a, b, c that satisfy $a > b$, $b > c$ and $a > c$ ($a > b$ denotes a subsuming b), the long-term relationship $a > c$ will be ignored. The reason is that we prefer to include the subsuming word b into the summary before including the subsumed word c. With the transitive reduction, c will not directly reach by a, and the order $a>b>c$ can be preserved in the summarization process. In practice, the transitive reduction is performed by checking the existing parents of the subsumed word when a new association between two words is to be established. The identification process of transitive reduction is given below and an example hierarchy is also provided to illustrate its effect in Figure 8.

```
Rank the word list W = {w1, …, wN} by their importance scores;
For i from 1 to N
    For j from i+1 to N
        If (COV(wj|wi) >λ)
            Judge that wj is subsumed by wi;
            For each existing parent wk of wj
                If (wi is previously judged to be subsumed by wk)
                    Remove the relation between wj and wk;
```

Comparing the new DAG to the one in Figure 7, we can see that it becomes more concise and compact after transitive reduction. An example redundant relation

is "*Hun Sam*", which is removed due to the existences of "*Hun Sen*" and "*Sen Sam*".

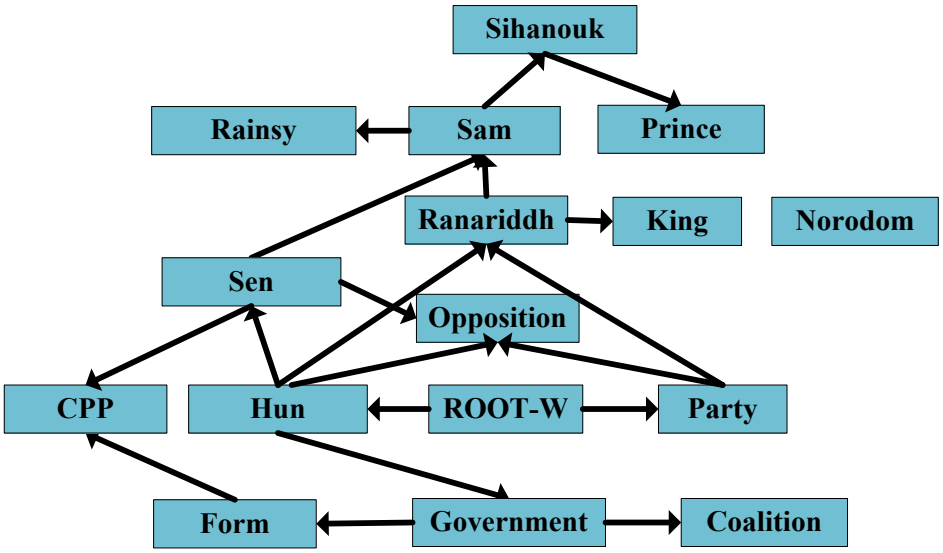

Figure 8. An example word DAG with the transitive reduction

4.2.1.2 Improving the identification method with HAL distance

In real data sets, sentence-level co-occurrences are much sparser than document-level co-occurrences due to the shorter length of sentences. Therefore, the sentence-level coverage between two words is usually much smaller than the document-level coverage. This makes it harder to identify word relations with sentence-level information. To discover more word relations, one possible solution is to extend relations between the words in the same sentence to the words in adjacent sentences. In another word, the cross-sentence co-occurrences need to be considered. Here we adopt the hyperspace analogue to language model (**HAL**) proposed in (Lund & Burgress, 1996). In the HAL model, the relevance between two tokens in different sentences is determined by the distance between the sentences. We follow this idea to develop a function to measure the relevance between two words instead of the co-occurrence-based measure. First of all, a function *f* is used to measure the

relevance of two tokens by the distance between them (denoted as d). Here we adopt the inverse proportion function $f(d)=1/(d+1)$. For an appearance of word w, its relevance to another word w_0 is calculated by its distance to the closest appearance of w_0. The total coverage of w is then calculated as the average relevance of all its appearances, i.e., $COV_{HAL}(w \mid w_0) = \frac{\sum_i f(d_i)}{|SPAN(w)|}$, where d_i is the minimum distance of the ith appearance of w to w_0. Based on the COV_{HAL} measure, we obtain a new relation identification method that is similar to the one proposed in the last section. The difference lies in the substitution of COV by COV_{HAL}. An example word DAG generated with the HAL-based distance measure is provided below in Figure 9.

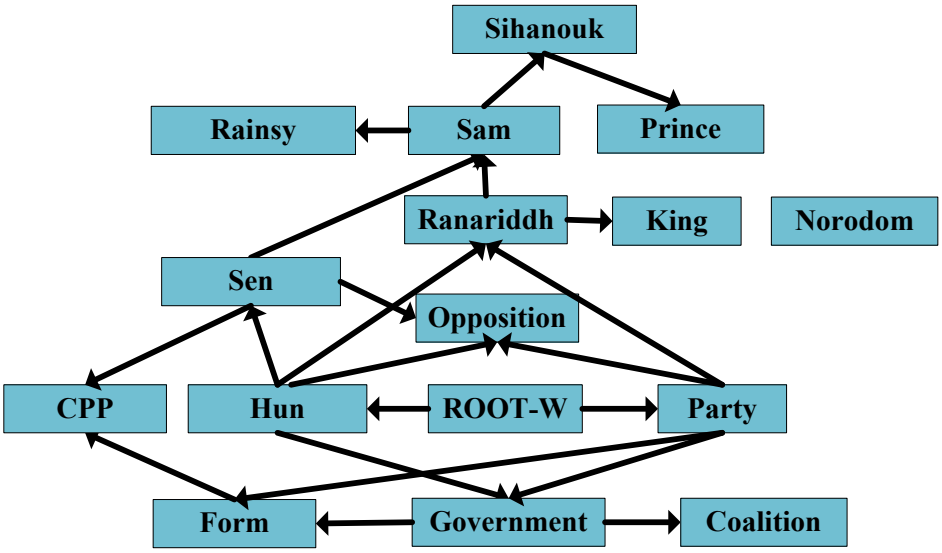

Figure 9. An example word DAG with the HAL distance

Compared to the DAG in Figure 7, the new DAG built with the HAL distance contains more relations under the same threshold λ (0.8), such as the relation between "Party" and "Government". However, it should be noticed that cross-sentence co-occurrences may also lead to wrong word associations since the

relevance between the words in adjacent sentences is less confident than the relevance between the words in the same sentence.

4.2.1.3 Improving the identification method with set-based coverage measure

Another strategy to address the sentence-level sparseness is to introduce set-based coverage measures. By examining the data sets, we observe that a word should be allowed to be covered by multiple words than a single word. For example, there are two common phrases "*King Norodom*" and "*Prince Norodom*" in a DUC 2004 document set. All the appearances of the word "*Norodom*" are almost perfectly split into two sets in which it co-occurs with "*King*" or "*Prince*" respectively. In the input documents, the coverage of "*Norodom*" to "*King*" or "*Prince*" is not larger than the given threshold λ (0.8 here) and thus "*Norodom*" is not recognized as being subsumed by any one of the two words under the concept coverage measure defined in Section 4.2.1.1. On the other side, "*Norodom*" is almost perfectly covered by the set {"*King*", "*Prince*"}. Inspired by this, we'd like to not only consider the coverage between words, but also between words and word sets. With the set-based coverage, more word relations are expected to be discovered.

Given an non-empty word set $W=\{w_1, w_2, \ldots, w_n\}$, the ***Concept Coverage*** (also use *COV* for short) of a new word w over the set W is devised to reflect to what extent w brings new information compared against the known information already provided in W. Similarity to the word-based coverage, $COV(w|W)$ is defined as the proportion of sentences in $SPAN(w)$ that are also in $SPAN(W)$, i.e., $COV(w|W) = |SPAN(w) \cap \cup_i SPAN(w_i)| / |SPAN(w)|$.

Now we use the set-based coverage to identify the pair-wise relations between

words. With this new method, we again follow the progressive style to discover all the relations. For each word, we compare it to the former words to identify the word(s) that can form a subsumption relation with it.

Before introducing the method, let's first discuss the case of comparing a new word w to a former word w_0 which already subsumes a set of words S. To assign a subsumption relation between w and w_0, firstly w should be recognized as being covered by w_0. Moreover, it should not be recognized as being subsumed by any word in S to ensure the transitive constraint. Therefore, there are two conditions in the new method to judge the subsumption relation between w and w_0: (1) $COV(w|w_0) \geq \lambda_1$; (2) $COV(w|S) < \lambda_2$. Here λ_1 and λ_2 are two thresholds used to control the strength of the conditions. For λ_2, we set a constant value (0.8 initially). For λ_1, we consider a word-dependent scheme in which λ_1 is calculated by the maximum coverage of w to each candidate word, i.e., $\lambda_1 = \lambda'_1 * \text{Max}_j COV(w|w'_j)$. Like λ_2, λ'_1 is fixed for all the words. For example, if λ'_1 equals 0.5, it means that a word is regarded as subsuming w when its ability of covering w is at least half of the word that best covers w. According to the definitions, both λ'_1 and λ_2 range from 0 to 1. The effects of the parameters will be examined in the experiment part.

The graph construction process can be also viewed as a process of inserting words into the word hierarchy. The words are inserted into the hierarchy by following the descending order by importance. For each new word w, it is compared to *ROOT-W* as a starting point of the whole process. Then following a top-down approach, it is compare to the existing words in the hierarchy until its final position is found. Once both coverage conditions passes for a word w_0, w will not go any deeper in the hierarchy and is inserted as a new child of w_0; otherwise, it will be compared to the children of w_0. The process continues until w cannot go any deeper

in the hierarchy. In fact, this insertion process is similar to the node inserting algorithm in a B-tree. The algorithm for constructing the whole DAG based on the set-based coverage is given below, in which $DESC(w)$ indicates all the descendants of word w in the current DAG.

Rank the word list $W = \{w_1, \ldots, w_N\}$ by their importance scores;
For i from 1 to N
    Calculate $COV(w_i|\{ROOT\text{-}W\} \cup DESC(ROOT\text{-}W))$; // Check $ROOT\text{-}W$ for w_i
    If $COV(w_i|\{ROOT\text{-}W\} \cup DESC(ROOT\text{-}W)) \geq \lambda_1$
        Calculate $COV(w_i|DESC(ROOT\text{-}W))$;
            If $COV(w_i|DESC(ROOT\text{-}W)) < \lambda_2$
                $ROOT\text{-}W$ is a target parent;
            Else
                For each child node w'_j of $ROOT\text{-}W$
                    Check each w'_j for w_i;

Note that we actually use $COV(w_i|\{w\} \cup DESC(w))$ to substitute $COV(w_i|w)$ in the above algorithm. The idea here is that the concept indicated by a word contains not only itself, but also all of its current descendants. An example of word DAG with the set-based coverage is provided in Figure 10.

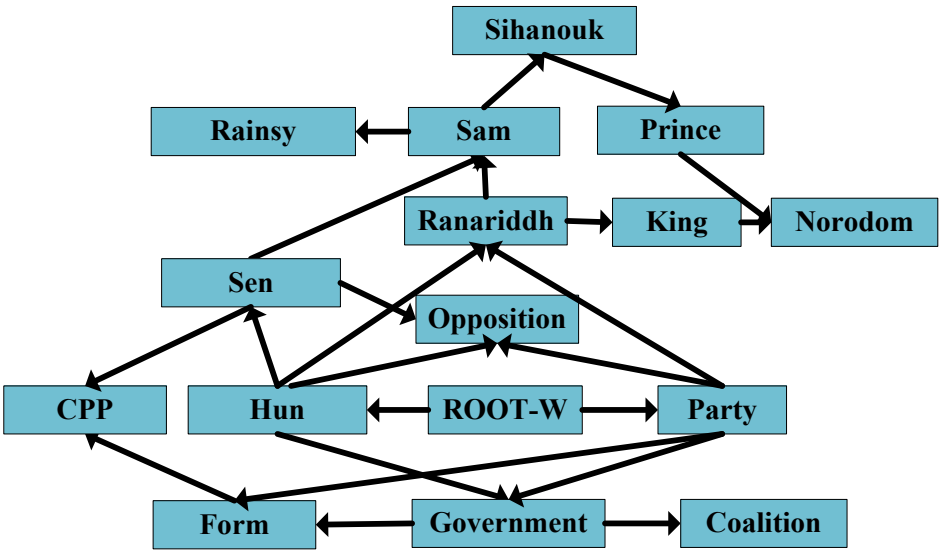

Figure 10. An example word DAG with the set-based coverage

Compared to figure 7, more relations are discovered by using the set-based coverage, such as the relations "*Prince Norodom*"and "*King Norodom*".

4.2.2 Definition of the subsumption sentence relationship

Based on the identified word relations, the subsumption relationship between two sentences is determined by measuring how tightly a sentence *s* can be attached to another sentence *s'*. This is done by considering the words in *s* that can be attached to *s'*.

Denoting the word set of *s* as $W=\{w_1, \ldots, w_l\}$ and the word set of *s'* as $W'=\{w'_1, \ldots, w'_m\}$, we first define the concept of "***Connected Word***". A word w_i in W is regarded to be "connected" to a word w'_j in W' if it satisfies the following condition: $\exists \{w_{l1}, \ldots, w_{lk}\} \subseteq W \cup W'$, s.t. $w_i < w_{l1} \wedge w_{l1} < w_{l2} \wedge \ldots \wedge w_{lk-1} < w_{lk} \wedge w_{lk} < w'_j$. Intuitively, it means that w_i can be locally reached to w'_j in the sub-graph consisting of the words in W and W' only. The weight of the edge that directly connects to w_i formulates the strength of the connection between w_i and w'_j (denoted as

$CON(w_i|w'_j)$). If a word in s is connected to at least one word in W', this word is regarded as being connected to s'.

Based on the definition of the "connected word", the *Conditional Saliency* (***CS*** for short) of s to s' is calculated by the weighted sum of the importance scores of all the "connected words" in s, i.e.,

$$CS(s|\, s') = \sum w_i \in s \text{ Log } (Max\{w'_j \in s_t CON(w_i|w'_j)\} * score(w_i))$$

The conditional saliency is an asymmetric and non-overlapping relationship between two sentences. It measures the probability of extracting sentence s given the condition that sentence s' is already extracted. The Figure 11 below provides an example to illustrate the idea. The first sentence s_1 is "*Cambodia's bickering political parties broke a three-month deadlock Friday and agreed to a coalition government leaving strongman Hun Sen as prime minister*" and the second sentence s_2 is "*Negotiations to form the next government have become deadlocked, and opposition party leaders Prince Norodom Ranariddh and Sam Rainsy are out of the country following threats of arrest from strongman Hun Sen*". As illustrated, the connected words of s_2 to s_1 are "*Rainsy*", "*Rannariddh*", "*Sam*", "*form*", "*Sihanouk*", "*prince*" and "*Norodom*".

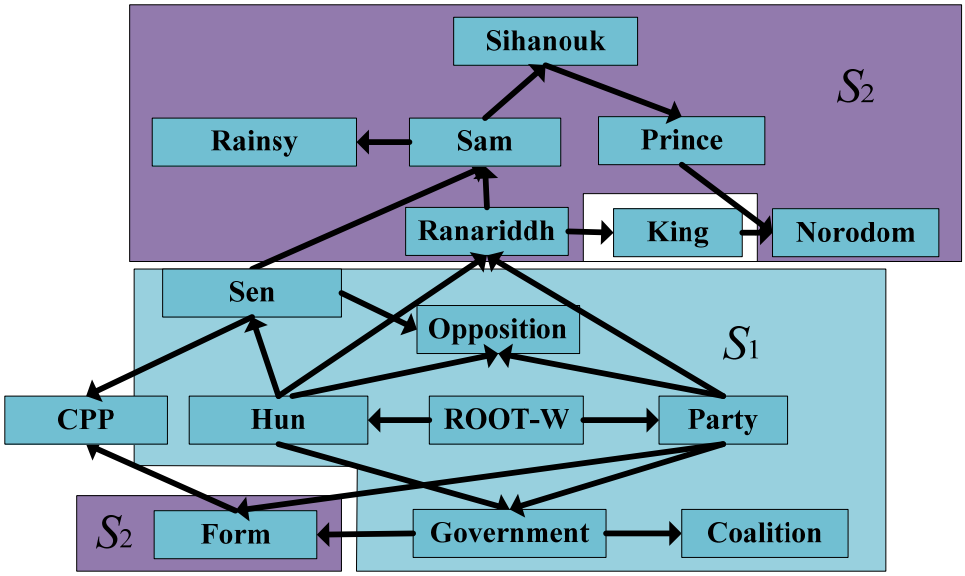

Figure 11. An example of the connected words between two sentences

4.3 The Hierarchical Summarization Method

In this section, we introduce the hierarchical summarization method based on the subsumption sentence relationship.

4.3.1 A conditional sentence selection process

Based on the sentence relationship, we consider a conditional sentence selection process. It can be viewed as a random walking process on the word DAG from central words to their connected words. In the process, summary sentences are selected to cover the central words first and then the other words which are reachable through the word relations. The center of the DAG *ROOT-W* serves as the starting point of the random walking process, which is defined above as a virtual word that spans the whole sentence set. With the virtual root word *ROOT-W*, we further define a virtual sentence *ROOT-S* that contains *ROOT-W* only. This virtual sentence *ROOT-S* is regarded as being selected at the beginning of the sentence selection

process. In fact, the sentences that are connected to *ROOT-S* are the general sentences because the words connected to *ROOT-W* are the general words. Thus we have just unified the identification of general sentences and supporting sentences by using the virtual root word and root sentence. This makes the sentence selection process more consistent.

The sentence selection process is as follows: first setting *ROOT-S* as the initial summary, and then iteratively adding the sentences that best support current summary sentence(s) (denoted as S_{old}). In each round, we calculate a score for every unselected sentence based on its maximum conditional saliency to every selected sentence, i.e., $Max_{s_t \in S_{old}} \{CS(s, s_t)\}$. The maximum conditional saliency reflects how much supporting information provided in s can be brought into the current summary. Therefore, the sentence with the largest saliency score is the one with the most supporting information and thus should be selected into S_{old} in this round (denoted as s_0). Moreover, the existing sentence to which s_0 has the maximum saliency is regarded as the one subsuming s_0 and this information is kept to preserve the relationship between summary sentences.

In this new summarization framework, we still need to consider other influencing factors to achieve better performances. Here we apply the experiences learned from the study in the last chapter. First of all, the length-based normalization is included to tackle the length limit condition. Specific factors in different tasks are also considered. For example, it is proved that position information is very important in generic summarization and thus a position-based modification is included in the hierarchical summarization method. Therefore, the final scoring function based on the conditional saliency for generic summarization tasks is defined as

$$score(s|S_{old}) = Max_{s_t \in S_{old}}\{CS(s, s_t)\} * 1/len(s) * (1\text{-}pos(s))$$

where *len*(*s*) is the total number of words in *s* and *pos*(*s*) is the normalized value of the position of *s* in the document. The adjustments reflect the preferences for sentences with shorter lengths or appearing earlier.

4.3.2 Redundancy control

To control the information redundancy in summarization, we follow the word damping method proposed in Section 3.5. Because the conditional saliency measure is a weighted sum of word scores, the score damping scheme can be directly applied to it. It is also cast as a dynamic modification on the word saliency scores during the conditional sentence selection process. Once a sentence is selected into the hierarchy, the score of each word in it is damped by α, i.e., $score(w_i) = \alpha * score(w_i)$. Therefore, α still indicates the penalizing degree to the covered words. Extremely, when α equals 0, an effective "connected word" is required not to appear in any selected sentence.

4.3.3 Sentence hierarchy construction

Recall that in the conditional sentence selection process, we also keep the information of the subsumption relationship between summary sentences. Therefore, all the summary sentences can be connected to construct a sentence hierarchy. As a matter of fact, it is not necessary to insert all the sentences into the hierarchy when the summary is confined to a given length. When the inserted sentences are enough for composing the summary, we can already stop the construction process. An example sentence hierarchy and the corresponding hierarchical summary are illustrated below.

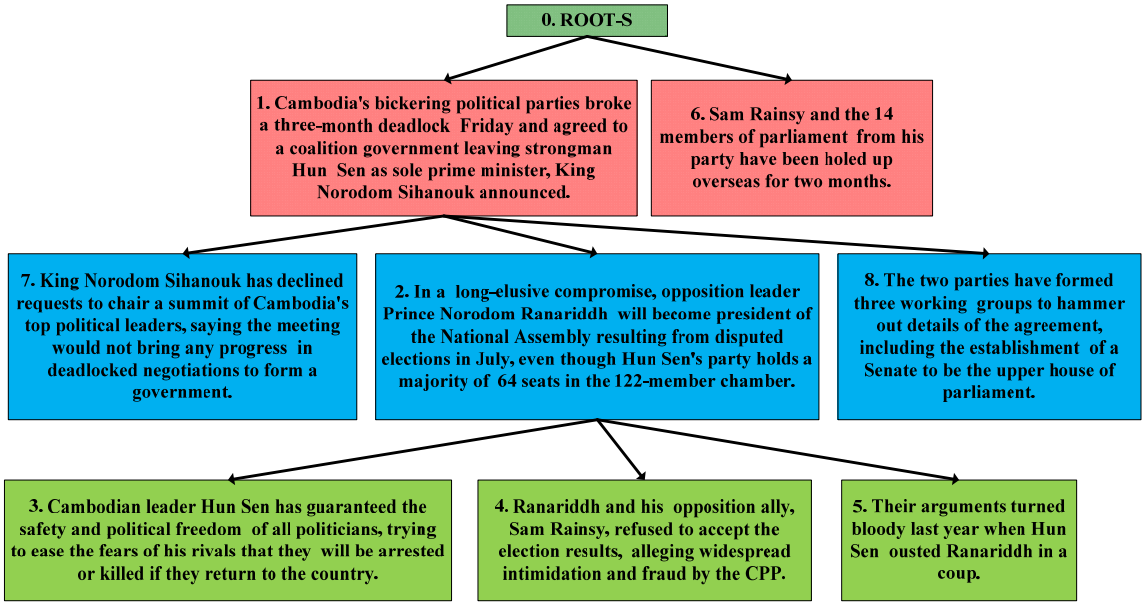

Figure 12. An example sentence hierarchy

A hierarchical summary
Cambodia's bickering political parties broke a three-month deadlock Friday and agreed to a coalition government leaving strongman Hun Sen as sole prime minister, King Norodom Sihanouk announced. In a long-elusive compromise, opposition leader Prince Norodom Ranariddh will become president of the National Assembly resulting from disputed elections in July, even though Hun Sen's party holds a majority of 64 seats in the 122-member chamber. Ranariddh and his opposition ally, Sam Rainsy, refused to accept the election results, alleging widespread intimidation and fraud by the CPP. Their arguments turned bloody last year when Hun Sen ousted Ranariddh in a coup. Cambodian leader Hun Sen has guaranteed the safety and political freedom of all

politicians, trying to ease the fears of his rivals that they will be arrested or killed if they return to the country.

Rainsy and the 14 members of parliament from his party have been holed up overseas for two months.

King Norodom Sihanouk has declined requests to chair a summit of Cambodia's top political leaders, saying the meeting would not bring any progress in deadlocked negotiations to form a government.

The two parties have formed three working groups to hammer out details of the agreement, including the establishment of a Senate to be the upper house of parliament.

4.3.4 Hierarchical summary generation

With the sentence hierarchy at hand, the generation of the hierarchical summary is quite straightforward. For each document set, the hierarchical summarization first follows the above methods to successively construct the word hierarchy and the sentence hierarchy. Then, the summary is composed by the top-level sentences in the sentence hierarchy. Because the sentences are inserted into the hierarchy by maximum conditional saliency, summary sentences can be extracted directly following the order of the inserting sequence until the length limit is reached.

Similar to the frequency-based method, the damping factor method is used for redundancy control here. Therefore, redundancy and saliency are also well integrated in the hierarchical summarization framework. In fact, the hierarchical method is advanced in balancing saliency and redundancy. As discussed in Section 2.2 of the literature review, most existing methods still consider the mentioned concepts when

measuring the saliency scores of new sentences because they rely on a few core words to ensure the saliency of most summary sentences. Experimental results on the damping factor in Section 3.5.4.3 also proves that it is not a good choice to simply ignore the mentioned concepts in the sentence selection process even for the proposed word-based methods that are typical sequential summarization methods.

Different from the word-based methods, the hierarchical summarization method is able to directly examine the uncovered parts of a sentence to measure its saliency. The core strategy to avoid the possible saliency decrease is the usage of word relations. In the hierarchical method, a supporting word is selected by the summary only if it can be connected to a general word that is already selected. Therefore, the saliency of the supporting word is not only ensured by itself, but also by the general word that is more salient. Thus the average saliency of the words in the summary can be improved. In the experiment section, we will show that the hierarchical method is able to achieve satisfactory performances when the damping factor is set to 0.

Since the sentences are selected according to the subsumption relationship in the hierarchical framework, we can also expect that the coherency between these sentences is better than independently selected sentences with sequential methods. Moreover, we can use the sentence relationship to improve the fluency of the summary. An intuitive idea is used here, i.e., the details should be introduced right after the main idea in the summary. In the hierarchical method, this is achieved by listing all the summary sentences by the depth-first order of the sentence hierarchy. Under the depth-first order, the subsumed sentence is placed in the next position of the subsuming sentence and thus the general-to-specific sentence flow persists in the summary.

In fact, the effect of the depth-first ordering strategy is similar to the

cluster-based ordering method introduced in (Barzilay et al., 2002), which grouped similar sentences together in the summary. Moreover, the order between two related sentences is also provided in our method. As stated in Chapter 1, the hierarchical framework has a greater potential for generating fluent summaries than sequential summarization frameworks because it "selects the sentences for fluency" instead of "improving the fluency for the selected sentences".

Because saliency, coverage and fluency are all improved based on the sentence relationship in the proposed hierarchical framework, we regard it as a new way to achieve the target of integrated summarization.

4.4 Experimental Results

We again conduct the experiments on the DUC data sets to evaluate the hierarchical summarization framework. The experiment is first done on the DUC 2004 generic multi-document summarization data set and then extended to the DUC 2005-2007 query-focused multi-document summarization data sets. Various features of the framework are evaluated, including how well the resulting systems discover the important information, control the redundancy, and organize the selected sentences, etc.

4.4.1 A sequential summarization method for comparison with the hierarchical method

We refer the summarization methods introduced in Chapter 3 as sequential methods since they do not consider the relationship between summary sentences. Here we also propose a sequential method which can be directly compared with the

hierarchical method. The sentence scoring function in this method is similar to the one used in the hierarchical method. The only difference is that it sums the importance of all the words in the sentence instead of the "connected words", i.e.,

$$score(s) = \sum w_i \in s \; Score(w_i) * 1/len(s) * (1\text{-}pos(s))$$

In this sequential method, the damping factor α is also applied for redundancy control. Since the main difference between the hierarchical method and the sequential method is the inclusion of sentence relationship or not, the comparison between them can provide a good view on the effectiveness of integrating sentence relationship in document summarization.

Actually, the sequential method implemented in the experiments below is similar to the word-based method introduced in Section 3.5. The main difference is that we only used frequency information in Section 3.5 to prove the effectiveness of the log-frequency hypothesis. This sequential method can be expected to perform better because it has also incorporated position information now.

4.4.2 Experiments on generic summarization

4.4.2.1 A comparison of the hierarchical method and the sequential method on the DUC 2004 data set

The first experiment is conducted on the DUC 2004 generic multi-document summarization data set. Table 12 below provides the average ROUGE recall scores and the corresponding 95% confidence intervals of the hierarchical systems (labeled as **Hier)**, the sequential systems (labeled as **Seq**). Three hierarchical systems are included, which use the co-occurrence-based, HAL-distance-based, and set-based coverage measures for word relations identification (labeled as **Co, HAL, and Set**

respectively**)**. Two sequential systems are included, i.e., the one proposed in Chapter 3 (labeled as **P)** and the one proposed in this Chapter (labeled as **T)**. The parameters in these systems are set as: λ'_1=0.5, λ_2=0.8, α=0.5.

Table 12. Results of the hierarchical systems and the sequential systems on the DUC 2004 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Hier Co	0.3892 (0.3761-0.4021)	0.0992 (0.0898-0.1084)	0.1389 (0.1303-0.1471)
Hier HAL	0.3861 (0.3719-0.3938)	0.0984 (0.0866-0.1042)	0.1357 (0.1280-0.1432)
Hier Set	**0.3911 (0.3784-0.4034)**	**0.1004 (0.0914-0.1089)**	**0.1410 (0.1329-0.1490)**
Seq T	0.3868 (0.3730-0.3990)	0.0967 (0.0865-0.1060)	0.1367 (0.1282-0.1453)
Seq P	0.3788 (0.3667-0.3912)	0.0929 (0.0849-0.1013)	0.1333 (0.1261-0.1404)

The results show that the best hierarchical system **Hier Set** outperforms the best sequential system **Seq T** (more significant than 95% in paired sample *t*-tests). This clearly shows the advantages of the hierarchical framework in discovering the important document content by using word relations.

Comparing different hierarchical systems, the HAL-based system performs worse than the basic co-occurrence-based system. Based on the observations on the constructed word hierarchies, we attribute the reason to the fact that the HAL distance introduces too many noises to the identified word relations. Though more related words are discovered by the cross-sentence co-occurrences, many less related

words are also wrongly associated. Another result is that the HAL-based system even performs worse than the sequential system **Seq T**, which implies that wrong word relations are not helpful and even harmful for discovering the important words. In contrast, the set-based system does perform better than the basic system **Hier Co**, which shows the rationality of the set-based coverage in discovering more reliable word relations.

Comparing the two sequential systems, it is not surprising to see that the one with the additional position information performs better. In the following experiments, we will use the set-based hierarchical system **Hier Set** and the modified sequential system **Seq T** for comparison unless otherwise stated.

4.4.2.2 A Comparison of the methods under different length limits and damping factors

To further compare the hierarchical system and the sequential system, we run the experiments with different length limits and damping factor values. Table 13 below provides the average ROUGE-1 and ROUGE-2 scores (**R-1** and **R-2** for short) of the two systems with three α values (**0**, **0.5** and **1**) and three length limits (**100**, **200**, and **400 words**).

Table 13. Results of the systems under different lengths and damp factors

System	100 R-1	100 R-2	200 R-1	200 R-2	400 R-1	400 R-2
Hier 0	**0.3911**	**0.0920**	**0.5419**	**0.1406**	**0.6731**	**0.1963**
Seq 0	0.3790	0.0895	0.5320	0.1322	0.6714	0.1871
Hier 0.5	**0.3907**	**0.1004**	0.5394	**0.1495**	**0.6723**	**0.2088**

Seq 0.5	0.3861	0.0974	**0.5414**	0.1468	0.6710	0.2058
Hier 1	**0.3828**	**0.0993**	**0.5330**	**0.1485**	**0.6609**	**0.2105**
Seq 1	0.3809	0.0960	0.5287	0.1467	0.6595	0.2100

The results in Table 13 again confirm the advantages of the hierarchical system over the sequential system. The hierarchical system performs better in most cases. Now let's look at the effects of the damping factor. Actually, the results of the sequential system in Table 13 are similar to the results presented in Section 3.5.4.3. It is likely to obtain higher ROUGE-1 scores with larger penalty on repeating words. However, the ROUGE-2 scores may drop significantly. Also, the ROUGE-1 scores cannot be improved in the full penalty cases. In contrast, the hierarchical system can cope with very small α to improve the ROUGE-1 scores without sacrificing the ROUGE-2 scores too much. Since it requires each newly-selected sentence being attached to one of the selected sentences, the saliency of each new sentence can still be ensured even when imposing full penalty on repeating words.

4.4.2.3 Parameter tuning of the hierarchical method

In this set of experiments, the effects of two parameters λ'_1 and λ_2 used in word hierarchy construction are examined. Figures 4 and 5 illustrate the ROUGE-1 scores of the hierarchical system versus λ'_1 and λ_2 respectively both ranging from 0.1 to 1 with a step of 0.1. The results of the two hierarchical systems based on the co-occurrence-based coverage measure and the set-based coverage measure are illustrated in the figures below.

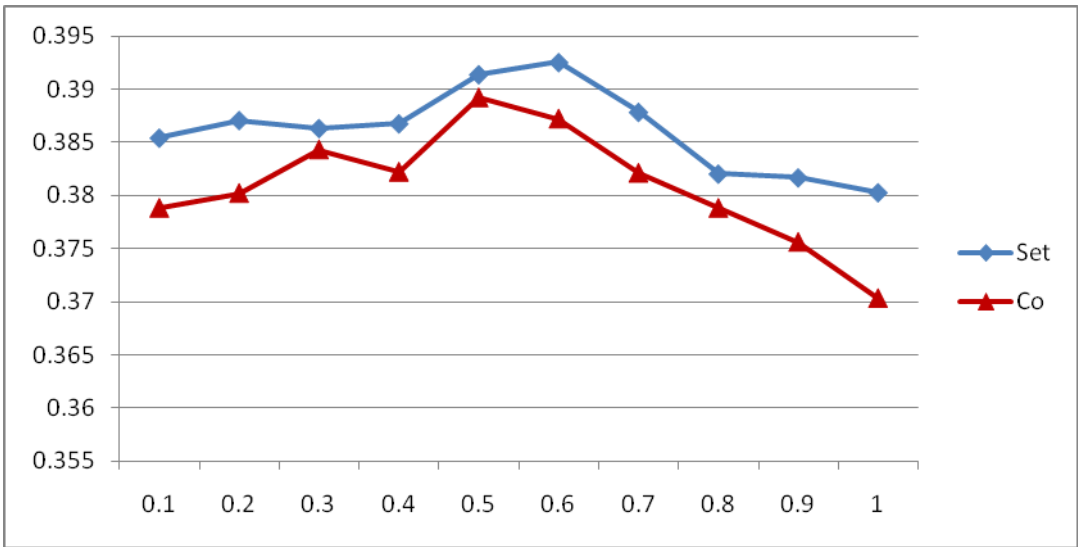

Figure 13. ROUGE-1 versus λ'_1

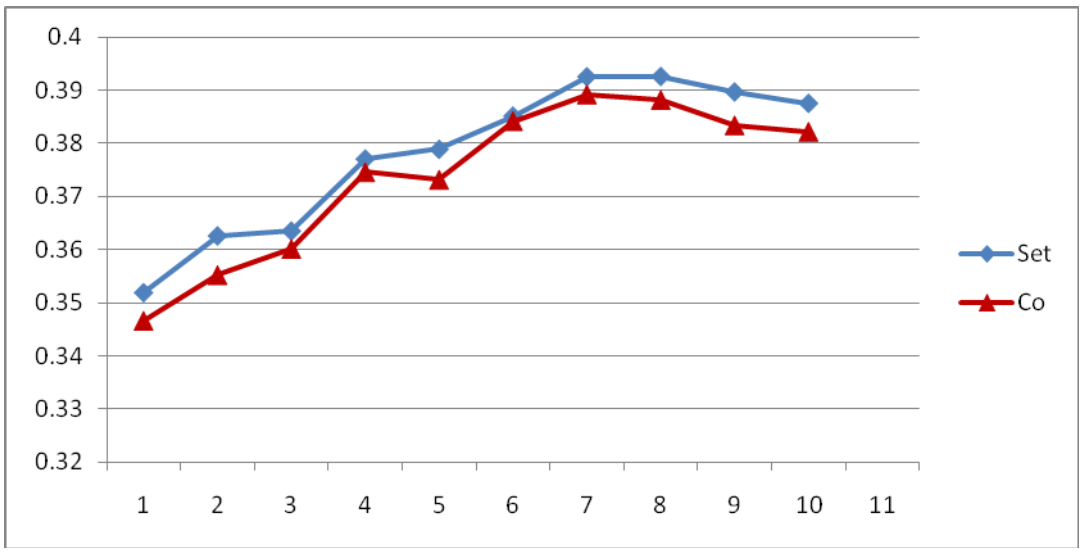

Figure 14. ROUGE-1 versus λ_2

As shown in Figure 14, λ_2 has a great influence on the performance of the hierarchical systems. A smaller λ_2 means a looser condition for judging whether the new word is covered by the descendents or not and thus more candidate words are compared to the new word. Consequently, more parent words may be discovered. However, if λ_2 is too small, many unrelated words may be wrongly associated, which will unavoidably impair the quality of the word hierarchy and lead to worse

performances. On the other hand, when using too large λ_2, the discovered word relations will be too limited and thus the advantages of the hierarchical system will be weakened. Therefore, a proper value of λ_2 is very important for the hierarchy construction method. From Figure 14, we can observe that the best λ_2 is around 0.7 and 0.8.

Similar to λ_2, λ'_1 also plays the role of balancing quantity and quality of the identified word relations. Therefore, a proper λ'_1 is also important. This is clearly shown in Figure 13 and the best value of λ'_1 is observed at around 0.5 and 0.6. However, its influence is not as significant as λ_2 when a good λ_2 is already chosen.

Finally, these two figures further demonstrate the advantages of the set-based coverage in discovering word relations. Its performances are consistently better than the co-occurrence-based measure when parameter values are changed.

4.4.3 Experiments on query-focused Summarization

In this section, we evaluate the effectiveness of the hierarchical framework on the DUC 2005-2007 query-focused summarization data sets. As introduced in Section 3.2, query-focused summarization refers to the task of summarizing a set of documents to serve the information need specified by a given query.

4.4.3.1 Query-driven modifications

Intuitively, it is necessary to consider the effect of the query in the sentence selection process for query-focused summarization. In the query-based hierarchical summarization method, we use the query to refine the word importance measure. A query-based importance measure is defined as the size of the ***Query-based Spanned Sentences Set***, which is the set of the sentences containing both the word under

concerned and at least one non-stop word in the query. In fact, this measure is about equal to the QTF feature used in the learning-based method, which is previously used to substitute the TF feature for query-focused summarization.

Since the words in the word hierarchy are inserted by the descending order of importance, the constructed hierarchy is actually sensitive to the importance measure. By using the query-based importance measure, the words are placed in more proper positions in the word hierarchy to reflect the effect of the query. Examples of an original hierarchy and a query-driven hierarchy are provided below to illustrate the difference. The hierarchies are constructed on the document set **D0701A** from the DUC 2007 set, which talks about "Southern Poverty Law Center and Morris Dee".

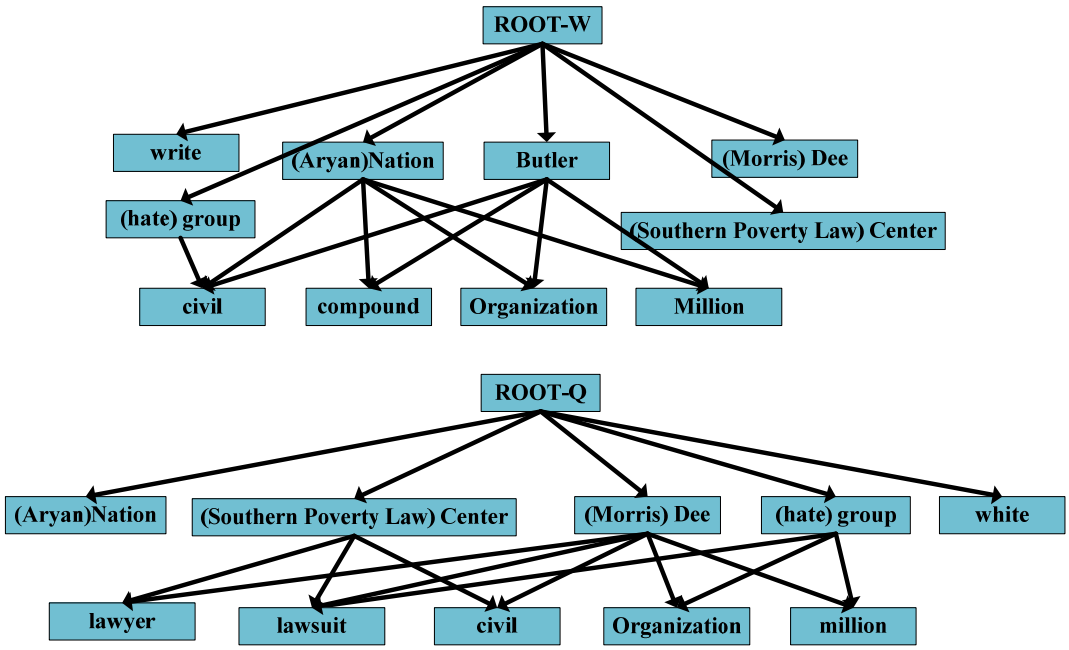

Figure 15. A comparison of the original word hierarchy (above) and the query-driven word hierarchy (below)

In the document set "D0701A", many documents are about the hate group "Aryan Nations" and thus the words such as "Aryan Nations" are first inserted into the hierarchy in the original method. Moreover, other words are likely to be

connected to these words. However, given the query "Describe the activities of Morris Dees and the Southern Poverty Law Center", the general-to-specific style is not well reflected in this hierarchy. On the other hand, we can see from Figure 15 that the query-driven hierarchy is much better in representing the content specified by the query. The top-level nodes are changed to "Morris Dees" and "Southern Poverty Law Center" and thus the hierarchy is more suitable for the query.

Another important modification is to use the query as the new starting point instead of the virtual root word ***ROOT-W*** which stands for the whole document set. The virtual root word now spans all the sentences that contain at least one non-stop word in the query (denoted as ***ROOT-Q*** in Figure 15).

Since the sentence selection process is a random walking process on the word hierarchy, it also becomes query-driven with the above modifications. Thus the bias of the summary to the query is well incorporated into the random walking process. Furthermore, we remove the position-based adjustment from the calculation of sentence saliency scores because it was shown that position information is ineffective in query-focused summarization (Ouyang et al., 2010).

4.4.3.2 Evaluating the effectiveness of query-driven modifications in query-focused summarization

We first examine the effectiveness of query-driven modifications on the DUC 2005-2007 query-focused multi-document summarization data sets. Table 14 below provides the ROUGE-1, ROUGE-2 and ROUGE-SU4 scores of the original hierarchical system (denoted by **Gene**) and the new system with the query-driven modifications (denoted by **Query**) on the DUC 2005, 2006 and 2007 data sets.

Table 14. Results of the hierarchical systems with/without the query-driven modifications on the DUC 2005-2007 data sets

System	ROUGE-1	ROUGE-2	ROUGE-SU4
05 Query	**0.3827 (0.3770-0.3884)**	**0.0742 (0.0709-0.0775)**	**0.1323 (0.1292-0.1355)**
05 Gene	0.3742 (0.3680-0.3803)	0.0704 (0.0671-0.0737)	0.1276 (0.1246-0.1310)
06 Query	**0.4129 (0.4066-0.4190)**	**0.0955 (0.0909-0.1000)**	**0.1494 (0.1454-0.1534)**
06 Gene	0.4059 (0.4000-0.4119)	0.0912 (0.0870-0.0955)	0.1445 (0.1409-0.1483)
07 Query	**0.4449 (0.4384-0.4517)**	**0.1202 (0.1154-0.1252)**	**0.1730 (0.1684-0.1778)**
07 Gene	0.4401 (0.4338-0.4466)	0.1149 (0.1102-0.1197)	0.1674 (0.1631-0.1717)

The experimental results clearly show that the query-driven modifications are effective in practice. The performances of the hierarchical system were consistently improved on all the data sets by applying the modifications.

Of course, the improvements in different years are different. We attribute the reason to the different similarity degrees between the document sets and the queries. Since the given documents are actually the ones retrieved to the query, the important words according to the spanned sentence set and the important words according to the query overlap in some extent. As a matter of fact, the more they overlap, the less information is brought by incorporating the query-driven modifications into the system. Consequently, the performance improvement may not be significant.

4.4.3.3 A Comparison of the hierarchical system and the sequential system in query-focused summarization

In this experiment, we compare the hierarchical and sequential systems in query-focused summarization. Table 15 below provides the ROUGE-1, ROUGE-2 and ROUGE-SU4 scores of the hierarchical system and the sequential system with the query-based modifications on the DUC 2005-2007 data sets. Moreover, the results of the best system from each year are also provided (denoted as **Best**) for reference.

Table 15. The results of the hierarchical system and the sequential system on the DUC 2005-2007 data sets

System	ROUGE-1	ROUGE-2	ROUGE-SU4
05 Hier	**0.3827 (0.3770-0.3884)**	0.0742 (0.0709-0.0775)	0.1323 (0.1292-0.1355)
05 Seq	0.3729 (0.3666-0.3789)	**0.0747 (0.0706-0.0787)**	0.1291 (0.1253-0.1333)
05 Best	0.3767 (0.3716-0.3818)	0.0738 (0.0711-0.0764)	**0.1326 (0.1299-0.1354)**
06 Hier	**0.4129 (0.4066-0.4190)**	**0.0955 (0.0909-0.1000)**	0.1494 (0.1454-0.1534)
06 Seq	0.4019 (0.3964-0.4077)	0.0940 (0.0898-0.0984)	0.1479 (0.1440-0.1521)
06 Best	0.4073 (0.4009-0.4137)	0.0950 (0.0907-0.0992)	**0.1534 (0.1494-0.1574)**
07 Hier	**0.4449 (0.4384-0.4517)**	0.1202 (0.1154-0.1252)	0.1730 (0.1684-0.1778)
07 Seq	0.4314 (0.4252-0.4372)	0.1195 (0.1147-0.1238)	0.1701 (0.1659-0.1743)

07 Best	0.4409 (0.4332-0.4481)	**0.1239 (0.1189-0.1288)**	**0.1750 (0.1701-0.1897)**

The results on the query-focused data sets again show the advantages of the hierarchical system in discovering the important content of the input documents. It consistently outperforms the sequential system on the three data sets. Notably, the hierarchical system performs very close to the best submitted DUC systems in each year.

4.4.3.4 A comparison to the results reported in previous studies

To illustrate the effectiveness of the proposed hierarchical summarization methods, we also compare them to the state-of-the-art methods reported in recent research studies in addition to the systems submitted during the DUC competitions. The methods in the following studies are included: (Wei et al., 2008) who adopted a three-level reinforcement chain model for sentence ranking; (Takamura & Okumura, 2009) who constructed the summary by modeling the summarization process as a maximum coverage problem; (Wan & Xiao, 2009) who considered graph-based multi-modality learning for sentence ranking; (Haghighi & Vanderwende, 2009) who proposed a system based on the hierarchical LDA model; (Wan, 2009) who considered additional topic analysis for graph-based summarization methods; (Wang et al., 2009a) who used hyper-graphs instead of general graphs for sentence ranking; (Wei et al., 2009) who proposed a co-ranking algorithm; (Wang et al., 2009b) who proposed a sentence-based topic model; (Cai et al., 2010) who considered a reinforcement scheme between clustering and ranking; (Shen & Li, 2010) who cast multi-document summarization as a minimum dominating set problem; (Celikyilmaz

& Hakkani-Tur, 2010) who used a hybrid method with hLDA-based training data construction and learning-based sentence ranking scheme. The tables below provide the results on each year, in which **Hierarchical** denotes our hierarchical system. Because not all the ROUGE scores are reported in every study, "-" is used to indicate the missing scores in the tables.

Table 16. Comparison to previous results on the DUC 2004 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
(Shen & Li, 2010)	-	0.0893	0.1314
(Cai et al., 2010)	0.3708	0.0835	-
(Wang et al., 2009b)	0.3907	0.0901	0.1322
(Takamura & Okumura, 2009)	0.385	-	-
Hierarchical	**0.3911**	**0.1004**	**0.1410**

Table 17. Comparison to previous results on the DUC 2005 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
(Wei et al., 2009)	**0.3880**	**0.0802**	**0.1373**
(Wei et al., 2008)	0.3868	0.0779	0.1366
(Wan, 2009)	0.3839	0.0737	0.1317
(Wan & Xiao, 2009)	0.3718	0.0676	0.1293
(Shen & Li, 2010)	-	0.0731	0.1306
Hierarchical	0.3827	0.0742	0.1323

Table 18. Comparison to previous results on the DUC 2006 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
(Wang et al., 2009a)	-	**0.0965**	**0.1525**
(Wan, 2009)	0.4101	0.0886	0.1420
(Wan & Xiao, 2009)	0.4031	0.0851	0.1400
(Shen & Li, 2010)	-	0.0930	0.1480
(Cai et al., 2010)	0.3953	0.0896	-
Hierarchical	**0.4129**	0.0955	0.1494

Table 19. Comparison to previous results on the DUC 2007 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
(Wan & Xiao, 2009)	0.4204	0.1030	0.1460
(Celikyilmaz & Hakkani-Tur, 2010)	**0.456**	0.114	0.172
(Haghighi & Vanderwende, 2009)	0.431	0.118	0.167
Hierarchical	0.4409	**0.1239**	**0.1750**

Our system is able to outperform most existing systems on the data sets, except in very few cases that our results are worse than the system from (Wei et al., 2008) on DUC 2005 and the system from (Wang et al., 2009a) on DUC 2006. Of course, the performance of a summarization system is not only determined by the sentence selection methods, but also depends on the pre-processing and post-processing methods. Therefore, the comparison between methods in different studies is not

absolutely fair. Nevertheless, we can still conclude that our system is able to achieve state-of-the-art performances giving the sufficient results listed above.

4.4.4 Manual experiments

It has been argued that ROUGE criteria only evaluate the content of summaries based on *N*-gram-based statistics. Manual evaluations are also included in our study for a closer look at the effects of incorporating sentence relationship into the summarization method. Overall quality and readability are the two measures to be manually evaluated in this study. The overall quality is defined by how well the summary can substitute the input documents. In fact, according to the experiences learned from the DUC competitions, the overall quality can be well reflected by the ROUGE scores though ROUGE only considers *N*-gram statistics. On the other side, the readability of a summary is referred to its overall quality as an independent text, despite of whether it conveys the important content of the input documents or not. The two measures involve more subsidiary measures, such as coherence, focus, fluency, etc. In our study, we only consider the overall quality and the readability of the summary instead of analyzing the sub-measures. Moreover, the readability is more concerned since it cannot be well reflected by the ROUGE scores.

First of all, we manually compare the overall quality of summaries generated by the hierarchical system and the sequential system. To this end, we consider a 1-5 standard criterion, in which 5 indicates that the summary is very good and 1 indicates that it is very bad. We evaluate on the 50 document sets from the DUC 2006 data set and report the average quality score of the 50 summaries generated by the two systems in Table 20 below.

Table 20. Manual results of the overall quality on the DUC 2006 data set

System	Overall Quality
Hierarchical	**2.9**
Sequential	2.46

We can see that the average quality of the summaries generated by the hierarchical system is better than the ones generated by the sequential system. It proves that the hierarchical system is more advanced in composing summaries for human readers. In fact, the average quality is quite close to the DUC 2006 manual evaluation results, in which the average quality scores of most systems are between 2 and 3.

In the DUC 2006 competitions, the differences between system summaries and human summaries in manual evaluations are much more significant than the differences among the ROUGE scores. In fact, the contents of the system summaries are already acceptable if they can cover most salient concepts in the input documents. However, their readability, especially in the organization of the sentences, is still far away from the level of human summaries.

Then we compare the average readability of the summaries generated by the two systems. In extractive summarization, the grammar correctness of the selected sentences is not a crucial issue since they are the original sentences extracted from the input documents. If the documents are well-written, the quality of the sentences should be high. Therefore, coherency and fluency of the summary are actually more important to the readability issue. In this experiment, three systems that follow the different sentence ordering strategies are compared. In the baseline sequential system (labeled as **Seq Baseline**), the sentences are ordered by the ranking order in the

sentence ranking results. Such an ordering strategy will inevitably lead to the poor readability of the resulting summary. In the second system, we consider a light-weighted ordering strategy for the specified data sets used in our experiments. In the DUC data sets, the input documents are newswire documents with explicit publishing dates, for example the document named "APW19990707.0181" was published in 1999/07/07. Based on this information, we consider a composite ordering strategy in which (1) all the summary sentences are first ordered by the publishing dates of the corresponding documents; (2) the sentences from the same document are further ordered by their original positions in the document. In fact, this strategy combines a chronological ordering method and a original ordering method. We use it to re-order the summaries generated by the sequential system and label the new system as **Seq Re-order**. In the hierarchical system, we use a depth-first search algorithm for traversing the sentence hierarchy to order the summary sentences in a more coherent way. However, for the two sentences with the same parent sentence, the order between them is still unknown. To this kind of sentence pairs, we again apply the above composite ordering strategy. The resulting hierarchical system is labeled as **Hierarchical**. Similar to the above experiment, we also consider a 1-5 standard criterion here. The results of the manual readability evaluation are provided in Table 21 below, still on the 50 document sets from the DUC 2006 data set.

Table 21. Manual results of the readability on the DUC 2006 data set

System	Readability
Hierarchical	**2.26**
Seq Re-order	1.82
Seq Baseline	1.28

The hierarchical system obtains the best average readability score of 2.26, which is close to the results reported in the DUC 2006 competition. Though the average readability of the hierarchical system is not a satisfactory result (less than 3, the borderline), it still outperforms the other two sequential systems. This shows that the subsumption relationship is able to improve the sentence ordering result. On the other side, the advantages of the modified sequential system to the baseline system also prove the effectiveness of the composite ordering strategy, which is based on both the original order and the publishing dates.

We further provide a detailed discussion on the effectiveness of the ordering methods. Firstly, let's analyze the composite ordering strategy by comparing the two sequential systems. From the generated summaries, we observe that this strategy is mostly effective in ordering two adjacent (or very near) sentences selected from the same document. The link between such sentences is usually very tight and thus it is suitable to use the original order for them. In contrast, when the two sentences are far away from one another in the document or even in different documents, the original order is not effective at all. However, it is actually very rare for the two extracted summary sentences to be in adjacent or near positions in the data set. For example, in the DUC 2006 data set, a document set contains 25 documents. Therefore, for a 250 word summary that usually consists of about 10 sentences, summary sentences are likely to come from different documents. Thus the effect of the original order method is quite limited on this data set. For the two sentences from different documents, the composite strategy orders them by the publishing date. However, the estimation of temporal information by the publishing date is very rough and it does not work in most cases. In conclusion, the improvement by the composite strategy mainly relies on adjacent sentences. Its failures on the long-distance sentences limit the

effectiveness.

On the other side, the hierarchical system utilizes the subsumption sentence relationship to help order the long-distance sentences. Two sentences from different documents are placed at adjacent positions if they are related by the subsumption relationship. By this, the readability of the summaries generated by the hierarchical system is further improved. However, we also observe that sometimes the hierarchical system may place adjacent sentences into separated positions unexpectedly. As a matter of fact, the sentence ordering is a very complicated problem and more studies are needed in the future to find a satisfying solution for the problem.

An important issue during the manual evaluation experiments is that it is even very hard for human summarizers to accurately rate the summaries. To make the comparison more credible, we also consider a comparative evaluation scheme to prove the advantages of the hierarchical system. The idea is that it may be easier for human summarizers to judge the preference between two summaries than to exactly measure how good a summary is. In this evaluation scheme, each pair of summaries generated by the two systems on the same document set is compared by human summarizers to judge whether one summary is obviously better than the other. The following Table 22 provides the comparison results: **Hierarchical** indicates the number of document sets in which the hierarchical system is preferred; **Sequential** indicates the opposite; **Tie** indicates the number of document sets in which the qualities of the two summaries are about equal.

Table 22. Manual results of the comparative experiments

System	Hierarchical	Tie	Sequential
Overall Quality	21	29	0
Readability	23	22	5

In the pair-wise comparison scheme, the hierarchical system is much more preferred by human summarizers. It appears that the hierarchical system is significantly superior to the sequential system for potential users of automatic summarization systems on the DUC 2006 data set.

Besides, we also include a pair of example summaries as a case study. Two summaries generated by different systems on the same data set are illustrated below.

Hierarchical summary
1. Cambodia's bickering political parties broke a three-month deadlock Friday and agreed to a coalition government leaving strongman Hun Sen as sole prime minister, King Norodom Sihanouk announced.
1.1. In a long-elusive compromise, opposition leader Prince Norodom Ranariddh will become president of the National Assembly resulting from disputed elections in July, even though Hun Sen's party holds a majority of 64 seats in the 122-member chamber.
1.1.1 Ranariddh and his opposition ally, Sam Rainsy, refused to accept the election results, alleging widespread intimidation and fraud by the CPP.
1.1.2 Their arguments turned bloody last year when Hun Sen ousted Ranariddh in a

coup. 1.1.3 Cambodian leader Hun Sen has guaranteed the safety and political freedom of all politicians, trying to ease the fears of his rivals that they will be arrested or killed if they return to the country.
Sequential summary
1. Cambodia's bickering political parties broke a three-month deadlock Friday and agreed to a coalition government leaving strongman Hun Sen as sole prime minister, King Norodom Sihanouk announced. 2. Cambodian leader Hun Sen on Friday rejected opposition parties' demands for talks outside the country, accusing them of trying to ``internationalize'' the political crisis. 3. Negotiations to form the next government have become deadlocked, and opposition party leaders Prince Norodom Ranariddh and Sam Rainsy are out of the country following threats of arrest from strongman Hun Sen. 4. Hun Sen complained Monday that the opposition was trying to make its members' return an international issue.

As illustrated, the sequential summary tends to include more general sentences since the summary sentences are selected from a general saliency-based ranking result. In contrast, the hierarchical summary goes to supporting sentences more quickly through the sentence relationship and thus it contains both general sentences and supporting sentences. Moreover, the parent-subsidiary relations between the summary sentences improve their relatedness and make the whole summary more

coherent.

By following the depth-first order, the related sentences in the hierarchical summary are placed in adjacent positions. As long as the subsumption relationship between the sentences is correctly identified, the fluency can be improved by following the general-to-specific style.

4.5 Chapter Summary

In this chapter, we consider the relations between the words in the input documents to incorporate more information into the summarization process beyond simply covering more salient words. We propose a hierarchical summarization framework that follows a general-to-specific process to select summary sentences. Under the framework, a word hierarchy is first constructed to structurally organize the salient words to form an overall understanding of the document content. Based on the word hierarchy, we define a conditional saliency measure to model the recommendation relationship between summary sentences, which is used to construct the sentence hierarchy and compose the hierarchical summary. The resulting hierarchical methods are compared to the traditional sequential methods on several DUC data sets. The results clearly demonstrate the advantages of the hierarchical summarization framework in composing more diverse, coherent and well-organized summaries. It is also shown that the framework can be well applied to different summarization tasks by adapting the general-to-specific summarization process.

The idea of hierarchical summarization is novel in the summarization area. We have published a paper (Ouyang et al., 2009) to report the preliminary study on

hierarchical summarization, in which a simpler summarization method was proposed. Nevertheless, the idea of hierarchical summarization is well reflected by the paper and hierarchical summarization is accepted as a new kind of summarization methods which can simultaneously improve the summary on various aspects.

In general, the effectiveness of the hierarchical summarization framework depends on how well the word hierarchy models the concepts in the documents and how well the sentences are inserted into the sentence hierarchy. As a matter of fact, a single word alone is often insufficient to represent a complex concept. Certainly, it may be more accurate to use phrases or other complex representations of concepts. In the next chapter, we will explore more sophisticated concept representations and aim to improve the hierarchical summarization method.

Chapter 5 Hierarchical Summarization Methods Beyond Words

In this chapter, we introduce several further studies on the hierarchical summarization framework, which adopt more sophisticated content representations for the input documents.

5.1 Chapter Overview

We have examined several content models to develop effective word-based summarization methods in Chapter 3 and introduced the use of word relations to develop hierarchical summarization methods in Chapter 4. In those two chapters, we mainly use words to approximately represent the concepts in the input documents. Experimental results in Section 4.4.3.4 show that word-based summarization methods can lead to very powerful systems with state-of-the-art performances. However, single words are often insufficient to represent complex concepts. In this connection, we conduct three further studies in this chapter, which consider content representations beyond single words. The studies are first briefly introduced below and then detailed in Section 5.2-5.4 respectively.

(1) Firstly, we investigate the use of phrases as a supplement of words. We follow a typical key phrase extraction process to identify the indicative phrases in the input documents based on syntactic parsing results and frequency-based statistics. The phrases are used to refine both the word hierarchy construction method and the

hierarchical summarization method.

(2) In the second study, we use WordNet to expand single words to synonym sets (**synset**), which are expected to be better in representing complex concepts. Then, the new hierarchical graph using synsets as vertices instead of words is constructed to develop a synset-based summarization method.

(3) In the last study, we use the hierarchical Latent Dirichlet Allocation (hLDA) (Blei et al., 2004) model to generate a set of hierarchically-organized topics for the input documents. Based on the topic hierarchy, we propose several hLDA-based summarization methods.

Since the definition of vertices is changed in these studies, other components of the hierarchical summarization method should be modified accordingly. Firstly, the relations between vertices are re-defined to obtain the edges in the new text graphs. After that, the sentence selection methods are modified based on the new graphs. In the following sections, we will introduce the methods sequentially, along with the corresponding experiments.

5.2 Phrase-based Methods

We now consider phrases as a supplement of words. As illustrated in Figure 9, the constructed word hierarchy may involve unnecessary relations when using words as vertices, for example, the relation between "*Hun*" and "*Sen*". In fact, "*Hun Sen*" is a person name and thus the two words together form a complete concept. Therefore, the general-subsidiary relation does not exist here. In practice, this kind of relations may yield bloated and less reasonable word hierarchies. One solution to this problem is to use phrases as vertices. Since phrase boundaries are not explicitly provided in

plain texts and thus an automatic phrase extraction process is needed in the phrase-based method.

5.2.1 Previous work on key phrase extraction

Existing phrase extraction methods involve two fundamental steps: the candidate phrase identification step and the key phrase selection step. Usually, the candidate identification step is cast as a filtering process in which the unavailable *N*-grams are removed from the candidates. For example, in (Frank et al., 1999), the *N*-grams with stop-words at the beginning or the end, those appearing only once in the document, or single proper nouns, were all eliminated. Medelyan and Witten (2006, 2008) further used thesaurus to filter the unavailable candidates. Parsing was also considered for phrase boundary identification (Barker & Corrnacchia, 2000). In (Wan & Xiao, 2008), simple POS-based patterns were used instead of parsing.

After the candidate identification step, key phrases are extracted from the candidates according to the importance scores which are usually estimated by multiple features, such as word frequencies, phrase frequencies, POS-tags, etc. The features are normally combined by either handcrafted heuristic rules or learned importance scoring functions (Turney, 1999; Medelyan & Witten, 2008).

When we participated in the key phrase extraction task of SemEval-2, we proposed a method based on keyword extraction and word expansion. From our experiences learned from the SemEval-2 task, the performances of current key phrase extraction methods are still not satisfactory, especially on the exact key phrase boundary determination. On the other hand, as shown in previous experiments, the proposed hierarchical summarization method is quite sensitive to the precision of word relations. In this connection, we consider a light-weighted

modification in the study instead of substituting all the words by phrases. We mainly consider collocations, a category of phrases that can be more accurately identified, to improve the hierarchical summarization method.

As a matter of fact, collocations are depended on the local context of the input documents. For example, two words "*southern*" and "*center*" are collocated in a document set about "*Southern Poverty Law Center*", but they may not be related in other document sets. Therefore, we choose frequency-based statistics to identify collocations for this local problem. Moreover, in order to improve the efficiency, we include a candidate identification process based on syntactic parsing results, which can greatly reduce the scope of candidate phrases.

5.2.2 Key phrase identification

We follow the typical two-step process to discover collocations in the input documents. Firstly, we use the syntactic parsing results of sentences for candidate phrase identification. For each sentence, we use the Stanford-Parser[5] to obtain its parsing tree and also the POS tags of the words in it. As an example, the parsing result of the sentence "*King Sihanouk declined to chair talks in either place.*" is illustrated in Figure 16 below.

[5] Available at http://nlp.stanford.edu/software/lex-parser.shtml

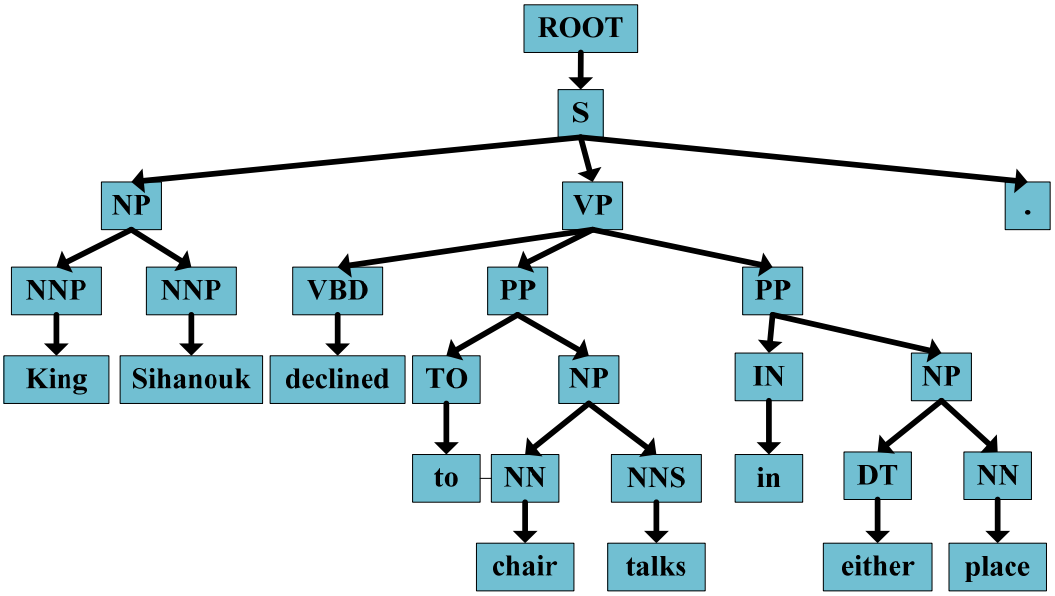

Figure 16. An exmaple parsing result by the Stanford-Parser

For each sentence, we only consider NPs and VPs in the parsing results as candidate phrases. Moreover, only content words are considered, i.e., nouns, verbs, adjectives and adverbs. Though quite simple, the heuristics are able to filter most *N*-grams that are not suitable as key phrases. In the above example, the candidate phrases are "*King Sihanouk*" and "*chair talks*".

We then measure the candidate phrases by their frequencies in the input documents to decide if they are collocated words. Usually, the structures of collocations are very stable. Based on this idea, collocations can be judged by the size of the sentence set in which the distances between the words are uniform. Denote the distance between the two words w_1 and w_2 in a sentence s as $DIST(w_1, w_2, s\}$. A distance-based coverage measure COV^D is defined as

$$COV^D(w_2|w_1, d)=|\{s|\ s\in SPAN(w_1) \wedge s\in SPAN(w_2) \wedge DIST(s, w_1, w_2)= d\}| \ / \ |SPAN(w_2)|$$

In fact, $COV^D(w_2|w_1, d)$ is the proportion of co-occurrences with a fixed distance d between the words.

Then, we regard that w_2 collocates with w_1 if the following condition holds:

$\text{Max}_{d \in \boldsymbol{Z}} COV^{D}(w_2|w_1, d) > 0.8$, $\boldsymbol{Z}$ is the integer set. In another word, two words are regarded as a collocation when the distances between most of their co-occurrences are uniform.

5.2.3 Phrase-based modifications on the hierarchical summarization framework

Using the identified phrases, we consider an absorption strategy to reduce the unnecessary edges in the original word DAG. For two collocated words, the less important one is absorbed by the other one and it is no longer explicitly appear in the word DAG. This is visually illustrated in Figure 17 below, where the absorbed words are enclosed within the round brackets.

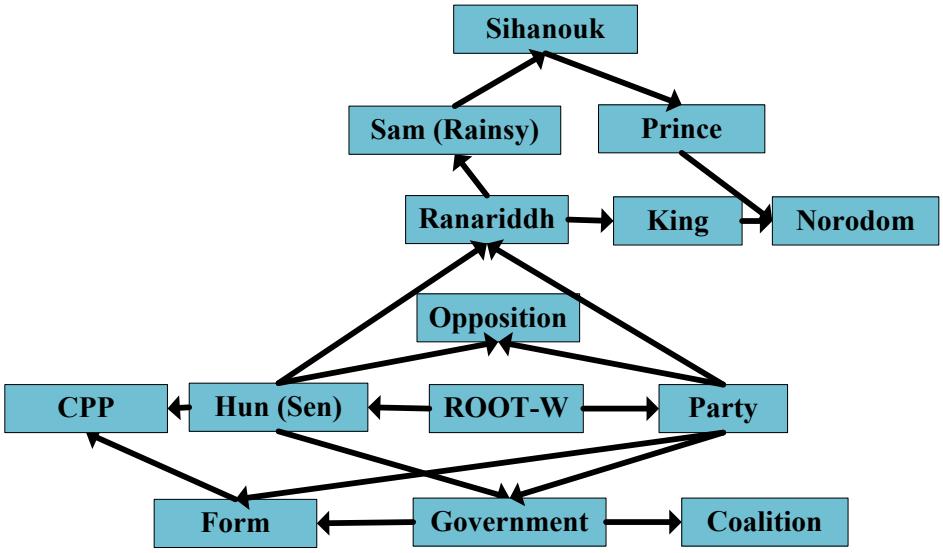

Figure 17. An example of the modified word DAG with the absorption strategy

As illustrated in the above example, unnecessary relations such as "*Hun Sen*" or "*Sam Rainsy*" are removed and thus the word DAG becomes more compact and accurate.

The summarization process using the new word hierarchy is almost the same as

the original process. The difference is that the absorbed words are now ignored in the random walking process on the word hierarchy. Once the core word of a node is covered, other words in the node are regarded as being covered as well.

5.2.4 Experimental results

To examine the effectiveness of the phrase-based modification, we conduct a comparative experiment on the original hierarchical system (denoted as **Word**) and the modified system with the word absorption strategy (denoted as **Phrase**). The same parameters are used for the two systems except the absorption strategy. Thus the gap between performances can directly reflect the effectiveness of the strategy.

5.2.3.1 Experimental results on the generic summarization data set

Experiments are first conducted on the DUC 2004 data set. The average recall scores of the ROUGE criteria of the two systems are reported in Table 23 below, along with the corresponding 95% confidence intervals.

Table 23. Results of the phrase-based system on the DUC 2004 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Word	0.3911 (0.3784-0.4034)	0.1004 (0.0914-0.1089)	0.1410 (0.1329-0.1490)
Phrase	**0.3924 (0.3787-0.4054)**	**0.1038 (0.0943-0.1133)**	**0.1439 (0.1353-0.1523)**

Looking at the results, the phrase-based system does perform better than the word-based system, though not very significant (*P* value equals 0.176 under the

pair-wise *T*-test on the ROUGE-1 scores). A fact to be noticed is that the number of identified phrases is much less than the number of words. This means that the two systems are very close in most cases and this may explain why the improvement is not very significant. In Figure 18 below, we further provide the ROUGE-1 results of the two systems on each document set.

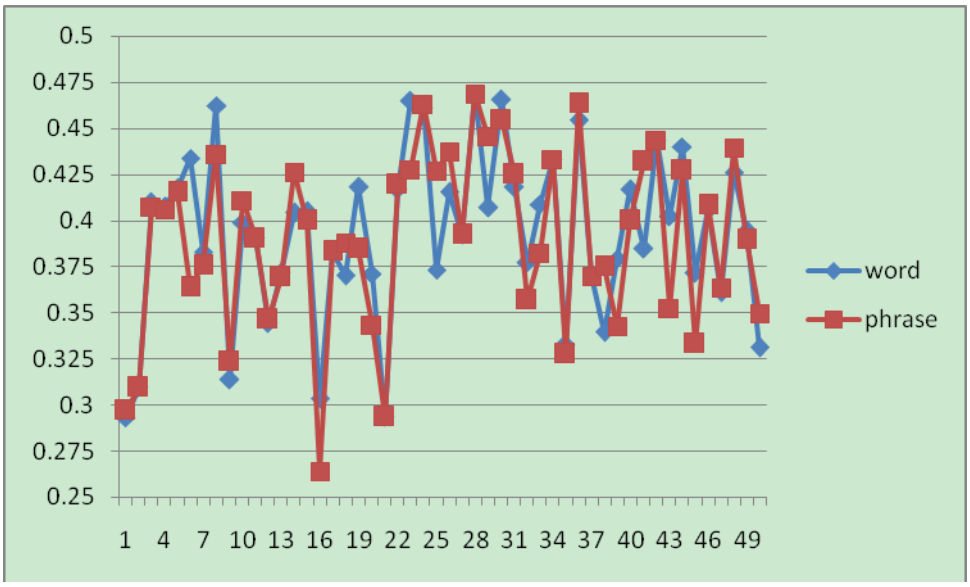

Figure 18. The ROUGE-1 scores of the phrase-based system on each set

From the figure, we can observe that the performances of the two systems are equal on most document sets. Moreover, the phrase-based modification is not always better. We attribute the variety of the performances to the different accuracies of the phrase identification results in the document sets.

5.2.3.2 Experimental results on the query-focused summarization data sets

To further test the effectiveness of the phrase-based modification, we also

compare the phrase-based system to the original system on the query-focused summarization data sets. For query-focused summarization, the query-based modifications introduced in Section 4.4.2 are included in both systems. The average recall scores of the ROUGE criteria of both systems are reported in Table 24 below, along with the corresponding 95% confidence intervals.

Table 24. Results of the phrase-based system on the DUC 2005-2007 data sets

System	ROUGE-1	ROUGE-2	ROUGE-SU4
05 Word	0.3827 (0.3770-0.3884)	0.0742 (0.0709-0.0775)	0.1323 (0.1292-0.1355)
05 Phrase	**0.3833 (0.3775-0.3888)**	**0.0753 (0.0722-0.0784)**	**0.1327 (0.1296-0.1359)**
06 Word	0.4129 (0.4066-0.4190)	0.0955 (0.0909-0.1000)	0.1494 (0.1454-0.1534)
06 Phrase	**0.4132 (0.4076-0.4184)**	**0.0968 (0.0925-0.1009)**	**0.1505 (0.1469-0.1543)**
07 Word	**0.4449 (0.4384-0.4517)**	0.1202 (0.1154-0.1252)	0.1730 (0.1684-0.1778)
07 Phrase	0.4446 (0.4382-0.4512)	**0.1233 (0.1185-0.1283)**	**0.1735 (0.1691-0.1779)**

The results again indicate that the phrase-based system outperforms the original system. And similarly, the improvement is also not significant. We still attribute it to the sparseness and imperfectness of the phrase identification result. Nevertheless, the effectiveness of the phrase-based modification is further proved by the consistently better performances of the modified system on the three data sets.

5.3 WordNet-based Methods

Although incorporating collocations into the word hierarchy is able to improve the performance, its effectiveness is confined by the limited number of collocations recognized from the input documents. In our study, we also consider the WordNet synsets as an alternative way for single word extension. WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets. It also provides general definitions and records the various semantic relations between the synonym sets. A synset consists of a set of words or phrases with the same semantic senses, such as {"*dog*", "*domestic dog*", "*Canis familiaris*"} that stands for the concept "*dog*". In our study, we use the WordNet synsets to solve the problem of the variety of word choices for expressing the same concept in the input documents.

5.3.1 The mapping scheme from words to synsets

To apply the synsets to the hierarchical summarization framework, we need to develop a scheme to map from words in the input documents to synsets. As a matter of fact, not all the words in the input documents can be found in WordNet as it is impossible for WordNet to cover all the concepts in the real world (this is somehow the limitation of WordNet). Such a problem should be addressed in the mapping scheme. Moreover, we also add a constraint into the mapping scheme, i.e. a word can only be mapped to one synset. This constraint is required to avoid ambiguity during sentence ranking. If a word is allowed to be mapped to multiple synsets, it will be ambiguous to estimate its importance in different synsets unless we can accurately identify the ambiguous senses, which is actually very hard in practice. Therefore, we

choose to avoid this problem by adding the unique mapping constraint.

In practice, we use a progressive algorithm for synset mapping, which discovers the appropriate synsets for the words in the input documents following the descending order of word importance. For each word, if it is covered by the synset of a previous word, this synset is used as the target synset. Otherwise, the word is searched in WordNet to find the target synset. In cases there are no synset found in WordNet, we simply use itself as a pseudo synset. The details of the mapping scheme are given below.

Rank the word list $W = \{w_1, \ldots, w_N\}$ by their importance scores;

For i from 1 to N

    If w_i is not covered by existing synsets

        Search in WordNet and add the synset syn_i;

    Else if w_i is covered by synset syn_j

        Add w_i to syn_j ;

5.3.2 Synset-based hierarchical summarization method

With the word mapping scheme, a synset-based summarization framework is developed, which is actually very similar to the word-based framework. The main difference is that words in text graphs are substituted by synsets based on the mapping scheme. Here we re-explain the hierarchical framework at synset-level to propose the synset-based summarization method.

Similar to the word frequency feature, we use the size of the spanned sentence

set of a synset as its saliency score. Since a synset is a set of words, the spanned sentence set of a synset *syn* can be defined as

$$SPAN(syn) = \cup_{w \in syn} SPAN(w)$$

And the coverage measure between a synset *syn* and multiple synsets $SYN = \{syn_1, \ldots, syn_n\}$ can be defined as

$$COV(syn \mid SYN) = \mid SPAN(syn) \cap (\cup_i SPAN(syn_i)) \mid / \mid SPAN(syn) \mid .$$

With the re-defined measures, we can follow the algorithm introduced in Section 4.2 to identify the relations between synsets. The only difference to the word hierarchy construction algorithm is that all the actions are carried out on synsets instead of words. An example synset-based DAG is provided below to illustrate the idea.

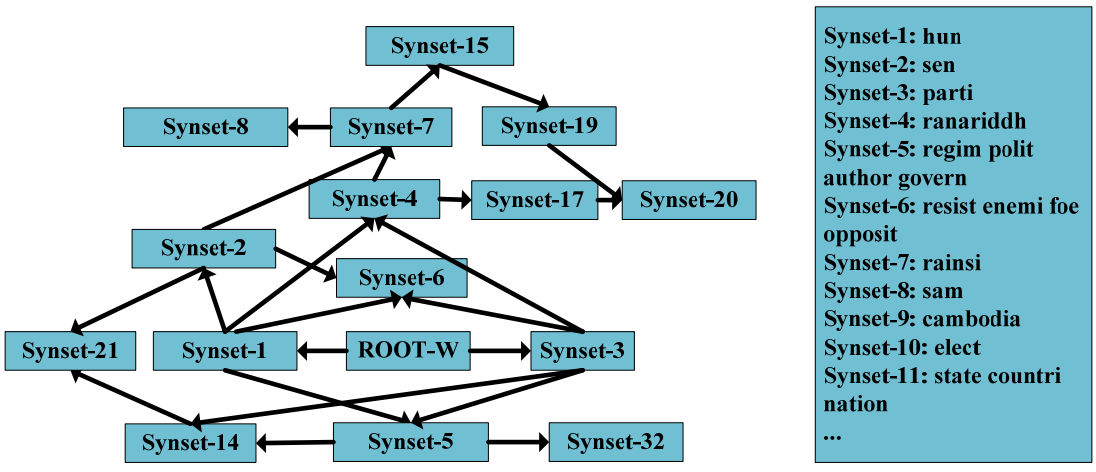

Figure 19. An example of the synset-based DAG

With the synset-based DAG, a synset-based summarization method is used to summarize the input documents. The sentences are selected to cover the synset hierarchy by following the general-to-specific order, which is similar to the covering process of the word hierarchy. Correspondingly, the criterion for sentence selection

is modified as

$$CS(s \mid s_t) = \sum_{syn_i \in s} \log(Max_{syn'_j \in s_t} CON(syn_i \mid syn'_j) * score(syn_i))$$

Finally, the damping factor is applied on the synsets instead of words. When a sentence is selected, the saliency scores of the corresponding synsets are penalized by multiplying the damping factor, i.e.,

$$score(syn_i) = \alpha \cdot score(syn_i)$$

Using other words, these modifications on the word-based method can be simply viewed as a substitution of words by synsets in the input documents. For example, a text segment "*Hun Sen and the opposition party*" is substituted by "*Synset-1 Synset-2 # # Synset-6 Synset-3*", in which the new words such as "*Synset-*" indicate the corresponding synsets and # stands for the stop-words ignored in the mapping scheme. In this view, the synset-based method is almost the same to the original word-based method. The only difference is to substitute words with synsets, which enables the unions of different words with the same semantics. Therefore, the vertices in the synset-based DAG are expected to be closer to real concepts and thus may yield better summaries.

5.3.3 Experimental results

To test the effectiveness of the synset-based modifications, we conduct the experiments to compare the original word-based hierarchical system and the synset-based hierarchical system on DUC 2004-2007 data sets.

5.3.3.1 Experimental results on the generic summarization data set

Again, experiments are first conducted on the DUC 2004 data set. The average

recall scores of the ROUGE criteria of both systems are reported in Table 25 below, along with the corresponding 95% confidence intervals.

Table 25. Results of the synset-based system on the DUC 2004 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Word	0.3911 (0.3784-0.4034)	0.1004 (0.0914-0.1089)	0.1410 (0.1329-0.1490)
Synset	**0.3918 (0.3763-0.4046)**	**0.1014 (0.0928-0.1102)**	**0.1415 (0.1330-0.1501)**

Similar to the phrase-based system, the performance of the synset-based system is slightly improved from the word-based system. However, the improvement is even less significant (*P* value equals 0.43 under the pair-wise *T*-test on the ROUGE-1 scores). Therefore, the effectiveness of the synset-based modifications is more doubtable and needs further research.

5.3.3.2 Experimental results on the query-focused summarization data sets

To further examine the effectiveness of the synset-based modifications, experiments are extended to the DUC 2005-2007 query-focused summarization data sets. The average recall scores of the ROUGE criteria of both systems are reported in Table 26 below, along with the corresponding 95% confidence intervals.

Table 26. Results of the synset-based system on the DUC 2005-2007 data sets

System	ROUGE-1	ROUGE-2	ROUGE-SU4
05 Word	**0.3827 (0.3770-0.3884)**	**0.0742 (0.0709-0.0775)**	**0.1323 (0.1292-0.1355)**
05 Synset	0.3573 (0.3516-0.3623)	0.0627 (0.0601-0.0652)	0.1174 (0.1145-0.1200)
06 Word	**0.4129 (0.4066-0.4190)**	**0.0955 (0.0909-0.1000)**	**0.1494 (0.1454-0.1534)**
06 Synset	0.4051 (0.3991-0.4107)	0.0862 (0.0821-0.0900)	0.1415 (0.1380-0.1451)
07 Word	**0.4449 (0.4384-0.4517)**	**0.1202 (0.1154-0.1252)**	**0.1730 (0.1684-0.1778)**
07 Synset	0.4223 (0.4164-0.4293)	0.0996 (0.0955-0.1041)	0.1544 (0.1502-0.1585)

Surprisingly, the synset-based system performs consistently worse than the original system on all the three data sets. This suggests that synsets are even less effective than single words in representing concepts on these data sets. We attribute the reason to the clash between the local characteristics of the input documents and the global senses of WordNet synsets. In our task, the senses of words are determined under the local context. In contrast, WordNet is built for general applications. Without a disambiguation process, locally-unrelated words may also be combined by WordNet synsets and thus unimportant words maybe wrongly recognized as important words. Another issue is that the length limit of summaries is 250 words in the query-focused summarization task, thus the wrongly-recognized words in these data sets may become more than those in the DUC 2004 generic data set which requires 100-word summaries only. This is a possible reason of the greater failures of the synset-based system on the DUC 2005-2007 data sets.

5.4 hLDA-based Methods

In the above sections, we have attempted to use phrases and synsets to improve the hierarchical content representation of the input documents. Experimental results show that better representations of documents can lead to better summaries, but worse representations may yield worse performances. In fact, the main differences of these methods to the word-based summarization method are how to define and formulate vertices in text graphs. The algorithms for relation identification and sentence selection are not much changed yet. In this section, we consider breaking the limit of the word-based framework by using more free-style text graphs. To achieve this objective, two crucial issues should be considered:

(1) the assignment of all the appearances of one word to multiple nodes; and

(2) the relation between two nodes with common words.

Here we consider the hierarchical Latent Dirichlet Allocation (hLDA) model, the hierarchical version of Latent Dirichlet Allocation (LDA) (Blei et al., 2003; 2004), as a possible solution. LDA is a type of generative models that allow sets of observations to be explained by unobserved groups, revealing why some parts of the data are similar. For a set of documents, hLDA is able to automatically discover a set of latent semantic topics, which are represented as word distributions on the vocabulary of the input documents. Moreover, hLDA organizes all the semantic topics as a hierarchical tree. Therefore, the hLDA model naturally addresses the two issues mentioned above. In the following sections, we will introduce the hLDA model and propose several hLDA-based summarization methods.

5.4.1 Hierarchical Latent Dirichlet Allocation

Given a set of documents $D=\{d_1, d_2, \ldots, d_N\}$ and its vocabulary $W = \{w_1, w_2, \ldots, w_M\}$, LDA assumes that there are a set of latent topics $Z=\{z_1, z_2, \ldots, z_K\}$ (K is pre-specified). Each document d_j is viewed as a mixture of the topics in Z and each topic z_k is a distribution over the word vocabulary W. Two kinds of distributions, the per-document topic distributions $p(Z|d_j)$ and the per-topic word distributions $p(W|z_k)$, are both modeled by multinomial distributions with Dirichlet priors. Different from the well-known bag-of-words model, the words are assumed to be independent given the topics in LDA instead of given the documents, i.e.,

$P(w_i) = \Sigma_i P(w_i|z_k) P(z_k)$

In the hierarchical version of the LDA model (hLDA), the *nested Chinese Restaurant Process* (nCRP) is used to model the topic distribution instead of the multinomial distribution. Under the nCRP, topics are organized by a hierarchical tree instead of a sequential topic list. According to the description in (Blei et al., 2004), an nCRP can be defined by imagining the following scenario. Suppose that there are an infinite number of infinite-table restaurants in a city, and one restaurant is regarded as the root restaurant. On each table in the root restaurant, there is a card with the name of another restaurant. On each table in these restaurants, there are also cards with the names of even more restaurants, and this structure repeats infinitely. A constraint held here is that each restaurant is referred exactly once by the cards. By this way, all the restaurants in the city are organized into an infinitely-branched tree. Note that each restaurant is associated with a level in this tree. The benefit of using nCRP is that we do not need to pre-define the structure of the tree, but just give the maximum level.

Different from the original LDA model in which words are generated from single topics, hLDA generates a word from all the topics in a path of the hierarchical tree, from the root topic to one of the leaf topics. The full generative model of hLDA is given below.

1. Let c_1 be the root restaurant.
2. For each level $l \in \{2,...,L\}$: (L is the maximum depth of the tree)
   (a) Draw a table from restaurant c_{l-1} using the nested Chinese restaurant process.
3. Draw an L-dimensional topic proportion vector θ from Dir(α).
4. For each word $l \in \{1,...,N\}$:
   (a) Draw $Z \in \{1,...,L\}$ from Mult(θ).
   (b) Draw w_n from the topic associated with restaurant c_z.

Since the hLDA model is too complex for extract inference, approximating inference algorithms are usually used. In our study, we follow the implementation of the Mallet toolkit[5], which uses the Gibbs sampling process to sample the posterior of the nCRP and the latent topics. The Gibbs sampler provides a method for simultaneously exploring the parameter space (the topics of the corpus) and the model space (the L-level tree). From the sampling process, we can obtain the latent tree, the level assignment for every word and the path assignment for every input document. Here we just use the Mallet toolkit to obtain the hLDA topics. More details of the hLDA model can be found in (Blei et al., 2004). From the output results of Mallet, we choose the word assignment results of topics to construct the

[5] Available at http://mallet.cs.umass.edu/

topic hierarchy, in which each topic is denoted by a set of words assigned in the sampling process. An example topic hierarchy is given below to illustrate the results generated by Mallet.

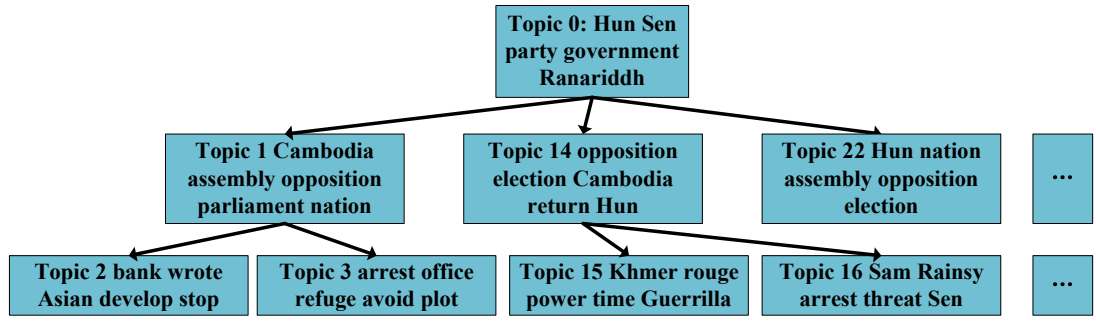

Figure 20. Exmaples of a hLDA topic hierarchy

In our study, we intend to use the hLDA model to explain sentence-level co-occurrences between words. Therefore, we use all the sentences in the input documents as the input to hLDA. We also consider a filtering strategy to make the co-occurrence information more accurate, i.e., only the valuable sentences in the input documents are considered as valid inputs. A sentence is regarded as valuable when it satisfies the following conditions: (1) the total number of words in it is not less than 6; (2) it contains at least one of the non-stop words in the query (for query-focused summarization only).

5.4.2 The hLDA-based summarization methods

In this section, we introduce the hLDA-based summarization methods. An initial idea is to directly apply the topic hierarchy to the proposed hierarchical summarization method, i.e., we need to model the hLDA topics as the vertices of text graphs. This can be achieved by choosing the most dominative words in each topic as a vertex. Thus the topic hierarchy is transformed to a graph with similar characteristics to the word-based and synset-based graphs and similar summarization process can be applied.

However, the generative model of hLDA actually follows a different idea with the proposed summarization method. In hLDA, word co-occurrences are explained by the paths in the topic hierarchy. So the two words in the same topic may be unrelated. Moreover, the information of the words except the dominative words in a topic is lost when converting topics to vertices in this way. Therefore, we believe that it is not suitable to mechanically apply the hLDA topic hierarchy to the hierarchical method proposed in Chapter 4 directly. In this connection, we develop several particular summarization methods adapted to the hLDA model.

5.4.2.1 Method 1: summarizing from the root topic

In (Haghighi & Vanderwende, 2009) who applied hLDA to summarization, they used the root topic only to select summary sentences, considering that the root topic represents the most general concepts in the input documents. We follow their idea to develop the first hLDA-based summarization method that also only uses the root topic for sentence selection. As illustrated in Figure 20, the frequencies of the sampled words in the topic are provided in the Gibbs sampling results. Based on this

information, we use the topic-based frequency to measure the word saliency instead of the global **TF** feature for the word-based ranking method proposed in Section 3.5. This time, the score of a sentence *s* to a topic *Z* is calculated as

$$score(s \mid Z) = \frac{\sum_{w_i \in s} \log freq(w_i \mid Z)}{\mid s \mid}$$

The above formula is used to calculate the importance score to the root topic as the ranking score. With the ranking scores, the summarization process of this hLDA-based method is almost the same to the word-based method proposed in Section 3.5.

5.4.2.2 Method 2: summarizing from each topic

In the study by Wang et al. (2009b) who used LDA to discover semantic topics for summarization, they adopted a sentence selection strategy that selected one sentence from each topic following the descending order of topic importance. We follow this idea to develop the second hLDA-based method, which also selects summary sentences from topics one after another. According to the hierarchical structure of hLDA, we consider a general-to-specific sentence selection process that selects summary sentences from the root topic to the leaf topics. We use the concept of "*acting topics*" to carry out the sentence selection process. Here the acting topics indicate those topics that are considered in a particular round of sentence selection.

The sentence scoring function in Method 1 is used to measure the saliency of a sentence to a particular topic here. The ranking score of a sentence is then calculated by its maximum saliency score to every "acting topic". In other words, the sentence is selected to cover one of the acting topics as much as possible. When a topic is covered, it is removed from the acting topic list and all its children topics are added

into the list. By this way, the topics will be covered from root to leaves. The details of the selection process are described below.

Set root topic Z_0 as the initial acting topic;

While the summary does not exceed the length limit

Rank the sentences by Max_i $score(s|Z_i)$, in which only the Z_is in acting topic list are considered;

Select the top-ranked sentence s_0 and assume that the corresponding topic is Z_k;

Foreach subtopic Z_j of Z_k

Add Z_j to the acting topic set;

Remove Z_k from the acting topic set;

5.4.2.3 Method 3: summarizing from each path

In hLDA, an input document is sampled from all the topics in a path of the topic hierarchy rather than a single topic. Therefore, we also consider a strategy that selects summary sentences from each path instead of each topic. For a leaf topic Z_i, we combine all the topics on the path from the root topic Z_0 to Z_i. All the sampled words in these topics are assembled together to form a composite topic $Z_{c,i}$. Then, the sentences are ranked by the maximum saliency score to each composite topic, i.e.,

$$Max_j \frac{\sum_{w_i \in s} \log freq(w_i \mid Z_{c,j})}{|s|}$$

5.4.2.4 Method 4: summarizing from root topic to leaf topics

We also consider another general-to-specific summarization method by

gradually traversing the paths in the topic hierarchy. Similar to Method 2, when a topic is successfully used to select a sentence, we will consider the child topics of it in the next round. As a matter of fact, the parent topic may not be fully covered by the selected sentences yet and thus we still need to consider the remaining words of the parent topic when selecting new sentences from the child topics. This is done by passing the remaining words in the parent topic to each child topic.

We use a sentence selection process that is similar to the one used in Method 2, i.e., first setting the root topic as the initial acting topic and then moving to subtopics when summary sentences are being selected. The modification is: when a topic Z_j contributes to a sentence selection action and is removed from the acting topic list, the remaining words that are not covered by the selected sentence are passed to each subtopic Z_k of Z_j, i.e.,

$$freq(w \mid Z_k) = freq(w \mid Z_k) + freq(w \mid Z_j)$$

This new Method 4 is advanced compared to Method 2. Because an hLDA topic is distributed on the whole word vocabulary of the input documents, one single sentence is not able to cover the whole topic. Therefore, the strategy used in Method 2 may miss many important words when it only selects one sentence from one topic. Differently, in Method 4 the uncovered words in the topic can still be reached in the next rounds of sentence selection by passing words along. Thus the problem of missing important words can be relieved.

5.4.3 Experimental Results

5.4.3.1 Experimental results on the generic summarization data set

We first compare the four methods on the DUC 2004 data set (denoted as

Method 1-4). Notice that Method 1 and Method 3 both follow the sequential summarization fashion. In contrast, Method 2 and Method 4 involve a hierarchical summarization process. Therefore, we include both the sequential method and the hierarchical method proposed in the last Chapter for reference. The average recall scores of the ROUGE criteria of all the systems are reported in Table 27 below, along with the corresponding 95% confidence intervals.

Table 27. Results of the hLDA-based systems on the DUC 2004 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Hierarchical	**0.3911 (0.3784-0.4034)**	**0.1004 (0.0914-0.1089)**	**0.1410 (0.1329-0.1490)**
Sequential	0.3868 (0.3730-0.3990)	0.0967 (0.0865-0.1060)	0.1367 (0.1282-0.1453)
Method 1	0.3823 (0.3689-0.3955)	0.0946 (0.0866-0.1031)	0.1365 (0.1286-0.1441)
Method 2	0.3521 (0.3379-0.3663)	0.0783 (0.0701-0.0865)	0.1197 (0.1122-0.1275)
Method 3	0.3863 (0.3717-0.4004)	0.0970 (0.0876-0.1062)	0.1401 (0.1318-0.1491)
Method 4	0.3881 (0.3741-0.4021)	0.0983 (0.0891-0.1071)	0.1399 (0.1316-0.1488)

From the above table, several results are observed:

(1) Comparing the two hLDA-based sequential systems (with Method 1 and Method 3) and the reference sequential system, the system with Method 1 that follows the strategy proposed in (Haghighi & Vanderwende, 2009) performs the worst. Therefore, we argue that it may not be appropriate to consider the words in the root topic only since the information in other topics is missed. In contrast, the

performance of the system with Method 3 that summarizes from each path is better than the reference system. Since sentences are actually generated from each path in the generative model of hLDA, theoretically this method is more sound and thus the better performance can be achieved.

(2) Comparing the two hLDA-based hierarchical systems (with Method 2 and Method 4) and the reference hierarchical system, the system with Method 2 that follows the strategy that hierarchically summarizes from each topic performs very badly. Its performance is significantly worse than all other systems. The reason is that it also ignores many important words. Compared to Method 1, it may even miss the words in the root topic since only one sentence is selected for each topic. Because the words in the root topic are especially important in the input documents, the word missing problem is even more serious in this method. This causes the deterioration of the performance. On the other hand, the system with Method 4 avoids the word missing problem by passing the uncovered words in an "activated" topic to its subtopics. The results clearly prove the soundness of this strategy. The system with Method 4 performs much better than the system with Method 2. However, it still cannot outperform the hierarchical system proposed in the last Chapter.

(3) The hierarchical system with Method 4 performs slightly better than the sequential system with Method 3, which is viewed as another proof of the advantages of hierarchical summarization over sequential summarization.

5.4.3.2 Experimental results on the query-focused summarization data sets

To go one step further, the hLDA-based methods are also examined through the

experiments conducted on the DUC 2005-2007 data sets. The average recall scores of the ROUGE criteria of all the systems are reported in Tables 28-30 below, along with the corresponding 95% confidence intervals.

Table 28. Results of the hLDA-based systems on the DUC 2005 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Hierarchical	0.3827 (0.3770-0.3884)	0.0742 (0.0709-0.0775)	0.1323 (0.1292-0.1355)
Sequential	0.3729 (0.3666-0.3789)	0.0747 (0.0706-0.0787)	0.1291 (0.1253-0.1333)
Method 1	0.3831 (0.3781-0.3882)	0.0724 (0.0693-0.0753)	0.1313 (0.1284-0.1342)
Method 2	0.3397 (0.3343-0.3450)	0.0567 (0.0543-0.0590)	0.1096 (0.1070-0.1124)
Method 3	0.3851 (0.3796-0.3901)	0.0760 (0.0730-0.0787)	0.1344 (0.1315-0.1373)
Method 4	**0.3862 (0.3810-0.3912)**	**0.0763 (0.0732-0.0792)**	**0.1344 (0.1313-0.1375)**

Table 29. Results of the hLDA-based systems on the DUC 2006 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Hierarchical	**0.4129 (0.4066-0.4190)**	**0.0955 (0.0909-0.1000)**	**0.1494 (0.1454-0.1534)**
Sequential	0.4019 (0.3964-0.4077)	0.0940 (0.0898-0.0984)	0.1479 (0.1440-0.1521)
Method 1	0.4098 (0.4047-0.4157)	0.0903 (0.0861-0.0944)	0.1467 (0.1433-0.1503)
Method 2	0.3661 (0.3606-0.3717)	0.0695 (0.0660-0.0728)	0.1240 (0.1207-0.1272)

Method 3	0.4076 (0.4019-0.4133)	0.0933 (0.0889-0.0975)	0.1458 (0.1425-0.1493)
Method 4	0.4093 (0.4038-0.4147)	0.0946 (0.0909-0.0985)	0.1460 (0.1424-0.1497)

Table 30. Results of the hLDA-based systems on the DUC 2007 data set

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Hierarchical	**0.4449 (0.4384-0.4517)**	**0.1202 (0.1154-0.1252)**	**0.1730 (0.1684-0.1778)**
Sequential	0.4314 (0.4252-0.4372)	0.1195 (0.1147-0.1238)	0.1701 (0.1659-0.1743)
Method 1	0.4327 (0.4259-0.4390)	0.1123 (0.1078-0.1168)	0.1658 (0.1613-0.1702)
Method 2	0.3961 (0.3894-0.4029)	0.0904 (0.0861-0.0954)	0.1418 (0.1376-0.1463)
Method 3	0.4310 (0.4242-0.4378)	0.1190 (0.1142-0.1238)	0.1703 (0.1659-0.1748)
Method 4	0.4319 (0.4252-0.4387)	0.1195 (0.1147-0.1244)	0.1708 (0.1664-0.1753)

The experimental results on the query-focused summarization data sets are similar to the results observed on generic summarization data sets. The systems with Method 3 and Method 4 perform better than the systems with Methods 1 and 2. Meanwhile, the gaps between different systems became smaller. Note that Methods 3 and 4 even performe better than the word-based hierarchical system on the DUC 2005 data set. This result indicates the potentials of hLDA in developing effective summarization methods. It can be alternative way to explore hierarchical summarization besides the word-based methods introduced in Chapter 4.

5.4.3.3 Discussion

The main problem of applying the hLDA model to document summarization is insufficiency of input data. As a probabilistic model, the statistical characteristics of hLDA are ensured by large corpora and it is initially used on large collections of documents. However, in the document summarization data sets, a document set usually contains just a few documents, for example, 10 documents per set in DUC 2004 and 25-50 documents per set in DUC 2005-2007. This is why we consider sentences as the input of hLDA instead of documents. The same strategy was also adopted in most existing methods that used hLDA for document summarization (Haghighi & Vanderwende, 2009; Celikyilmaz & Hakkani-Tur, 2010). However, sentences are much shorter than documents and thus sentence-level information may be less reliable than document-level information. Abnormal sentences are common in data sets, which are likely to be noises. However, it is harder to identify abnormal sentences than abnormal documents due to the smaller granularity. So the use of hLDA on document summarization still needs much more studies in future.

5.5 Chapter Summary

In this chapter, we conduct several extended studies upon the hierarchical summarization framework proposed in the last chapter. We mainly consider the problem of refining the definitions of vertices and edges of the text graphs that are used to model the input documents.

In the first study, we investigate word collocations in the input documents in order to remove unnecessary relations in the word hierarchy. Phrase candidates are first identified from the parsing results of sentences and then verified by a

frequency-based measure. Then we consider an absorption strategy that treats the words in a collocation as a whole unit in the summarization process. Experimental results show that the absorption strategy does improve the performance of the hierarchical summarization method. It is also observed that the effectiveness is quite depended on the accuracy of the phrase identification result.

In the second study, we investigate the WordNet synsets as another possible content representation beyond words. Intuitively, synsets should be better in representing concepts than single words and thus can be expected to be more suitable for constructing concept graphs. However, experimental results show that synsets are actually ineffective in our task. From the analysis on summarization data sets, we attribute the reason to the clash between the global definition of the synsets and the local context of the input documents.

Besides, we also try to adopt hLDA to model the local context of the input documents. We develop four methods that follow different strategies to select summary sentences based on the topic hierarchy generated by hLDA.

From the studies, we derive some important conclusions based upon the hierarchical summarization framework. For the definition of vertices in the hierarchical content representations, the fact that WordNet-based and hLDA-based methods fail to outperform the word-based method shows that the word-based method is actually quite successful in implementing the idea of hierarchical summarization. It seems to be unnecessary to employ too complex concept definitions to model the document content. On the other hand, the success of the phrased-based refinement shows that it is effective to introduce additional collocation relation. This result suggests that exploring more kinds of relations may be a future direction for developing better hierarchical summarization methods.

Chapter 6 Conclusion and Future Work

This dissertation presents a series of studies on document summarization. The studies begin with estimation of the saliency of the words in the input documents in order to develop effective word-based summarization methods. It is our next objective to incorporate the word relations into the summarization process. A hierarchical summarization framework that takes into account the subsumption sentence relationship is developed. After that, several further studies are conducted with the aim to improve the proposed hierarchical framework.

The main studies and contributions of the dissertation include:

(1) We propose a learning framework for word saliency estimation and compare three kinds of machine learning models. Experiments are conducted on three authoritative data sets to evaluate the learning framework and many valuable results are observed. The main contribution of this study is the introduction of regression models to document summarization. We show that regression models are actually better than classification models and learning-to-rank models for word saliency estimation, which have been used in most previous researches.

From the initial results, we further develop a word-based summarization method based on frequency information only. It is not only very effective with state-of-the-art performances, but also very efficient with the simple and solid methodology. It captures the log-linear relationship between the frequency in the input documents and the real word saliency. We believe that this relationship exists in most summarization tasks and thus the method can be applied to different summarization tasks. Moreover, because it takes into account the frequency

information only, it can serve as a good prototype method, which has great potentials for further refinement in specific summarization tasks.

Another important characteristic of the method is that it has actually integrated two different summarization objectives, i.e., saliency and redundancy. The two objectives are well combined in the single sentence selection process. Therefore, this study is also regarded as an important step to the main target of the dissertation, i.e., to integrate different summarization objectives.

(2) Based on the frequency-based method, we further consider the use of word relations in document summarization. This is actually the main contribution of this dissertation.

Since word relations are not given in the input documents, we need to develop word relation identification methods. We analyze the characteristics of document summarization in order to design more suitable methods for the problem in our study. For example, we require the word relations to be transitive to reduce the redundancy relations. Also, a set-level coverage measure is proposed to model the relations among multiple words. These considerations are all proved to be effective in the experiments conducted on the DUC data sets.

Then, we define the subsumption sentence relationship based on the identified word relations, which is crucial for the hierarchical summarization framework. In the framework, the relationships between the selected and unselected sentences are used to develop a conditional sentence selection method. In the conditional sentence selection process, a novel conditional saliency measure is defined based on the sentence relationship. Experimental results clearly show that it is more effective than traditional global saliency measures in discovering the important content in the input documents.

In the study, we also manually evaluate the quality of the summaries generated by the novel hierarchical summarization framework. It is shown that the sentence relationship is not only able to improve the saliency of the output summary, but also improve the other aspects, such as coherence and fluency.

With the proposed methods and the experimental results, we introduce a new type of summarization methods, i.e., hierarchical summarization. We also proved that hierarchical summarization has many good characteristics. Therefore, it can be regarded as a type of method with great potentials worth further studying in the future.

(3) In the further studies on hierarchical summarization, we improve the word-based hierarchical method by refining the word hierarchy with phrases. We also try to make use of the WordNet dictionary and the hLDA model to obtain more kinds of hierarchical content representations for the input documents. The results show that the hLDA model is another flavour of hierarchical summarization. Compared to the word-based summarization systems, the hLDA-based systems can perform comparably well. Nevertheless, the performance of the word-based method is still slightly better. From these studies, we again confirm the power of the proposed word-based summarization method. Moreover, we show a possible direction to improve the method, i.e., to introduce more kinds of relations into the summarization process.

Although the extensive studies in this dissertation have told a continuous story on hierarchical document summarization and resulted in very powerful summarization methods, a number of issues still need to be addressed.

(1) In learning-based word saliency estimation, we use the same scoring function for all the document sets. However, different document sets may have

different characteristics and thus a single scoring function may not be always suitable. This problem is not well studied in previous researches because it is very hard to perfectly model the influence factors of the true word saliency. In our future work, we will consider adaptive learning models as a possible solution.

(2) In Chapter 4, we mainly investigate the subsumption relations between the words for hierarchical summarization. After that, we consider word collocations in Chapter 5. Collocations can just be viewed as one type of word relations. Experimental results show that the performance is improved by considering additional collocation relations. As a matter of fact, there are more types of relations between the words as well as the relations between sentences, which can be explored. If we can introduce more types of relationships, such as the cause and effect relationship or the follow-up relationship so that the subsequent summarization process can be closer to the human summarization process, better summaries can be expected. There were existing studies that tried to model the various sentence relationships, such as (Zhang et al., 2003). However, in these studies, the relationships were just classified according to sentences overlapping. The accuracy of the classification does not live up to the requirements of our summarization methods.

To improve the hierarchical summarization method by the above idea, two main issues should be given high priority in future work, i.e., a proper relationship definition to model the recommendations between sentences and a reasonably good method that can convincingly identify the relationships.

(3) Regarding the hierarchical summarization framework, currently we follow the extractive style to construct the hierarchical summary. However, a selected sentence may unavoidably contain unexpected words besides the “connected words”

because the candidate sentences are all from the input documents. A selected sentence may not always be ideal to express the concepts embedded in the "connected words". However, due to the limitation of natural language generation techniques, an automatic summarization system still cannot freely compose ideal sentences as human summarizers do. In the future, we would also like to investigate some other means to break the limitation of using the original sentences in the input documents, such as to try sentence compression or fusion techniques, which can generate additional candidate sentences and may express the "connected words" more accurately.

(4) In Chapter 5, we consider a WordNet-based method to replace words with synsets for vertices in the text graphs. However, experimental results show that the synsets are somehow not suitable to solve the problem. The clash between the global definitions of WordNet synsets and the local context of input documents is the main cause of the ineffectiveness. On the other hand, the probabilistic hLDA model does a better job in representing the local context and thus better performances are achieved with the hLDA-based summarization methods. In the future work, we'd like to examine other language models and statistical models to see if they can be used to better represent the content of the input documents in order to develop better hierarchical summarization methods.

(5) Through the whole dissertation, we follow a bottom-up order to consider the document summarization problem, i.e., starting with words, then sentences, and finally the whole summary. As a matter of fact, there is also other additional sentence-level, document-level and even summary-level information, which may also be beneficial for composing better summaries. The question of how to incorporate such kind of information into the hierarchical framework is also an

important issue to be considered in the future.

In conclusion, the summarization framework that considers sentence relationship, such as the hierarchical summarization framework, is a novel way to integrate different objectives of document summarization. Our studies presented in this thesis have sought to make a distinctive contribution towards this goal.

Bibliography

Aker, A., Cohn, T., Gaizauskas, R.. 2010. Multi-document summarization using A* search and discriminative training. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), pages 482-491.

Amini, M. R., Usunier, N., Gallinari, P.. 2005. Automatic text summarization based on word-clusters and ranking algorithms. In Proceedings of the 27th European Conference on Information Retrieval (ECIR 2005), pages 142–156.

Amini, M. R., Usunier, N.. 2009. Incorporating prior knowledge into a transductive ranking algorithm for multi-document summarization. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2009), poster session, pages 704-705.

Banerjee, S. & Pedersen, T.. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In the Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLING-2002). pages 136-145.

Barker, K. & Cornacchia, N.. 2000. Using noun phrase heads to extract document keyphrases. In Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence (AI 2000), pages 40-52.

Barzilay, R. & Elhadad, M.. 1997. Using lexical chains for text summarization. In Proceedings of the ACL 1997 Workshop on Intelligent Scalable Text Summarization, pages 10-17.

Barzilay, R., Mckeown, K., Elhadad, M.. 1999. Information fusion in the context of multi-document summarization. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999), pages 550-557.

Barzilay, R., Elhadad, N., McKeown, K. 2002. Inferring strategies for sentence ordering in multidocument news summarization. In *Journal of Artificial Intelligence Research*, 17(1), pages 35–55.

Barzilay, R., & Lee, L.. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), pages 113-120.

Baxendale, P. B.. 1958. Man-made index for technical literature—An experiment. In *IBM Journal of Research and Development*, 2(4), pages 354-361.

Blei, D. M., Ng, A. Y., Jordan, M., Lafferty, J.. 2003. Latent Dirichlet allocation. In *Journal of Machine Learning Research*, 3, pages 993–1022.

Blei, D. M., Jordan, M., Griffiths, T., Tenenbaum, J.. 2004. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*.

Bollegala, D., Okazaki, N., Ishizuka, M.. 2005. A machine learning approach to sentence ordering for multidocument summarization and it's evaluation. In Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005), pages 624-635.

Bollegala, D., Okazaki, N., Ishizuka M., 2006. A bottom-up approach to sentence ordering for multi-document summarization. In Procedings of the 44th annual

meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 2006), pages 385–392.

Cai, X., Li, W., Ouyang, Y.. 2010. Simultaneous Ranking and Clustering of Sentences: An Reinforcement Approach to Multi-Document Summarization. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pages 134-142.

Carbonell, J. G. & Goldstein, J.. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1998), short paper session, pages 335-336.

Celikyilmaz A., & Hakkani-Tur, D.. 2010. A Hybrid Hierarchical Model for Multi-Document Summarization. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), pages 815-824.

Chuang, W. & Yang, J.. 2000. Extracting sentence segments for text summarization: a machine learning approach. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2000), pages 152-159.

Conroy, J., Schlesinger, J., O'Leary, D.. 2006. Topic Focused Multi-document Summarization Using an Approximate Oracle Score. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006), pages 152-159.

Dang, H. T.. 2005. Overview of DUC 2005. In Proceedings of Document Understanding Conference 2005 (DUC 2005), http://duc.nist.gov.

Daumé III, H. & Marcu, D.. 2006. Bayesian Query-Focused Summarization. In

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL 2006), pages 305-312.

DeJong, G.. 1982. An overview of the FRUMP system. In *Strategies for natural language processing*, pages 149-176.

Dunning, T.. 1993. Accurate Methods for Statistics of Surprise and Coincidence. In *Computational Linguistics*, 19(1), pages 61-74.

D'Avanzo, E., & Magnini, B.. 2005. A Keyphrase-Based Approach to Summarization: the LAKE System at DUC-2005. In Proceedings of Document Understanding Conference 2005 (DUC 2005), http://duc.nist.gov.

Edmundson, H. P.. 1969. New methods in automatic extracting. In *Journal of the Association for Computing Machinery*, 16(2), pages 264–285.

Erkan, G. & Radev, D.. 2004. LexPageRank: Prestige in Multi-Document Text Summarization. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pages 365-371.

Fellbaum, C., editer. 1998. WordNet: an electronic lexical database. In The MIT Press, Cambridge London.

Filatova, E. & Hovy, E.. 2001. Assigning time-stamps to event-clauses. In TASIP'01 Proceedings of the Workshop on Temporal and Spatial Information Processing.

Filatova E. & Hatzivassiloglou, V.. 2004. A formal model for information selection in multisentence text extraction. In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), pages 397–403.

Fisher, S. & Roark, B.. 2006. Query-focused summarization by supervised sentence ranking and skewed word distributions. In Proceedings of Document Understanding Conference 2006 (DUC2006), http:/duc.nist.gov.

Frank, E., Paynter, W., Witten, I., Gutwin, C., Nevill-Manning, G.. 1999. Domain Specific Keyphrase Extraction. In Proceedings of the 16th international joint conference on Artificial intelligence (IJCAI 1999), pages 668-673.

Gillick, D., Favre, B., Hakkani-Tur, D., Bohnet, B., Liu, Y., Xie, S.. 2009. The ICSI/UTD Summarization System at TAC 2009. In Proceedings of Text Analysis Conference 2009 (TAC 2009), http://www.nist.gov/tac/.

Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J.. 1999. 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 (SIGIR 1999), pages 121-128.

Gong Y. & Liu X.. 2001.. 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 (SIGIR 2001), pages 19-25.

Gunn, S. R.. 1998. Support vector machines for classification and regression. Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton.

Gupta, S., Nenkova, A., Jurasky, D.. 2007. Measuring Importance and Query Relevance in Topic-focused Multi-document Summarization. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), poster session, pages 193-196.

Haghighi A. & Vanderwende, L.. 2009. Exploring Content Models for Multi-Document Summarization. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2009), pages

362-370.

Hatzivassiloglou, V., Klavans, J., Holcombe, M., Barzilay, R., Kan, M., McKeown, K. 2001. Simfinder: A flexible clustering tool for summarization. In Proceedings of the NAACL 2001 Workshop on Automatic Summarization.

Hirao, T. & Isozaki, H.. (2002) Extracting important sentences with support vector machines. Proceedings of the 19th International Conference on Computational Linguistics, pages 342-348.

Hovy, E. & Lin, C.. 1998. Automated text summarization and the SUMMARIST system. In ACL workshop on text summarization, pp 197-214.

Israel, L. Q., Han, H., Song, I.. 2010. Focused multi-document summarization: human summarization activity vs. automated systems techniques, Journal of Computing Sciences in Colleges, v.25 n.5, pages10-20.

Jagarlamudi, J., Pingali,P., Varma, V.. 2006. Query Independent Sentence Scoring Approach to DUC 2006. In Proceedings of Document Understanding Conference 2006. http://duc.nist.gov.

Ji, D. & Nie, Y.. 2008. Sentence Ordering based on Cluster Adjacency in Multi-Document Summarization. In Proceedings of the Third International Joint Conference on Natural Language Processing.

Jin, F., Huang, M., Zhu, X.. 2010. A Comparative Study on Ranking and Selection Strategies for Multi-Document Summarization. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pages 525-533.

Jing, H.. 1998. Summary generation through intelligent cutting and pasting of the input document. Technical Report, Columbia University.

Jing, H. & Mckeown, K.. 2000. Cut and paste based text summarization. In

Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000), pages 178-185.

Joachims, T.. 1999. Making large-Scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, MIT-Press.

Joachims, T.. 2002. Optimizing search engines using clickthrough data. In Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2002), pages 133-142.

Jones, Karen S.. 2007. Automatic summarising: The state of the art. In *Information processing & management*, 43(6), pages 1449-1481.

Lund, K. & Burgess, C.. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. In *Behavior Research Methods, Instruments, and Computers*, (28), pages 203-208.

McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., Eskin, E.. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In Proceedings of the 16th national conference on Artificial intelligence and the 11th Innovative applications of artificial intelligence conference innovative applications of artificial intelligence (AAAI/IAAI 1999), pages 453–460.

Katragadda, R. & Varma, V.. 2009. Query-focused summaries or query-biased summaries?. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009), short paper session, pages 105-108.

Kummamuru, K., Lotlikar, R., Roy, S., Singal, K., Krishnapuram, R.. 2004. A hierarchical monothetic document clustering algorithm for summarization and

browsing search results. In Proceedings of the 13th international conference on World Wide Web (WWW 2004), pages 658-665.

Knight, K. & Marcu, D.. 2000. Statistics-based summarization - Step one: Sentence compression. In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000), pages 703-710.

Kupiec, J. M., Pedersen, J., Chen, F.. 1995. A trainable document summarizer. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1995), pages 68-73.

Lacatusu, F., Hickl, A., Aarseth P., Taylor, L.. 2005. MMR variation Lite-GISTexter at DUC 2005. In Proceedings of Document Understanding Conference 2005 (DUC2005), http:/duc.nist.gov.

Lawrie, D., Croft, W., Rosenberg, A.. 2001. Finding Topic Words for Hierarchical Summarization. In Proceedings of the 24th international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2001), pages 349-357.

Leskovec, J., Grobelnik, M., Milic-Frayling, N.. 2004. Learning sub-structures of document semantic graphs for document summarization. In Proceedings of the 7th International Multi-Conference Information Society, pages 133–138.

Li, L., Zhou K., Xue G., Zha H., Yu Y.. 2009. Enhancing diversity, coverage and balance for summarization through structure learning. In Proceedings of the 18th international conference on World Wide Web (WWW 2009), pages 71-80.

Li, W., Li, W., Li, B., Chen, B., Wu, M.. 2005. The Hong Kong Polytechnic University at DUC 2005. In Proceedings of Document Understanding Conference 2005 (DUC2005), http:/duc.nist.gov.

Li, W., Xu, W., Wu, M., Yuan, C., Lu, Q. 2006. Extractive summarization using inter- and intra- event relevance. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-COLING 2006), pages 369-376.

Shen, C. & Li, T.. 2010. Multi-Document Summarization via the Minimum Dominating Set. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pages 984-992.

Lin, C. & Hovy, E.. 2000. The Automatic Acquisition of Topic Signatures for Text Summarization. In Proceedings of The 18th International Conference on Computational Linguistics (COLING 2000), pages 495-501.

Lin, C. & Hovy, E.. 2001. Neats: A multidocument summarizer. In Proceedings of Document Understanding Conference 2001 (DUC2001), http:/duc.nist.gov.

Lin, C. & Hovy, E.. 2002. Manual and automatic evaluation of summaries. In Proceedings of Document Understanding Conference 2002 (DUC2002), http:/duc.nist.gov.

Liu, F. & Liu, Y. 2009. From extractive to abstractive meeting summaries: can it be done by sentence compression?. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009), short paper session, pages 261-264.

Luhn, H.. 1958. The automatic creation of literature abstracts. In *IBM Journal of Research and Development*, 2(2), pages 159-165.

Mani, I. & Bloedorn, E.. 1998. Machine learning of generic and user-focused summarization. In Proceedings of the fifteenth national/tenth conference on

Artificial intelligence/Innovative applications of artificial intelligence (AAAI/IAAI 1998), pages 820-826.

Marcu, D.. 1999. Discourse Trees Are Good Indicators of Importance in Text. In *Advances in Automatic Text Summarization*, pages 123–136. MIT Press, Cambridge.

Martins, A. & Smith, N.. 2009. Summarization with a joint model for sentence extraction and compression. In ACL 2009 workshop on Integer Linear Programming for Natural Language Processing.

McKeown, K. R. & Radev, D. R.. 1995. Generating summaries of multiple news articles. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1995), pages 74-82.

McKeown, K., Robin, J., Kukich, K.. 1999. Generating concise natural language summaries. In *Information Processing & Management*, 31 (5), pages 703-733.

Medelyan, O. & Witten, I. H.. 2006. Thesaurus based automatic keyphrase indexing. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL 2006), pages 296-297.

Medelyan, O. & Witten, I. H.. 2008. Domain independent automatic keyphrase indexing with small training sets. In *Journal of American Society for Information Science and Technology*, 59 (7), pages 1026-1040.

Metzler, D. & Kanungo, T.. 2008. Machine learned sentence selection strategies for query-Biased summarization. In the SIGIR 2008 Workshop on Learning to Rank for Information Retrieval.

Mihalcea, R. & Tarau, P.. 2004. TextRank – bringing order into texts. In the Proceedings of the 2004 Conference on Empirical Methods in Natural

Language Processing (EMNLP 2004), pages 404-411.

Mihalcea, R. & Tarau, P.. 2005. An Algorithm for Language Independent Single and Multiple Document Summarization. In Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005).

Nenkova, A.. 2005. Automatic text summarization of newswire: lessons learned from the document understanding conference. In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005), pages 1436-1441.

Nenkova, A. & Vanderwende, L.. 2005. The Impact of Frequency on Summarization. MSR-TR-2005-101, Microsoft Research Technical Report.

Neto, J., Santos, A., Kaestner, C., Freitas, A.. 2000. Document Clustering and Text Summarization. In Proceedings of the 4th International Conference on Practical Applications of Knowledge Discovery and Data Ming (PADD 2000), pages 41-55.

Neto, J. L., Freitas, A. A., Celso A. A.. 2002. Kaestner. Automatic text summarization using a machine learning approach. In Proceedings of the 16th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence, pages 205-215.

Nomoto, T. & Matsumoto, Y.. 2001. A new approach to unsupervised text summarization. In the 24th international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2001), pages 26-34.

Nomoto, T.. 2007. Discriminative sentence compression with conditional random fields. In *Information Processing & Management*, 43(6), pages 1571-1587.

OtterBacher, J., Erkan, G., Radev, D.. 2005. Using Random Walks for Question-focused Sentence Retrieval. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural

Language Processing (HLT-EMNLP 2005), pages 915-922

Ouyang, Y., Li, S., Li, W.. 2007. Developing learning strategies for topic-based summarization. Proceedings of the 16th ACM conference on Conference on information and knowledge management (CIKM 2007), pages 79-86.

Ouyang, Y., Li, W., Wei, F., Lu, Q.. 2009. Learning Similarity Functions in Graph-Based Document Summarization. In Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy (ICCPOL 2009), pages 189-200

Ouyang, Y., Li, W., Zhang, R., Lu, Q.. 2010a. A Study on Position Information in Document Summarization. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pages 919-927.

Ouyang, Y., Li, W., Zhang, R., Lu, Q.. 2010b. Applying regression models to query-focused multi-document summarization. In *Information Processing & Management*.

Porter, M. F.. 1980. An algorithm for suffix stripping. In *Program*, 14(3), pages 130−137.

Radev, R., Hongyan J., Malgorzata B.. 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In Proceedings of the 2000 NAACL-ACL Workshop on Automatic Summarization, pages 21-30.

Radev, D., Hovy, E., McKeown K.. 2002. Introduction to the special issue on summarization. In *Computational Linguistics*, 28 (4), pages 399-408.

Radev, D., Otterbacher, J., Qi, H., Tam, D.. 2003. MEAD ReDUCs: Michigan at DUC 2003. In Proceedings of Document Understanding Conference 2003

(DUC2003), http:/duc.nist.gov.

Salton, G., Singhal, A., Mitra, M., Buckley, C.. 1997. Automatic text structuring and summarization. In *Information Processing & Management*, 33(2), pages 193-207.

Sanderson, M. & Croft, W. B.. 1999. Deriving concept hierarchies from text. In Proceedings of the 22nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1999), pages 206–213.

Schiffman, B., Nenkova, A., McKeown, K.. 2002. Experiments in Multidocument Summarization. In Proceedings of the 2nd international conference on Human Language Technology Research (HLT 2002), pages 52-58.

Schilder F. & Kondadadi R.. 2008. FastSum: fast and accurate query-based multi-document summarization. In Proceedings of the 46th annual meeting of the Association for Computational Linguistics on Computational Linguistics ACL 2008, short papers session, pages 205-208.

Shen, D., Sun, J., Li, H., Yang, Q., Chen, Z.. 2007. Document summarization using conditional random fields. In Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI 2007), pages 2862-2867.

Svore, K., Vanderwende, L., Burges, C.. 2007. Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), pages 448-457.

Takamura, H. & Okumura, M.. 2009. Text summarization model based on maximum coverage problem and its variant. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL

2006), pages 781–789.

Tombros, A. & Rijsbergen, C. J.. 2004. Query-Sensitive Similarity Measures for Information Retrieval. In *Knowledge and Information Systems*, 6, pages 617-642.

Toutanova, K. et al.. 2007. The PYTHY summarization system: Microsoft research at DUC 2007. In Proceedings of Document Understanding Conference 2007 (DUC2007), http:/duc.nist.gov.

Turney, P.. 1999. Learning to Extract Keyphrases from Text. Technical Report ERB-1057, National Research Council, Institute for Information Technology.

Wan, X., Yang, J., Xiao, J.. 2006. Using Cross-Document Random Walks for Topic-Focused Multi-Document Summarization. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pages1012-1018.

Wan, X., Yang, J., Xiao, J.. 2007. Manifold-Ranking Based Topic-Focused Multi-Document Summarization. In Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI 2007), pages 2903-2908.

Wan, X. & Yang, J.. 2008. Multi-Document Summarization Using Cluster-Based Link Analysis. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008), pages 299-306.

Wan, X. & Xiao, J.. 2008. Single document keyphrase extraction using neighborhood knowledge. In Proceedings of the 23rd national conference on Artificial intelligence (AAAI 2008), pages 885-860.

Wan, X.. 2009. Topic Analysis for Topic-Focused Multi-Document Summarization. In Proceeding of the 18th ACM conference on Information and knowledge

management (CIKM 2009), pages 1609-1612.

Wan, X. & Xiao, J.. 2009. Graph-Based Multi-Modality Learning for Topic-Focused Multi-Document Summarization. In Proceedings of the 21st international joint conference on Artificial intelligence (IJCAI 2009), pp 1586-1591.

Wang, C., Jing, F., Zhang, L., Zhang, H.. 2007. Learning query-biased web page summarization. In Proceedings of the 16th ACM conference on Conference on information and knowledge management (CIKM 2007), pages 552-562.

Wang D., Li, T., Zhu, S., Ding, C.. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008), pages 307-314.

Wang, W., Wei, F., Li, W., Li, S.. 2009a. HyperSum: hypergraph based semi-supervised sentence ranking for query-oriented summarization. In Proceeding of the 18th ACM conference on Information and knowledge management (CIKM 2009), pages 1855-1858.

Wang, D., Zhu, S., Li, T., Gong, Y.. 2009b. Multi-document summarization using sentence-based topic models. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009), short papers session, pages 297-300.

Wei, F., Li, W., Lu, Q., He, Y.. 2008. Query-Sensitive Mutual Reinforcement Chain and Its Application in Query-Oriented Multi-Document Summarization. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008), pages

283-290.

Wei, F., Li, W., He, Y.. 2009. Co-Feedback Ranking for Query-focused Summarization. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009), pages 117-120.

Wiebe, J., O'Hara, T., Ohrstrom-Sandgren, T., McKeever, K. 1998. An empirical approach to temporal reference resolution. In *Journal of Artificial Intelligence*, 9, pages 247–293.

Witbrock, M. & Mittal, V.. 1999. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR 1999), pages 315 - 316.

Xie, X. & Liu, Y.. 2008. Using Corpus and Knowledge-based Similarity Measure in Maximum Marginal Relevance for Meeting Summarization. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), pages 4985 - 4988.

Vapnik, V. N.. 1995. *The nature of statistical learning theory*. Springer.

Ye, S., Chua, T., Kan, M., Qiu, L.. 2007. Document concept lattice for text understanding and summarization. In *Information Processing & Management*, 43(6), pages 1643-1662.

Yih, W., Goodman, J., Vanderwende, L., Suzuki, H.. 2007. Multi-document summarization by maximizing informative content-words. In Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI 2007), pages 1776-1782.

Zajic, D., Dorr, B., Lin, J., Schwartz, R.. 2007. Multi-candidate reduction: Sentence

compression as a tool for document summarization tasks. In *Information Processing & Management*, 43(6), pages 1549-1570.

Zha, H.. 2002. Generic Summarization and Key Phrase Extraction using Mutual Reinforcement Principlae and Sentence Clustering. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), pages113-120

Zhang, Z., Otterbacher, J., Radev, R. D.. 2003. Combining labeled and unlabeled data for learning cross-document structural relationships. In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP 2004), pages 32-41.

Zhao, L., Wu, L., Huang, X.. 2005. Fudan university at DUC 2005. In Proceedings of Document Understanding Conference 2005 (DUC2005), http:/duc.nist.gov.

Zhou, L. & Hovy, E.. 2003. A web-trained extraction summarization system. In Proceedings of Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003), pages 205-211.