Similarity Measures and Diversity Rankings for Query-Focused Sentence Extraction

A Thesis

Submitted to the Faculty

of

Drexel University

by

Palakorn Achananuparp

In partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

May 2010

ACKNOWLEDGEMENTS

This thesis would not have been complete without the assistance from many people over the course of the doctoral study. Especially, I would like to thank my academic advisor, Xiaohua (Tony) Hu, for his guidance, encouragement, and support throughout the years. I have benefited greatly from his impeccable knowledge of data mining and text mining research. I am also grateful to the other committee members, Drs. Lisa Ulmer, Xia Lin, Christopher Yang, Il-Yeol Song, Eileen Abels, and Yuan An, for their constructive comments and feedbacks which greatly help shaping up the quality of this thesis. In particular, I thank Dr. Chris Yang for his helpful comments about the sentence ranking models and his career advice. I am extremely grateful for Dr. Abels for her time and effort in proof reading the initial draft of the thesis and her suggestions on improving its organization. Next, I sincerely thank Dr. Ulmer for her consistent supports, financially and intellectually. Without her, this work would have been much more difficult to accomplish. I thank Dr. Lin for his early guidance which helps refine my research focus later on.

Several other iSchool faculty members also mentored me during my early years of study. I especially thank Drs. Scott Robertson, Hyoil Han, Bob Allen, Katherine McCain, and Mike Atwood, for their supports and advices. I tremendously enjoyed collaborating with them in many research projects. Many core research ideas in this thesis were the products of my collaboration with two colleagues, Xiaohua Zhou (Davis) and Xiaodan Zhang whom I have worked closely with in the last three years. I also thank Lifan Guo for his assistance in several phases of the experiments.

Lastly, I owe the greatest gratitude to my parents, Dr. Surakiat and Mrs. Lalida Achananuparp, for their unconditional love and support throughout my life. One cannot find any better role models in life than both of them. This thesis is dedicated to them.

The research in this thesis was supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), PA Dept of Health Grant (No. 239667), and NSF CCF 0905291 and NSFC 90920005 "Chinese Language Semantic Knowledge Acquisition and Semantic Computational Model Study."

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Similarity Measures and Diversity Rankings
for Query-Focused Sentence Extraction
Palakorn Achananuparp
Supervisor: Xiaohua Hu, Ph.D.

Query-focused sentence extraction generally refers to an extractive approach to select a set of sentences that responds to a specific information need. It is one of the major approaches employed in multi-document summarization, focused summarization, and complex question answering. The major advantage of most extractive methods over the natural language processing (NLP) intensive methods is that they are relatively simple, theoretically sound – drawing upon several supervised and unsupervised learning techniques, and often produce equally strong empirical performance. Many research areas, including information retrieval and text mining, have recently moved toward the extractive query-focused sentence generation as its outputs have great potential to support every day's information seeking activities. Particularly, as more information have been created and stored online, extractive-based summarization systems may quickly utilize several ubiquitous resources, such as Google search results and social medias, to extract summaries to answer users' queries.

This thesis explores how the performance of sentence extraction tasks can be improved to create higher quality outputs. Specifically, two major areas are investigated. First, we examine the issue of natural language variation which affects the similarity judgment of sentences. As sentences are much shorter than documents, they generally contain fewer occurring words. Moreover, the similarity notions of sentences are different than those of documents as they tend to be very

specific in meanings. Thus many document-level similarity measures are likely to perform well at this level. In this work, we address these issues in two application domains. First, we present a hybrid method, utilizing both unsupervised and supervised techniques, to compute the similarity of interrogative sentences for factoid question reuse. Next, we propose a novel structural similarity measure based on sentence semantics for paraphrase identification and textual entailment recognition tasks. The empirical evaluations suggest the effectiveness of the proposed methods in improving the accuracy of sentence similarity judgments.

Furthermore, we examine the effects of the proposed similarity measure in two specific sentence extraction tasks, focused summarization and complex question answering. In conjunction with the proposed similarity measure, we also explore the issues of novelty, redundancy, and diversity in sentence extraction. To that end, we present a novel approach to promote diversity of extracted sets of sentences based on the negative endorsement principle. Negative-signed edges are employed to represent a redundancy relation between sentence nodes in graphs. Then, sentences are reranked according to the long-term negative endorsements from random walk. Additionally, we propose a unified centrality ranking and diversity ranking based on the aforementioned principle. The results from a comprehensive evaluation confirm that the proposed methods perform competitively, compared to many state-of-the-art methods.

CHAPTER 1: INTRODUCTION

"If we knew what it was we were doing,

it would not be called research, would it?"

--German-born American physicist

& Person of The Century

The thesis examines an extractive approach to generate a set of sentences that responds to a specific information need. In particular, two major areas are explored: the similarity measure of sentences and the ranking principle for a set of sentences. In the area of sentence similarity measure, we focus on improving the similarity judgments of generic sentences as well as interrogative sentences. Next, in the area of sentence ranking principle, we investigate several approaches for diversifying a set of sentences in chapter 2. Chapter 3 investigates a task of identifying similar question pairs in the context of question reuse or question retrieval. Chapter 4 focuses on the variability of natural language expression problem which affects the sentence similarity judgments. Chapter 5 examines the effects of sentence similarity measures and diversity ranking methods in different sentence extraction contexts. Lastly, chapter 6 presents a comprehensive evaluation of various state-of-the-art ranking models in promoting diversity of a sets of sentences.

This chapter starts with a motivating example that emphasizes the issues and problems explored in the thesis. Then, it describes the main research questions, the terminology, and the data sets used in the evaluations. Finally, the thesis organization is outlined.

1.1 Motivating Example

Suppose that a young man named Bill has recently become an Apple fan. With a run-away success of the iPad, he could not help but be curious about the history of the company he greatly admired. He wanted to know more about its founders – aside from Steve Jobs, the year and location in which the company first established, and the stories about its wonderful products and inspirations. Being in an internet age, he quickly searched for "history of apple computer inc." on Google. The search results returned to him, shown in figure 1.1, appeared to come from various sources.

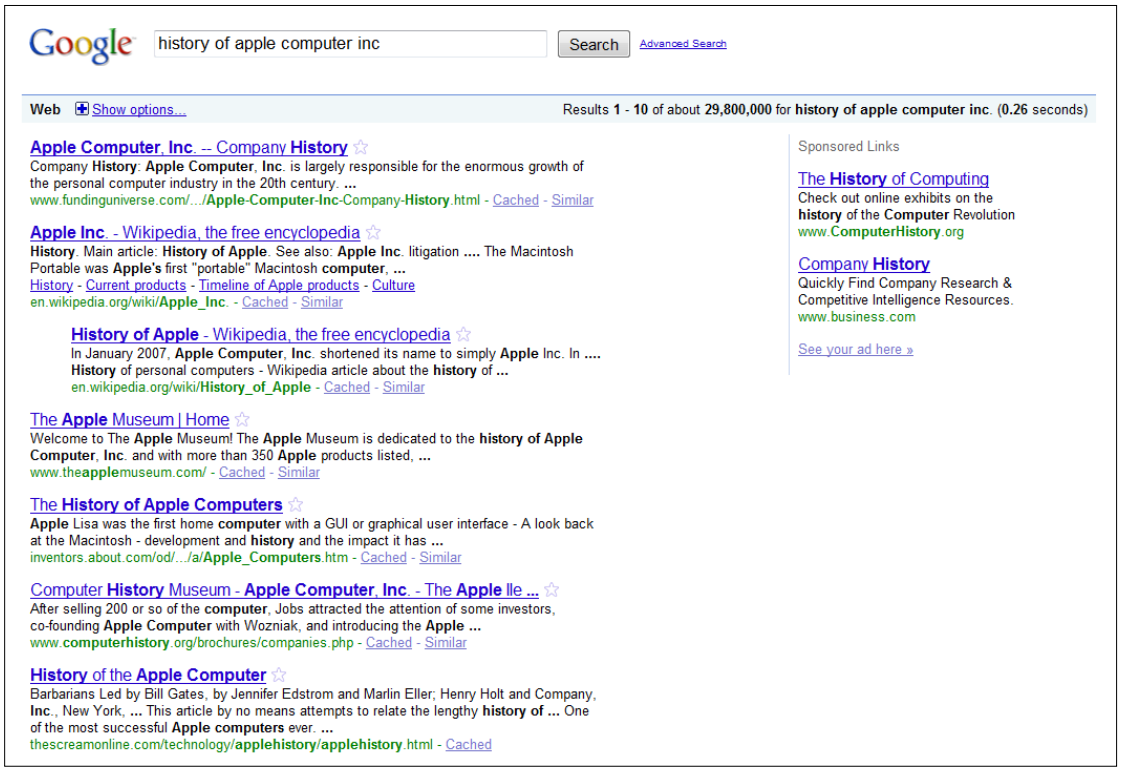


Figure 1.1. Search results for Bill's query "history of apple computer inc."

If we are to create a brief summary for Bills' query by extracting the top five relevant sentences from each top-five retrieved result, a list of candidate sentences will look something like that in figure 1.2.

Apple Computer, Inc. -- Company History
http://www.fundinguniverse.com/company-histories/Apple-Computer-Inc-Company-History.html

1. Apple Computer, Inc. is largely responsible for the enormous growth of the personal computer industry in the 20th century.
2. The introduction of the Macintosh line of personal computers in 1984 established the company as an innovator in industrial design whose products became renowned for their intuitive ease of use.
3. Though battered by bad decision-making during the 1990s, Apple continues to exude the same enviable characteristics in the 21st century that catapulted the company toward fame during the 1980s.
4. The company designs, manufactures, and markets personal computers, software, and peripherals, concentrating on lower-cost, uniquely designed computers such as iMAC and Power Macintosh models.
5. Apple was founded in April 1976 by Steve Wozniak, then 26 years old, and Steve Jobs, 21, both college dropouts.

Apple Inc. - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Apple_Inc.

1. Apple Inc. is an American multinational corporation that designs and manufactures consumer electronics, computer software, and commercial servers.
2. The company's best-known hardware products include Macintosh computers, the iPod, the iPhone and the iPad.
3. Apple software includes the Mac OS X operating system; the iTunes media browser; the iLife suite of multimedia and creativity software; the iWork suite of productivity software; Aperture, a professional photography package; Final Cut Studio, a suite of professional audio and film-industry software products; and Logic Studio, a suite of audio tools.
4. As of January 2010 the company operates 284 retail stores in ten countries, and an online store where hardware and software products are sold.
5. Established in Cupertino, California on April 1, 1976 and incorporated January 3, 1977, the company was called Apple Computer, Inc.

The History of Apple Computers
http://inventors.about.com/od/cstartinventions/a/Apple_Computers.htm

1. On April Fool's Day, 1976, Steve Wozniak and Steve Jobs released the Apple I computer and started Apple Computers.
2. The Apple I was the first with a single circuit board used in a computer.
3. The first home computer with a GUI or graphical user interface was the Apple Lisa.
4. The very first graphical user interface was developed by the Xerox Corporation at their Palo Alto Research Center (PARC) in the 1970s.
5. Steve Jobs, visited PARC in 1979 (after buying Xerox stock) and was impressed and influenced by the Xerox Alto, the first computer ever with a graphical user interface.

Computer History Museum - Apple Computer, Inc.
http://www.computerhistory.org/brochures/companies.php

1. Cupertino, California based high school friends Steve Wozniak and Steve Jobs produced their first computer, the single-board Apple I, in a garage workshop in 1976.
2. After selling 200 or so of the computer, Jobs attracted the attention of some investors, co-founding Apple Computer with Wozniak, and introducing the Apple II computer in 1977.
3. It was the engineering skill of Wozniak (known affectionately as “Woz”) the marketing ability of Jobs, and the hard work of many of the early employees that contributed to Apple’s early success.
4. Apple was the first company to mass market the graphical user interface in their Macintosh computer, introduced in 1984, a product that re-defined personal computing.

History of the Apple Computer
http://thescreamonline.com/technology/applehistory/applehistory.html

1. On April 1, 1976, the Apple computer was born. Steven Wozniak, a high school drop-out who worked for Hewlett-Packard, dabbled in computer-design and created what would become the Apple I.
2. His high school buddy Steven Jobs, also a drop-out, worked for Atari and convinced him that the two should form a company to market the new computer, which eventually took off in 1977 with the Apple II.
3. By 1980, the Apple III was released and their company employed several thousand workers.
4. So begins the rocky, but enormously successful, story of the most revolutionary computer in history.
5. The early Mac's user-friendly interface, with such features as the trash can, windows, drag-and-drop file moveability, and plug-in-and-play compatibility, predated by far the efforts of those developing the PC.

Figure 1.2. Top five relevant sentences from the top five results from Google.

Upon a closer inspection of all candidate sentences, we realize that some of them contain redundant information. For example, a lot of sentences in the retrieved results mention Steve Jobs and Steve Wozniak as Apple's co-founders. Thus for the purpose of creating the most informative summary that has as many key points as possible, we proceed to organize them into six distinct groups and one miscellaneous group, shown in figure 1.3, based on their shared meanings. Group 1 focuses around the first graphical-user interface (GUI) based Macintosh computer. Group 2 gives a general introduction to Apple's hardware products. Group 3 talks about the Apple I computer and the founding of Apple. Group 4 focuses on the Apple II computer and the incorporation of Apple. Group 5 mentions the relationship between Xerox PARC and Apple. And finally, group 6 talks about the company's rough yet successful journey. The other sentences that do not fit into any of these main points are clumped together in the miscellaneous group.

Still, the clustering is far from perfect as there are some partial overlaps between sentences in different groups. For example, the second sentence in group 3 "*Established in Cupertino, California on April 1, 1976 and incorporated January 3,*

1977, the company was called Apple Computer, Inc." also mentions the year the company was incorporated, which is one of the main points in group 4. In addition, some sentences are tangentially related or provide supporting information to the group's meanings. These include the seventh sentence of group 3 "*The Apple I was the first with a single circuit board used in a computer*" and the first sentence of group 5 "*The very first graphical user interface was developed by the Xerox Corporation at their Palo Alto Research Center (PARC) in the 1970s."*

Group 1: The first GUI-based Macintosh computer

1. The introduction of the Macintosh line of personal computers in 1984 established the company as an innovator in industrial design whose products became renowned for their intuitive ease of use.
2. The first home computer with a GUI or graphical user interface was the Apple Lisa.
3. Apple was the first company to mass market the graphical user interface in their Macintosh computer, introduced in 1984, a product that re-defined personal computing.
4. The early Mac's user-friendly interface, with such features as the trash can, windows, drag-and-drop file moveability, and plug-in-and-play compatibility, predated by far the efforts of those developing the PC.

Group 2: The Introduction of Apple's hardware products

1. The company designs, manufactures, and markets personal computers, software, and peripherals, concentrating on lower-cost, uniquely designed computers such as iMAC and Power Macintosh models.
2. Apple Inc. is an American multinational corporation that designs and manufactures consumer electronics, computer software, and commercial servers.
3. The company's best-known hardware products include Macintosh computers, the iPod, the iPhone and the iPad.

Group 3: The Apple I computer and the founding of Apple

1. Apple was founded in April 1976 by Steve Wozniak, then 26 years old, and Steve Jobs, 21, both college dropouts.
2. Established in Cupertino, California on April 1, 1976 and incorporated January 3, 1977, the company was called Apple Computer, Inc.
3. On April Fool's Day, 1976, Steve Wozniak and Steve Jobs released the Apple I computer and started Apple Computers.
4. Cupertino, California based high school friends Steve Wozniak and Steve Jobs produced their first computer, the single-board Apple I, in a garage workshop in 1976.
5. On April 1, 1976, the Apple computer was born.
6. Steven Wozniak, a high school drop-out who worked for Hewlett-Packard, dabbled in computer-design and created what would become the Apple I.
7. The Apple I was the first with a single circuit board used in a computer.

Group 4: The Apple II computer and the incorporation of Apple

1. After selling 200 or so of the computer, Jobs attracted the attention of some investors, co-founding Apple Computer with Wozniak, and introducing the Apple II computer in 1977.
2. His high school buddy Steven Jobs, also a drop-out, worked for Atari and convinced him that the two should form a company to market the new computer, which eventually took off in 1977

with the Apple II.
Group 5: Xerox PARC and Apple 1. The very first graphical user interface was developed by the Xerox Corporation at their Palo Alto Research Center (PARC) in the 1970s. 2. Steve Jobs, visited PARC in 1979 (after buying Xerox stock) and was impressed and influenced by the Xerox Alto, the first computer ever with a graphical user interface.
Group 6: Apple's rough yet successful journey 1. Though battered by bad decision-making during the 1990s, Apple continues to exude the same enviable characteristics in the 21st century that catapulted the company toward fame during the 1980s. 2. So begins the rocky, but enormously successful, story of the most revolutionary computer in history.
Miscellaneous Group 1. Apple Computer, Inc. is largely responsible for the enormous growth of the personal computer industry in the 20th century. 2. Apple software includes the Mac OS X operating system; the iTunes media browser; the iLife suite of multimedia and creativity software; the iWork suite of productivity software; Aperture, a professional photography package; Final Cut Studio, a suite of professional audio and film-industry software products; and Logic Studio, a suite of audio tools. 3. As of January 2010 the company operates 284 retail stores in ten countries, and an online store where hardware and software products are sold. 4. It was the engineering skill of Wozniak (known affectionately as "Woz") the marketing ability of Jobs, and the hard work of many of the early employees that contributed to Apple's early success. 5. By 1980, the Apple III was released and their company employed several thousand workers.

Figure 1.3. The candidate sentences organized into seven groups.

The challenges of the extractive summarization task is beginning to show. First, the sentences in each group demonstrate that the same key point can be linguistically formulated in various ways. For instance, all four sentences in group 1 convey the same key point about the first Macintosh computer using variations of words and syntactic compositions. Moreover, not all of them are completely equivalent to one another. Apart from the first and the third sentences, some sentences contain extra information that are not expressed in the others. The first sentence "*The introduction of the Macintosh line of personal computers in 1984 established the company as an innovator in industrial design whose products became renowned for their intuitive ease of use*" seems to be the only one that is semantically equivalent to the third sentence "*Apple was the first company to mass market the*

graphical user interface in their Macintosh computer, introduced in 1984, a product that re-defined personal computing." The *variability of natural language* is one of the challenges in text summarization as well as other high level applications, such as question answering, information extraction, and machine translation.

The second challenge is diversity of information or facts in the summary. Without knowing a priori the aspects of Bill's information needs, it becomes unclear as to how the representative sentences should be selected. This particular problem is regarded as extrinsic diversity. Figure 1.4 shows one possible way to extract an extrinsically diverse set of sentences which captures all of Bill's information needs -- Apple's cofounders, the year and location in which the company first established, and the stories about its products and inspirations. Alternatively, without assuming what aspects of Apple's history Bill was seeking, we still have to take into account an intrinsic diversity. That is, we need to make sure that the representative sentences are novel or factually distinct, compared to others. An ideal intrinsically diverse summary, as shown in figure 1.5, can be extracted by selecting one best representative sentence, the one which contains the most factual coverage, from each group.

1. Cupertino, California based high school friends Steve Wozniak and Steve Jobs produced their first computer, the single-board Apple I, in a garage workshop in 1976.
2. After selling 200 or so of the computer, Jobs attracted the attention of some investors, co-founding Apple Computer with Wozniak, and introducing the Apple II computer in 1977.
3. Steve Jobs, visited PARC in 1979 (after buying Xerox stock) and was impressed and influenced by the Xerox Alto, the first computer ever with a graphical user interface.
4. It was the engineering skill of Wozniak (known affectionately as "Woz") the marketing ability of Jobs, and the hard work of many of the early employees that contributed to Apple's early success.
5. Though battered by bad decision-making during the 1990s, Apple continues to exude the same enviable characteristics in the 21st century that catapulted the company toward fame during the 1980s.
6. As of January 2010 the company operates 284 retail stores in ten countries, and an online store where hardware and software products are sold.

Figure 1.4. An ideal extrinsically diverse summary.

1. The introduction of the Macintosh line of personal computers in 1984 established the company as an innovator in industrial design whose products became renowned for their intuitive ease of use.
2. The company's best-known hardware products include Macintosh computers, the iPod, the iPhone and the iPad.
3. Apple software includes the Mac OS X operating system; the iTunes media browser; the iLife suite of multimedia and creativity software; the iWork suite of productivity software; Aperture, a professional photography package; Final Cut Studio, a suite of professional audio and film-industry software products; and Logic Studio, a suite of audio tools.
4. Apple was founded in April 1976 by Steve Wozniak, then 26 years old, and Steve Jobs, 21, both college dropouts.
5. Steve Jobs, visited PARC in 1979 (after buying Xerox stock) and was impressed and influenced by the Xerox Alto, the first computer ever with a graphical user interface.
6. By 1980, the Apple III was released and their company employed several thousand workers.
7. Though battered by bad decision-making during the 1990s, Apple continues to exude the same enviable characteristics in the 21st century that catapulted the company toward fame during the 1980s.

Figure 1.5. An ideal intrinsically diverse summary.

1.2 Research Questions

The research presented in this thesis is motivated by the previous example. Because of the challenges in extracting a set of sentences that responds to a specific information need, the thesis focuses on answering the following research questions:

1.2.1 Research question 1 (RQ1)

What are the useful resources that helps improve the similarity judgment at sentence level? How can we incorporate them into the similarity function?

Since sentences are much shorter than documents, it contains less contextual information, e.g. word occurrences. In terms of semantics, sentences contain more specific expressions than documents. In addition, because of the natural language variability, sentences with the same meaning can be linguistically reformulated in various forms. This makes it much harder for most text similarity measures to make an accurate judgment. Therefore, in order to improve the similarity judgment, we want to find useful resources which can be incorporated into the sentence similarity function. Resources are broadly defined as any component which can be utilized. Specifically, they can be either lexical knowledge, e.g. dictionary, thesaurus, and/or tools. We focus on two broad classes of sentences, interrogative sentence (or question) and generic sentence (any syntactically formed text fragment). First, we explore the issue of comparing the similar interrogative sentences from the task of finding similar questions in question-answering archives. In particular, what are the components that can be integrated into the similarity function to identify semantically similar questions? For generic sentences, we are interested in answering the following questions: what semantic knowledge can be integrated into the sentence similarity measures? What are their effects on the similarity judgment of generic sentences? How effective are they in dealing with different similarity notions, compared to the existing methods?

1.2.2 Research question 2 (RQ2)

What is the effectiveness of the proposed similarity measure in different application contexts? How can we incorporate the proposed similarity method into sentence extraction methods?

Sentence similarity measures play a crucial role in many text mining applications. Specifically, extractive-based applications, such as text summarization and question answering, employ several similarity functions as part of the sentence extraction process. Most similarity functions compute the similarity scores based on co-occurrences or distributional similarity of words between two sentences. These functions include the Jaccard coefficient, cosine similarity, etc. We are interested in the effectiveness of these methods in the specific application contexts. Moreover, we examine whether the measures that perform well in sentence similarity evaluations improve the overall performance of the sentence extraction task. Particularly, does the proposed similarity measure improve the effectiveness of focused summarization and question answering?

1.2.3 Research question 3 (RQ3)

How can we apply a graph-based ranking model to intrinsically promote diversity of a set of sentences?

In the previous research questions, we examine how to identify semantically similar sentences in the context of focused summarization and question answering. The similarity measure allows us to find the representative sentences, those that uniquely describe a fact or information. Another related issue in sentence extraction is the diversity of the extracted set. Specifically, this work focuses on intrinsic diversity. While the similarity judgment focuses on the comparison between a pair

of sentences, the diversity focuses on the extracted set of sentences collectively. That is, each sentence in a diverse set should contain novel information with respect to others. There has been a significant amount of research regarding novelty, redundancy, and diversity in ranking (Zhu et al. 2007; Clarke et al. 2008; Li et al. 2009). However, very few methods have considered a graphical model to promote diversity. The previous works have demonstrated the effectiveness of the graphical models, such as random walks, in finding salient items from a sentence graph. Drawing upon research in the graphical models and diversity in ranking, we focus on answering the following questions: How can we incorporate novelty, the opposite of redundancy, into a sentence graph? How effective is the proposed graphical representation, compared to the traditional graph-based models? How effective is the proposed graphical model in focused summarization and question answering?

1.2.4 Research question 4 (RQ4)

What is the best way to incorporate diversity ranking into the graph-based ranking model while retaining the advantage of centrality ranking? How effective is the proposed diversity ranking model, compared to the similar state-of-the-art methods?

Many diversity promotion or redundancy reduction methods are typically applied to a set of items post-ranking. In some cases, diversity is considered as an implicit property of the ranking principle. As such, the traditional ranking methods treat saliency and diversity separately. Following the diversity in ranking issue, we investigate the performance of the graphical models which incorporate centrality and diversity in one unified process. To that end, we will conduct a comprehensive evaluation on the standard focused summarization and question answering tasks. In particular, we focus on answering the following questions: What are the performance

improvements, if any, of the unified centrality and diversity ranking models, compared to the models that consider diversity implicitly? What is the effectiveness of different diversity ranking principles in extracting a diverse set of sentences? What performance metrics should be used to evaluate the diversity of the sets of sentences? What are the agreements among different evaluation metrics?

1.3 Terminology

The thesis focuses on *extractive* approach, as opposed to *abstractive* approach, to generate a set of sentences. Abstractive approach refers to a new reproduction of content, while extractive approach generates the set of sentences by extracting or selecting sentences from the original sources.

Sentence extraction generally refers to an extractive approach of generating a set of sentences. *Query-focused sentence extraction* or *query-focused extraction* refers to a more specific sentence extraction task where a set of sentences is extracted to respond to a given information need. Since most methods in the thesis are proposed for query-focused sentence extraction task, we use *sentence extraction* as a shorter form of *query-focused sentence extraction* unless specified otherwise.

Sentence and *document* are two different levels of text units discussed throughout the thesis. A sentence is a syntactically formed sequence of words. It expresses a specific *fact* or *piece of information*. A document consists of one or more sentence. In the context of sentence graphs, we use *sentence*, *item*, *node*, *vertex*, and *state* interchangeably.

We use *similarity* generally to refer to various relations between text units, e.g. *relatedness*, *paraphrase*, *entailment*, and *topicality*. Methods discussed

throughout the thesis typically refer to the similarity at specific level of text units, either *word-level similarity*, *sentence-level similarity*, or *document-level similarity*.

In the context of sentence extraction tasks, we use *focused summarization*, *topic-focused summarization*, *query-focused summarization*, *goal-focused summarization*, and *task-focused summarization* interchangeably. This is the task of extracting a summary for a given information need. Similarly, *question answering* and *complex question answering* are used interchangeably, except in chapter 3 where question answering specifically refers to *factoid question answering*.

Lastly, *novelty* is used to represent the opposite of *redundancy* and a unit of *diversity*. See chapters 5 and 6 for more description of these terms).

1.4 The Overview of the Evaluation Data Sets

Six standard data sets were used in several experimental evaluations in the thesis. They are briefly summarized as follows.

1.4.1 TREC-9 Question Variants Key

This data set was employed in chapter 3's experimental evaluation. TREC (Text REtrieval Conference) is an annual conference , organized by the National Institute of Standards and Technology (NIST), to encourage research in information retrieval. It consists of several research tracks depending on various information retrieval tasks. For each track, NIST provides a set of test document collections and questions. The test data used in chapter 3 are taken from the question answering track of the ninth Text Retrieval Conferences (TREC-9). We selected a set of 193 question pairs from TREC-9 question variants key. The variants key consists of fifty four original questions and their variants. The original questions are a subset of test questions used in TREC-9 QA experiment and were taken from the actual users'

submissions. The question variants are the paraphrased questions that were constructed by human assessors to be semantically identical but syntactically different from the original questions. 386 question pairs are used as a test set -- 50% of which are the positive pairs.

1.4.2 Microsoft Research Paraphrase Corpus (MSRP)

MSRP data set was employed in chapter 4's evaluation section. It contains 5,801 pairs of paraphrased sentences (4,076 training pairs and 1,725 test pairs) which have been automatically extracted from various new sources on the web by Microsoft Research. Each sentence pair is judged by two human assessors whether they are semantically equivalent or not. In other words, a bi-directional semantic inference is required to judge paraphrase pairs. Positive examples comprise 67% of the total sentence pairs. Semantically equivalent sentences may contain either identical information or the same information with minor differences in detail according to the principal agents and the associated actions in the sentences. In contrast, non-paraphrased sentences may contain several word overlaps, but they are judged to be not equivalent if they do not the same key information, i.e. principal agents and actions. In addition, sentence that describes the same event but is a superset of the other is considered to be a dissimilar pair. Note that the latter rule is similar to the one used in text entailment task.

1.4.3 The Third PASCAL Recognizing Textual Entailment Challenge (RTE3) Data Set

RTE3 data set was employed in chapter 4's experimental evaluation. It consists of 800 pairs of entailment sentences from the development set and 800 pairs of entailment sentences from the test set used in the third Recognizing Textual

Entailment Challenge (RTE). Each pair comprises two small text segments, which are referred to as *text* and *hypothesis*. The text-hypothesis pairs are collected by human assessors from four subsets of application domains: information retrieval, multi-document summarization, question answering, and information extraction. Similarity judgment between sentence pairs is based on directional inference between text and hypothesis. If the hypothesis can be entailed by the text, then that pair is considered to be a positive example. On the other hand, a negative example indicates that the hypothesis cannot be inferred from the text.

1.4.4 Document Understanding Conferences 2006 & 2007 Data (DUC06 &DUC07)

DUC06 and DUC07 were employed in the evaluations in chapter 5 and 6. The two data sets were taken from the standard data sets used in the 2006 and 2007 Document Understanding Conferences (DUC). The tasks and test collections in DUC data sets are prepared by human experts at NIST to be used for evaluating document summarization systems. Each data set comprises a set of topics (50 topics for DUC06 and 45 topics for DUC07), a set of 25 relevant news articles, and a set of human-extracted summaries for each topic to be used as the reference. Each topic contains title and a brief narrative. The main task is to generate a 250-word summary corresponding to each summary topic description.

1.4.5 Complex Interactive Question Answering 2006 (CIQA06)

CIQA06 was employed in the evaluations in chapter 5 and 6. The data set was taken from the ciQA (complex, interactive question answering) task at TREC 2006 Question Answering track. Information needs in ciQA (referred to as topics) contain a canonical form of questions called template, and a free-form narrative describing

the specific aspects of users' information needs. In contrast to other forms of question answering task, e.g. factoid question answering where the answers to those questions are typically 50 characters or fewer, ciQA's information needs reflect those posed by intelligence analysts and require a paragraph-long answer.

1.4.6 The Subset of Yahoo! Answers Data (YahooQA)

YahooQA was employed in chapter 5's experimental evaluation. The data set contains subjective and ill-defined information needs formulated by the members of Yahoo! Answers community. The subjects of interests span widely from mathematics, general health, to wrestling. In this work, 100 questions and 10,546 answers were randomly selected from the top 20 most frequent question categories, measured in terms of responded answers, to use as a test set. The test set was re-formatted in order to make it consistent with the standard procedure used in ciQA tasks. To achieve that, a list of benchmark information nuggets for YahooQA is automatically created by matching the relevant answers with the corresponding questions. The best answer, chosen by asker, for each question is marked as vital nugget while the other answers are marked as okay nugget.

1.5 Thesis Organization

The thesis is presented in seven chapters. In chapter 2, we review the related works regarding various aspects of similarity of sentences, i.e. the notions of similarity, the similarity measures, and techniques used in ranking sentences. Then, we investigate the effectiveness of similarity functions in finding question paraphrases in chapter 3. Specifically, we propose a hybrid question similarity measure which incorporates semantic, syntactic, and categorical information of questions and evaluate it on question paraphrases identification task. Next, in chapter 4, we

further examine the performance of sentence similarity measures on two specific similarity notions, semantic equivalence and textual entailment. We focus on the issue of variability of natural language expression and its effects on the similarity judgments. To address the issue, we propose a method that incorporates the semantic structure of sentences and evaluate its performance on paraphrase identification and textual entailment recognition tasks. In chapter 5, we investigate the effectiveness of the semantic similarity measure, proposed in the previous chapter, in a context of query-focused sentence extraction. In addition to the proposed measure, we introduce a graph-based ranking model that focuses on reranking the extracted sentences based on their novelty. Specifically, the proposed ranking model is based on random walk over the negative-edge graph. We evaluate the effects of the methods on focused summarization and question answering tasks. Next, the issue of diversity in ranking is examined in chapter 6. As a follow-up to the findings from previous chapter, we propose a unified model of centrality and diversity ranking by extending the negative-edge based model. A comprehensive evaluation is conducted on focused summarization and question answering using several diversity-focused metrics. Lastly, we conclude the thesis by summarizing our contributions as well as discussing the implication for future works.

CHAPTER 2: LITERATURE REVIEW

"Research is to see what everybody else has seen,

and to think what nobody else has thought."

--Hungarian spy & discoverer of vitamin C

This chapter reviews the related works in two major areas. The first area focuses around the similarity of sentences. The discussions include the notions of similarity and measures employed to compute the similarity scores at the sentence level as well as the word level. The second area focuses around sentence selection and ranking methods applied to generic sentence extraction as well as focused sentence extraction.

2.1 Judgment of Text Similarity at Sentence Level

2.1.1 Notions of Sentence Similarity

In recent years, different notions of similarity between sentences have been proposed in various domains. For instance, in information retrieval (IR), Metzler et al. (2005) and Murdock (2006) proposed the spectrum of relevance and similarity in sentence retrieval. They suggested the more fine-grain notions of relevance and similarity based on multiple levels of specificity. The spectrum of relevance ranges from useful content, tangentially related, on the general topic, on the sub-topic, providing supporting information to satisfying the request directly. Similarly, the spectrum of similarity is broken down into exact match, matching at the synonym level, matching at the related term level, matching at the co-occurrence level, and unrelated. Apart from lexical similarity, sentences can be structurally similar. The

spectrum of structural similarity consists of identical construction, clauses reordered, matching n-grams, matching pattern, and unrelated.

In the natural language processing (NLP) community, two specific notions of sentence similarity are explored, semantic equivalence (Dolan et al. 2004) and textual entailment (Dagan et al. 2005). These two relationships are strongly intertwined, and therefore, a clear judgment is sometimes difficult to make. In general, sentences are considered to be semantically equivalent if they express the same meaning. That is, sentences are considered to be semantically equivalence if we can made a bidirectional inference between the them. Following this definition, a paraphrase is considered the most common form of semantic equivalent sentence.

In contrast, textual entailment can only be inferred directionally. For example, given two sentences “John bought a Toyota” and “John bought a car”, we can sufficiently infer the meaning of the latter from the former, but not the other way around. That is, if John bought a Toyota, we can safely conclude that he bought a car. On the other hand, if John bought a car, we don’t have sufficient information to conclude that he bought a Toyota. Many computational methods have been proposed to address the issue of semantic understanding of text units, several of them rely on word or n-gram distribution which is arguably insufficient to distinguish various notions of sentence similarity and semantic redundancy. For example, “John bought a car from Mike” and “Mike bought a car from John,” share virtually the same word occurrences. Thus, the word distribution approaches are likely to judge them as identical. However, from a semantic point of view, we can clearly see that they describe two different events having different subjects and objects.

In conclusion, it is important to distinguish between different notions of sentence similarity as they particularly pertain to sentence extraction tasks in many application domains, such as text summarization, question answering, etc. From sentence retrieval perspective, the degree to which sentences are similar is determined by topical specificity level. Next, Natural language processing research focuses on judgments at semantic level. Sentences are said to be semantically equivalent if bidirectional inference between them can be made. On the other hand, if an inference can be made in one direction only, we conclude that one sentence entails the other.

2.1.2 Notions of Word Similarity

Many approaches to compute the similarity of sentences rely on the similarity at word level. In general, similarity describes the quality between two objects or concepts which share common attributes. In computational linguistics, there are three terms that often used interchangeably to denote similarity of words: semantic similarity, semantic relatedness, and semantic distance (Budanitsky & Hirst, 2006). In many studies, semantic similarity usually refers to the notion of synonymy between words (Rubenstein & Goodenough 1965; Miller & Charles 1991; Landuaer & Dumais 1997) or sometimes uses to represent is-a relations (hypernym/hyonym) (Ponzetto & Strube 2007). For example, “automobile” and “car” are more semantically similar than “automobile” and “gas” in this notion. On the other hand, Budanitsky & Hirst (2006) and Resnick (1995) adopted a broader similarity judgment by distinguishing between the notion of semantic similarity and semantic relatedness. That is, semantic relatedness covers a wider range of word relations than semantic similarity by including other notions, such as meronymy (part-of

relation), antonymy (opposite relation), and other types of functional relations (is-made-of, is-an-attribute-of, etc.) (Ponzetto & Strube 2007). In other words, semantic similarity contains a subset of notions used in semantic relatedness. Finally, semantic distance typically refers to the opposite notion of similarity or relatedness. If two words are closely similar or related, they can be said to be highly distant. Given this example, the words "car" and "automobile" is more semantically similar to each other than "car" and "noodle". In other words, the semantic distance between "car" and "noodle" is greater than the semantic distance between "car" and "automobile." Nevertheless, sometimes semantic distance is also used to refer to the same notion as similarity and relatedness. The lack of consensus on the usage of these terms in the literatures often causes confusion to the readers.

2.1.3 Sentence Similarity Measures

Various techniques have been proposed to measure the similarity scores between pairs of sentences. First, in sentence retrieval application, probabilistic approaches have been adopted to identify topically similar sentences. One of the main issues of measuring sentence similarity is vocabulary mismatch problem. To address the problem, the sentence similarity task has been modeled as a statistical translation in a monolingual setting (Berger and Lafferty 1999). For example, Metzler et al. (2005) proposed a generalized framework of sentence similarity based on statistical translation models which can be parameterized to measure sentence similarity at different levels of relevance. Given an alignment of corresponding words between the query sentence and target sentence, and a distribution of term translation probabilities, the probability of translation is computed as a product of the translation probabilities of the aligned words. Murdock (2006) and Croft (2005) also

proposed a model of sentence similarity (referred to as Model-S) based on statistical translation techniques applicable to various sentence retrieval tasks. Specifically, the authors presented methods for smoothing Model-S and conditional models. In their subsequent work, Metzler et al. (2007) further considered different types of text representations, such as surface form, stemmed form, and expanded form, and applied them to various similarity measures for query-level judgment. They employed the negative KL-divergence as a ranking mechanism in their probabilistic framework. Recently, Balasubramanian et al. (2007) compared the performance of nine language modeling techniques in sentence retrieval task. They found that, despite their superiority in coping with the vocabulary mismatch problem, most probabilistic methods do not significantly outperform existing measures in sentence retrieval task.

Next, several approaches have been proposed to identify paraphrases. Previous works in paraphrase recognition focus on sentence alignment task in monolingual comparable corpus. For instance, Barzilay and Elhadad (2003) demonstrated that using a weak sentence similarity measure, such as cosine similarity, with contextual information is more effective than using sophisticated sentence similarity measures (Hatzivassiloglou et al. 1999; Jing 2002). Dolan et al. (2004) investigated two unsupervised techniques, edit distance and heuristic strategy, to find monolingual sentence-level paraphrases from multiple news sources over the web. They found that sentential paraphrase data extracted by edit distance is cleaner and easier to align than the heuristic data. Nevertheless, edit distance data lacks many lexical and syntactic variations. Next, lexical knowledge bases, such as WordNet, have been utilized in several unsupervised approaches. Mihalcea et al. (2006) suggested a hybrid corpus-based and knowledge-based method for measuring the semantic

similarity of sentences. To achieve that, they combined word specificity, as specified by inverse document frequency, and word-to-word semantic similarity, derived from WordNet-based semantic similarity measure. The paraphrase recognition experiment has shown that their method significantly outperformed many traditional lexical matching methods. Malik et al. (2007) adopted the similar measure for mapping new questions to existing questions in the automatic email response system.

Recently, the natural language processing community has focused on developing systems for recognizing textual entailment and paraphrase (Dagan et al. 2005; Giampiccolo et al. 2007). Various techniques, with varying degree of complexity, have been utilized in multiple system components, including WordNet, n-gram word similarity, syntactic matching, semantic role labeling, logical inference, corpus-based statistics, machine learning classification, anaphora resolution, and entailment-corpora background knowledge. For this task, systems that extensively employ deep natural language components and extensive background knowledge have shown significant improvement over relatively shallower approaches. For instance, the best overall system in RTE3 by Hickl and Bensley (2007) employed a background knowledge from a large corpora of entailment examples to train the classifier. The large training examples crucially contributed to their 80% accuracy. The second best system, scored at 72% accuracy, (Tatu and Moldovan 2007) utilized a sophisticated named entities analysis, especially of person names, as well as the extended WordNet knowledge base parsed from WordNet's glosses. Nevertheless, this comes with a trade off in computation cost and training time which render NLP-intensive systems currently impractical for a large text collection.

2.1.4 Word Similarity Measures

In many text application domains, it is necessary to quantify word similarity into computable values. There are several methods to compute word similarity. They can be grouped into various approaches according to specific criteria. For instance, from the source of semantic knowledge perspective, several methods employ external knowledge bases as the sources for semantic relations of words or concepts. The knowledge bases contain explicit relations of words or concepts with varying degree of formality. The ontological knowledge base organizes concept relations into hierarchies. In the literature, two most commonly used ontologies include WordNet (Fellbaum 1998) and MeSH (Medical Subject Headings). The less formal forms of knowledge bases are dictionary, thesauri, and encyclopedic entries. These include WordNet, Roget's thesaurus (McHale, 1998), and Wikipedia. Alternatively, instead of relying on external knowledge sources, semantic relations of words can be derived from corpus statistics. From the methods of computing word similarity perspective, word similarity measures can be categorized into two main approaches: path-based approach and word distribution approach.

The path-based approach, sometimes referred to as a knowledge-based or taxonomy-based approach, computes the word similarity scores from a taxonomical distance between their concepts in the concept hierarchy. Each word is represented as a concept node while its relations to others represented as links connecting to the other concept nodes. The similarity between two concept nodes is computed by counting a number of edges or vertices that form a specific path between them. The shorter the path length, the more similar the two concepts are. However, one major assumption of a path-based approach is the notion of the uniform distance of links in the hierarchy (Budanitsky & Hirst, 2006), which does not always hold in many

knowledge bases, e.g. WordNet. To solve this problem, many methods have included a scaling factor with respect to the depth of concepts in the hierarchy. Wu & Palmer (1994) proposed a conceptual similarity measure that computes the similarity of two concepts in WordNet hierarchy as a proportion of the depth of their least common subsumer (LCS) and the depth of the given concepts. Leacock and Chodorow (1998) used the maximum depth of the hierarchy as a scaling factor. Resnick (1995) introduced the measure that determines the similarity of two concepts by the information content of their least common subsume. Generally, the information content of a concept is obtained from corpus statistics. For instance, in Resnick's experiment, the information content of concepts were calculated from Brown Corpus data (Francis & Kucera 1982). A major shortcoming of Resnick's measure is that it only uses concept relations to find the LCS of a concept pair. However, if two concept pairs happen to share the same LCS, then it is unable to provide a finer granularity of the similarity between those two pairs. Two subsequent measures were proposed as a follow up to Resnick's algorithm. First, Jiang & Conrath (1997) put more emphasize of taxonomic relations in the concept hierarchy into Resnick's measure. Next, Lin (1998) proposed a universal similarity measure that is applicable to any objects or any forms of knowledge representation. Essentially, Lin's measure is Resnick's measure normalized by information content of the two given concepts.

All path-based measures discussed so far in this section consider synonymy and hypernymy (*is–a* relations) to compute word similarity. In contrast, Hirst and St-Onge (1998) proposed a measure that considers substantial types of word relations, including upward relations, i.e. hypernymy and meronymy, downward relations, i.e. hyponymy and holonymy, and horizontal relation, i.e. antonymy. Recently, Seco et al. (2004) proposed an intrinsic information content measure

employing the structure of the concept hierarchy. Unlike Resnick's formulation, the information content of a concept used in Seco's algorithm is defined as a function of its child nodes. Ponzetto & Strube (2007) proposed a novel method to compute word similarity based on Wikipedia's category hierarchy. The categories organize Wikiepdia articles into a hierarchical structure based on different classification schemes, e.g. topics, lists, projects, etc. Given the two words w_1 and w_2, they retrieved disambiguated Wikipedia articles p_1 and p_2, corresponding to each word. Then, they extracted the lists of Wikipedia categories C_1 and C_2 for p_1 and p_2, respectively. For each category pair c_1 and c_2, $c_1 \in C_1, c_2 \in C_2$, they extracted all possible paths connecting c_1 and c_2. Once a set of paths were extracted, several path-based measures could be applied to compute the similarity of c_1 and c_2. for example, they adapted Resnick' measure with intrinsic information content (Seco et al. 2004).

The next approach computes the similarity scores based on word distributions. First, the original word overlap measure, gloss overlap, was introduced by Lesk (1986) as a word sense disambiguation technique. To distinguish between different senses of words, gloss overlap measure compares different glosses (dictionary definition) of the target word with those of the other words. The sense of the target word which contains the most word overlap with the surrounding words' is then selected. Banerjee and Pederson (2003) proposed the extended gloss overlap measure by extending the original Lesk's algorithm to include the related concepts from the WordNet hierarchy. To achieve that, they considered the following word relations: hypernymy, hyponymy, meronymy, holonymy, troponymy, attribute, similar-to, and also-see. The glosses of input concept pairs were exhaustively compared with glosses of related concepts. The maximum score was selected from all

possible concept pairs. Ponzetto & Strube (2007) proposed a novel method to compute word similarity from Wikipedia articles. Instead of comparing WordNet dictionary definitions or *glosses*, like in Lesk's algorithm, they retrieved the first paragraph of the Wikipedia article corresponding to each word being compared. The similarity between two words was defined as a double-normalized overlap score in order to minimize the role of outliers. The first normalization was by the sum of text lengths while the second normalization was in the form of hyperbolic tangent function.

2.2 Sentence Selection and Ranking for Generic and Focused Extractions

Many sentence extraction methods rely on a ranking mechanism to quantify the importance or saliency of each candidate sentence. In generic multi-document summarization, Nenkova and Vanderwende (2005) first proposed a corpus-level frequency sentence scoring method called *SumBasic*. Their method is based on the observation that human-constructed summaries tend to contain highly frequent words. To compute SumBasic, each sentence is scored by the sum of the average probability of the words in the sentence. After the top sentence is chosen, the probability of words containing the selected sentence is updated to penalize redundancy. They empirically proved that word frequency significantly contributes to the extraction of salient sentences using the standard benchmark data sets. A few subsequent works tried to extend SumBasic into several directions. For example, Yih et al. (2007) employed sentence position in addition to the word frequency feature. In topic-focused summarization, Vanderwende et al. (2007) proposed SumFocus which computes a sentence score as a linear combination of the unigram

probabilities derived from the topic description and the unigram probabilities from the document.

Erkan and Radev (2004) proposed *LexRank*, an eigenvector centrality approach to find salient sentences for multi-document summarization. Their method is inspired by a well-known PageRank algorithm (Brin and Page 1998). In essence, LexRank defines a random walk over sentence graph where each vertex represents the individual sentence and each edge represents the similarity between sentences. The edge weight is determined by TFIDF-weighted cosine similarity score between sentence nodes. Since the sentence graph can be transformed into a stochastic matrix, it defines a Markov chain. Thus, each sentence can be ranked according to its stationary distribution. Important sentences are selected according to the highest stationary distribution. The LexRank method has been extended to several topic-focused sentence extraction tasks, such as question answering (Otterbacher et al. 2005) and focused summarization (Otterbacher et al. 2009). The extended approach, called topic-focused LexRank, is defined as a mixture model of the relevance of the sentence to the query and the similarity between sentences. Mihalcea and Tarau (2004) also incidentally proposed a similar eigenvector centrality approach for single-document summarization called TextRank. Their main idea is the same as LexRank in which a sentence graph is constructed and transformed into stochastic matrix. Then, sentences are selected according to their stationary distribution.

Some recent works have applied information distance to extractive summarization. Information distance is based on Kolmogorov complexity (Li and Vitanyi 1997) which is comparable to a well-known information theory developed by Claude Shannon. For instance, Long et al. (2009) proposed a conditional information distance based approach for extractive multi-document summarization. Two

methods used for estimating information distance were presented in their approach: approximation by compression and approximation by the coding theory. Topic models (Blei et al. 2003) have also been explored in the context of focused summarization. For example, Tang et al. (2009) focused on the problem of multi-topic based focused summarization. To address the problem, they proposed a statistical topic model to discover multiple topics in a document collection. Two strategies for incorporating the query information into the topic model were explored. The first strategy integrated the query information into the generative process of the topic model, resulting in a mixture of a document-specific topic distribution and a query-specific topic distribution. The second strategy involved the use of a regularization form to constrain the topic model by the query information. In essence, the query-specific topics were employed to bias the topic model.

Other summarization methods considered diversity as one of the major goals of the extractive-based generic and focused summarizations. Recently, there were a growing number of works which attempted to integrate diversity into the sentence ranking function itself. For example, Zhu et al. (2007) proposed a unified ranking algorithm called GRASSHOPPER which is based on random walks over an absorbing Markov chain. The representative sentences which have been selected into the summary become absorbing states, effectively transforming their transition probabilities to zero. The absorbing nodes will drag down the scores of the adjacent nodes as the walk gets absorbed. On the other hand, the nodes which are far away from the absorbing nodes still get visited by the random walk. Next, Li et al (2009) casts the diversity issue as the optimization under constraints problem. They propose a supervised method based on structural learning which incorporates diversity as a set of subtopic constraints. Then, they train a summarization model

and enforce diversity through the optimization problem. Wan et al. (2006) proposes a cross-document random walks to extract a focused summary with high information richness and novelty. They introduce a diversity penalty imposition step to remove redundancy after the initial list of representative sentences has been extracted. After each top-ranked sentence i in the initial list is selected into the summary, the scores of all adjacent sentences to i will be penalized.

In general, diversity is one of the most important topics in many related areas, particularly in information retrieval. Perhaps, the most well-known work is Maximal Marginal Relevance (MMR) (Carbonell and Goldstein 1998) in which redundancy reduction method is first introduced to rerank the search results. Since then, it has become the most commonly used method to reduce redundancy in text summarization. Subsequent works in information retrieval research attempt to establish a theoretical framework of diversity ranking and evaluation (Agrawal et al. 2009; Clarke et al. 2008; Zhai et al. 2003). Considering related works in text summarization, most graph-based ranking models (Chen et al. 2009; Otterbacher et al. 2005; Zhu et al. 2007) are inspired by the PageRank algorithm (Brin and Page 1998). Therefore, they employ eigenvector centrality to measure the importance of nodes in sentence graph. Under this model, a node is considered to be important if it is linked to other important nodes. Simply, it receives a high recommendation vote from the adjacent nodes.

In the context of graphical models, there have been several attempts to incorporate negative edges into the traditional graph representations. These include the areas such as trust/distrust ranking (de Kerchove and Dooren 2008; Guha et al. 2004) and social network mining (Kunegis et al. 2009; Yang et al. 2007). For example, de Kerchove et al. (2008) proposed the PageTrust algorithm as an

extension to the original PageRank algorithm by including negative links as the propagation of distrust among web pages. Their method ranks the nodes using both positive and negative links. Similarly, Kunegis et al. (2009) defined an eigenvector ranking method called signed spectral ranking which considers both positive and negative links to model friend and foe relationships in the social network.

Finally, the semantic structure of sentence has been applied in a few text mining and information retrieval applications. In text categorization, Shehata et al. (2007) propose conceptual term frequency as a new term weight scheme computing at sentence semantic level. It has been applied to text classification task. Next, Wang et al. (2008) utilized a simple structural composition of sentences to compute the similarity scores in multi-document summarization. Next, Bilotti et al. (2007) explored the use of semantic roles to create structural search queries. Most applications of sentence semantics were based on semantic role labeling research in natural language processing domain. Gildea and Jurafsky (2002) first introduced a machine learning approach to automatically label sentence constituents with proper semantic roles. Their classifier was trained on FrameNet data (Baker et al. 1998) using various linguistic features, such as verb, head nouns, syntactic category, active/passive voice label, and grammatical function. A subsequent work by Pradhan et al. (2004) explored a shallow semantic parsing approach to train a multi-class Support Vector Machines (SVM) classifier for semantic role labeling task.

CHAPTER 3: USING SEMANTIC, SYNTACTIC, AND CATEGORICAL INFORMATION TO FIND SIMILAR INTERROGATIE SENTENCES

"Exploratory research is really like working in a fog.

You don't know where you're going. You're just groping.

Then people learn about it afterwards and think how straightforward it was."

--Co-discoverer of DNA

& scientist until the bitter end

3.1 Introduction

Knowledge-sharing/question-answering communities, such as Yahoo! Answers, and digital reference services, such as IPL2 and Ask Dr. Math, have been collecting a significant number of questions. Given the current magnitude of questions and answers in their archive, it is likely that a newly submitted question has already been asked before by other users. However, finding such similar questions is ineffective due to the inherited limitation of the current search engines. Standard text retrieval approaches that compute the similarity of a document-level text are neither effective nor efficient for matching natural language questions. First, the fundamental principle of document similarity techniques is based on the degree of word overlaps. This notion works well in distinguishing similar documents since they are likely to contain sufficient number of words in common. On the other hand, the length of question phrases is relatively short and often contains very few word overlaps. Furthermore, due to the generative power of natural language, the same

question can be expressed in various ways. Hence, most questions are likely to receive a low similarity score from document similarity measures.

In this chapter, we investigate how question similarity judgment can be computationally improved. We define the notion of similarity between questions as those that share the same information need. The reliable measure needs to be able to match interrogative sentences according to their lexical semantic, and syntactic variations. Since an information need represents a far more specific notion of relevance than a topical relevance notion used in standard information retrieval approaches, traditional text similarity measures are not likely to perform well. To that end, we propose an approach to evaluate the similarity between questions based on semantic, syntactic, and question category information. Semantic information was derived from a lexical resource while syntactic information was provided by a shallow natural language parsing. We employ the information about the types of questions, provided by a trained text classifier, to further differentiate similar/dissimilar questions. These components are combined linearly.

The rest of the chapter is organized as follows. First, we discuss the research questions being tested in this chapter. In section 3.2 Then, we describe our approach to determine question similarity in section 3.3. In section 3.4, we describe the experimental set up and discuss the results in section 3.4. Finally, we conclude the chapter in section 3.6.

3.2 Research Question Tested

This chapter focuses on answering the first research question described as follows:

RQ1: *What are the useful resources that helps improve the similarity judgment at sentence level? How can we incorporate them into the similarity function?* The chapter explores the issues in the context of question similarity judgments. In particular, the work described in this chapter aims to find out what are the components that can be integrated into the similarity function to identify semantically similar questions? To achieve that, we introduce the hybrid sentence semantics and question category approach in the next section.

3.3 The Hybrid Sentence Semantics and Question Category Approach

We propose the hybrid method based on the combinations of three different components: semantic similarity, syntactic similarity and question category similarity. The combination of the first two components can be regarded as generic sentence similarity component, which is based on Li et al.'s method (2006). The third component, question category, provides question-specific information to the overall similarity measure. The idea behind the question category is that similar questions may share the same interrogative words. For example, location-related questions typically start with *where* while temporal-related questions usually start with *when*. To quantify the similarity between words in the sentence, semantic information was obtained from WordNet (Fellbaum 1998). A part-of-speech tagger was employed to analyze the syntactic information of the question phrases, i.e. word order and part of speech labels. While a deep natural language processing technique might provide greater syntactic information of the sentences, our reason to use shallow NLP technique, i.e. part of speech tagging, was to balance the tradeoff between the effectiveness and efficiency of the similarity measure. The equation below describes the question similarity measure between questions *q1* and *q2* as follow.

$$sim(q_1, q_2) = \beta \cdot (\alpha \cdot sem(q_1, q_2) + (1 - \alpha) \cdot syn(q_1, q_2)) + (1 - \beta) \cdot cat(q_1, q_2) \quad (3.1)$$

Two component coefficients were used to fine tune the similarity components. First, we optimized two sub-components within the sentence similarity component: semantic similarity (*sem*) and syntactic similarity (*syn*), through α. Then, we controlled the influence of sentence similarity and question category similarity (*cat*) components via β. All component coefficients have a real-number value ranging from 0 to 1.

To produce the actual question similarity score, each component will be replaced by the appropriate sentence similarity measures, which are described in section 3.2.2, and question category similarity measure, described in section 3.2.3. For example, either sentence vector similarity or part-of-speech semantic similarity measures can be plugged into the semantic similarity component. This results in a number of similarity measure combinations, which we described in section 3.2.4. Finally, most sentence similarity measures rely on the comparison of individual words between two sentences. Such comparison requires word similarity measures which is described in the next section.

3.3.1 Word Similarity Measures

First, we selected two candidate measures to compute word similarity scores: universal similarity (Lin 1998) and gloss overlap measures (Achananuparp 2007). The two measures were chosen because of their superior performance to the conventional path-based similarity measures. Mainly, Lin's measure combines local similarity judgment with global term information from information content value while gloss overlap measure only computes word similarity on a local basis. The

similarity value produced by both measures has a real-number value ranging from 0 (not similar) to 1 (identical).

Universal Similarity Measure In this measure, the similarity between two words, w_1 and w_2 is determined by their information content and the path distance in WordNet hierarchies. Here, we used Resnik's formulation (1995) of information content which defines the information content of concept c as the negative log likelihood function $-log(p(c))$, where $p(c)$ is the probability of encountering such concept c.

Gloss Overlap Measure The Gloss overlap approach for measuring word similarity was first introduced by Lesk (1986). Our variation of gloss overlap similarity between two words is defined as the overlap between their dictionary definitions or *glosses* and their direct hypernym and hyponym in WordNet hierarchies. The overall similarity measure is formulated as follows.

We empirically tested the correlation with human judgment for both measures on the selected noun pairs from the standard Rubenstein and Goodenough (R&G) data set used in (Li et al. 2006) and found that both correlated highly with human judgment. The universal similarity measure performed slightly better than the gloss-overlap measure (r_{lin}=0.924 and r_{gloss}=0.901). Therefore, we employ the universal similarity measure to compute the word-level semantic similarity.

3.3.2 Sentence Similarity Measures

All similarity measures used in this work rely on a pair-wise comparison between words in the two sentences. To select the best score for each word pairs, we performed a simple word sense disambiguation by choosing the maximum similarity

score. The similarity score generated by all three measures has a real-number value ranging from 0 (not similar) to 1 (identical).

Sentence Semantic Similarity The motivation behind semantic measures was to distinguish similar sentences beyond their surface form by utilizing semantic information of words in the sentences. Such information is typically obtained from linguistic resources such as WordNet (Fellbaum 1998). A few number of WordNet-based similarity measures have been proposed to calculate semantic similarity between words. For a comprehensive comparison of word similarity measures, we recommend the readers to the work done by Budanitsky and Hirst (2006). For sentence semantic measures evaluated in this work, we use the universal similarity measure defined in Lin (1998) to compute word similarity scores.

There are several approaches to utilize word semantic similarity scores to determine similarity between sentences. First, Li et al. (2006) suggested the semantic vector approach to compute sentence similarity. Like the traditional vector space model, sentences are transformed into feature vectors having words from sentence pair as a feature set. In contrast, term weights, in this case, are derived from the maximum semantic similarity score between words in the feature vector and words in a corresponding sentence. In addition, Li et al. also include the importance of words in term weight calculation by multiplying word similarity score with information contents of corresponding word feature and its associated word in the sentence. In this work, we simplify the sentence similarity measure by only using word similarity scores as term weights. Moreover, instead of exhaustively calculating word similarity scores for all possible word pairs, we only compute semantic similarity of words within the same part-of-speech class, e.g. noun vs.

noun, verb vs. verb, etc. Then, the semantic similarity between sentence pair is computed as the cosine similarity between semantic vectors of the two sentences.

$$sim_{ssv}(s_1, s_2) = \frac{sv_1 \cdot sv_2}{|sv_1||sv_2|} \tag{3.2}$$

where sv_1 and sv_2 are the semantic vectors for sentence s_1 and s_2, respectively. Another semantic similarity measure, proposed by Mihalcea et al. (2006), linearly combines word semantic similarity scores with word specificity scores. Given two sentences s_1 and s_2, the sentence similarity computation begins by finding the maximum word similarity score for each word in s_1 with words in the same part of speech class in s_2. Next, the same process is applied for each word in s_2 with the corresponding word in s_1. Then, the derived word similarity score is weighted with IDF scores that belong to the corresponding word. Formally, the IDF-weighted sentence semantic similarity measure is defined as follow.

$$sim_{sem,IDF}(s_1, s_2) = \frac{1}{2}\left(\frac{\sum_{w\in\{s_1\}}(\max sim(w, s_2) \times idf(w))}{\sum_{w\in\{s_1\}} idf(w)} + \frac{\sum_{w\in\{s_2\}}(\max sim(w, s_1) \times idf(w))}{\sum_{w\in\{s_2\}} idf(w)}\right) \tag{3.3}$$

where max $sim(w,s_i)$ is the maximum semantic similarity score of w and words in s_i that belong to the same part-of-speech as w while $idf(w)$ is an inverse document frequency of w. The reason for computing the semantic similarity scores only between words in the same part of speech class is that most WordNet-based measures are unable to compute semantic similarity of cross-part-of-speech words. Since there are no explicit relations between different concept hierarchies in WordNet, most path-based similarity measures will judge a cross part-of-speech word pair as unrelated.

Malik et al. (2007) have adopted a simplified variation of Mihalcea et al.'s measure (2006) by dropping the word specificity component. That is, they compute the sentence similarity score based on the sum of the maximum word similarity scores of words in the same part-of-speech class normalized by sentence lengths. Their formulation is described as follow.

$$sim_{sem}(s_1, s_2) = \frac{\sum_{w \in \{s_1\}} \max sim(w, s_2) + \sum_{w \in \{s_2\}} \max sim(w, s_1)}{|s_1| + |s_2|} \tag{3.4}$$

Word Order Similarity Apart from lexical semantics, word composition also plays a role in sentence understanding. Basic syntactic information, such as word order, can provide useful information to distinguish the meaning of two sentences. This is particularly important in many similarity measures where a single word token was used as a basic lexical unit when computing similarity of sentences. Without syntactic information, it is impossible to discriminate sentences that share the similar bag-of-word representations. For example, "the sales manager hits the office worker" and "the office manager hits the sales worker" will be judged as identical sentences because they have the same surface text. However, their meanings are very different. To utilize word order in similarity calculation, Li et al. (2006) define word order similarity measure as the normalized difference of word order between the two sentences. The formulation for word order similarity is described as follow:

$$sim_{wo}(s_1, s_2) = 1 - \frac{||r_1 - r2||}{||r_1 + r_2||} \tag{3.5}$$

where r_1 and r_2 is a word order vector of sentence s_1 and s_2, respectively. The steps to build a word order vector are similar to sentence vector's process. That is, a feature set of word order vector is taken from the individual words of the two sentences. Each entry in the word order vector is derived by comparing word feature w_i with each word in the sentence. If the two are identical, then we fill the entry of w_i with an index number (word position) of the corresponding word. Otherwise, we calculate word similarity score between w_i and the remaining words in the sentence and fill w_i entry with an index number of a matching word that gives a maximum similarity score.

3.3.3 Question Category Similarity Measure

In this measure, we focus on the interrogative words containing in question construction as they can be utilized as useful features to distinguish questions from any generic sentences. Moreover, these words can be used to determine aboutness of the questions. Given two questions constructed from a near-identical set of words, the interrogative words serve as surrogates of the question category that helps distinguish between the two. For instance, we can infer that "where was JFK assassinated?" and "when was JFK assassinated?" are two different questions judging by different *wh*-pronouns: where (location modifier) and when (temporal modifier). Thus, we defined the similarity measure around the idea that similar questions share the same interrogative words or categories.

That is, the proposed question category similarity is computed as a cosine similarity between the question category vectors. As it can be seen, the major step in our approach is constructing the question category vector. For the task of classifying questions into different types, we employed Support Vector Machine (SVM) as the

underlying classifier based on its overall performance in text classification (Zhang and Lee 2003). In this work, we used SVMLight (Joachims 1998) to train the classifier.

We developed SVM classifier using linear kernel to predict the question categories. The features for the classifier include unigram, multiword collocations, and the hypernyms of the head nouns (the head of the noun phrases). Specifically, we restricted the head nouns to those following the interrogative words. For instance, a head noun of the question "What tourist attractions are there in Reims?" is the word "tourist". A list of multiword collocations, including interrogative words, was compiled from the training example. For example, "how many", "how much", "what is a", "what is the" were automatically identified and extracted by the aforementioned tool. In the testing stage, we simply used exact string match to identify multiword collocations. The hypernyms of the head noun serve as semantic features which increase the chance of semantically-similar concepts sharing common features. The method of extracting head nouns and their hypernyms are the same as the one in Metzler and Croft (2005). The classifier was built on the UIUC dataset (Li and Roth 2002) which is a superset of the TREC QA track dataset. The UIUC dataset contains 5,500 training questions and 500 TREC-10 questions for testing. Their question class taxonomy contains two levels. The coarse level has six categories whereas the fine level has fifty categories. The classification precisions for coarse-grained and fine-grained taxonomies are 81.8% and 89.2%, respectively. In this study, we classified questions based on the fine-grained categories due to their superior performance. Moreover, we took a multi-label classification approach to categorize questions. As such, a question was classified into one or more categories.

3.3.4 The Combined Semantic and Syntactic Measures

Using the notion that both semantic and syntactic information contribute to the understanding of a sentence, Li et al. (2006) defined the sentence similarity measure as a linear combination of semantic vector similarity and word order similarity (equation 3.6). The relative contribution of semantic and syntactic measures is controlled by a coefficient alpha. It has been empirically proved (Li et al 2006) that a sentence similarity measure performs the best when semantic measure is weighted more than syntactic measure. This follows the conclusion from a psychological experiment conducted by (Landauer et al. 1997) which emphasizes the role of semantic information over syntactic information in passage understanding. We tuned the parameters α and β of the hybrid measure using 10% of the test data. Then, we selected the values that produced the optimal results. In this case, $\alpha = 0.8$ and $\beta = 0.5$.

We also experiment with a minor variation of the combined sentence similarity formulation by employing Malik et al.'s formulation to compute the sentence semantic similarity (equation 3.7). The same semantic coefficient value is applied.

$$sim_{ssv+wo}(s_1, s_2) = \alpha \cdot sim_{ssv}(s_1, s_2) + (1 - \alpha) \cdot sim_{wo}(s_1, s_2) \quad (3.6)$$

$$sim_{sem+wo}(s_1, s_2) = \alpha \cdot sim_{sem}(s_1, s_2) + (1 - \alpha) \cdot sim_{wo}(s_1, s_2) \quad (3.7)$$

3.4 Experimental Evaluation

3.4.1 Data Sets

To evaluate the performance of the question similarity measures, we selected a set of 193 question pairs from TREC-9 (Voorhees 2001) question variants key. The

variants key consists of fifty four original questions and their variants. The original questions are a subset of test questions used in TREC-9 QA experiment and were taken from the actual users' submissions. The question variants are the paraphrased questions that were constructed by human assessors to be semantically identical but syntactically different from the original questions. 386 question pairs are used as a test set -- 50% of which are the positive pairs. Although the data set is semi-artificial, it contains sufficient linguistic complexity to reflect the variability of nature language expressions. That is, there are various types of paraphrasing strategies (Tomuro 2003) exhibited in the question variants. For example:

- Lexical substitution: What kind of animal was Winnie the Pooh? vs. what species was Winnie the Pooh?
- Morpho-syntactic variations: What kind of animal was Winnie the Pooh? vs. Winnie the Pooh is what kind of animal?, who owns CNN? vs. CNN is owned by whom?
- Interrogative reformulation: How did Bob Marley die? vs. what killed Bob Marley?
- Semantic inference: What tourist attractions are there in Reims? vs. What do most tourists visit in Reims?

Over 50% of the paraphrases were categorized into multiple categories. Additional descriptive summary of the test set is displayed in table 3.2.

Table 3.1 The composition of paraphrase categories in TREC-9 question variants.

Paraphrase Category	Lexical Substitution	Morpho-Syntactic Variation	Interrogative Reformulation	Semantic Inference
# pairs	63	97	112	31

Table 3.2 Summary of TREC-9 data sets used in the experiment.

Number of sentence pairs	386
Number of unique words	252
Percentage of unique words covered by WordNet	84.5%
Average question length (in characters)	39.35
Degree of symmetry between two comparing questions (in characters)	4.32

3.4.2 Preprocessing

The preprocessing steps are described as follows. First, we broke down each question into single-word tokens and assigned a part-of-speech tag for each token. To preserve word meaning, we did not stem the token. Next we filtered out any functional words -- words that do not contain semantic content such as articles, pronouns, prepositions, conjunctions, auxiliary verbs, modal verbs, and punctuations. All cardinal numbers were kept. Then, we compute the word similarity scores for all possible word pairs and the results were cached for later use.

3.4.3 Evaluation Criteria

We are specifically interested in comparing the effectiveness of different similarity measures in predicting positive cases (semantically-equivalent questions) and negative cases (unrelated questions). To achieve that, we define six evaluation metrics, *recall*, *precision*, *rejection*, F_1, and f_1, based on the predicted positive and negative judgments as follows.

Recall is a proportion of correctly predicted question paraphrases compared to all question paraphrases. Precision is a proportion of correctly predicted question paraphrases compared to all predicted question paraphrases. Rejection is a proportion of correctly predicted unrelated questions compared to all unrelated questions. Accuracy is a proportion of all correctly predicted questions compared to all questions. F_1 is a uniform harmonic mean of precision and recall. Lastly, f_1 is a uniform harmonic mean of rejection and recall. Following a similar experiment done by Mihalcea et al (2006), the scoring threshold for positive pairs is set at 0.5. Note that precision and rejection are the two similar metrics whose values change in the same direction. Although precision compares the ratio of true positive while rejection compares the ratio of true negative, both metrics rely on a number of false positive cases as part of the denominator. A high number of false positive cases result in low precision and low rejection. We include rejection and f_1 metrics in addition to the standard precision-recall based metrics as it presents another aspect of the performance based on the tradeoff between true positive and true negative judgments.

3.5 Results and Discussion

Table 3.3 Comparison of the performance of different similarity measures on TREC9 data set.

Similarity Measures	Prec.	Rec.	Rej.	F_1	f_1	Acc.
jaccard	1	0.383	1	0.554	0.554	0.691
cosine	1	0.762	1	0.865	0.865	0.881
sim_{wo}	0.644	0.487	0.731	0.555	0.584	0.609
sim_{ssv+wo}	0.68	0.979	0.539	0.803	0.695	0.759
sim_{sem+wo}	0.963	0.933	0.964	0.948	0.948	0.948
$sim_{sem+wo+cat}$	**0.987**	**0.98**	**0.93**	**0.983**	**0.954**	**0.986**

The performance of different similarity measures is shown in table 3.3. By weighing the contributions of sentence similarity and question category similarity components equally, the hybrid question similarity measure ($sim_{sem+wo+cat}$) performed the best across all evaluation metrics. It produced the optimal scores of 0.983, 0.954, and 0.986 on F_1, f_1, and accuracy metrics, respectively. This confirms our expectation that the addition of question category information helps improve the overall effectiveness of question similarity judgment. Compared to the performance of the second-best measure, F_1 and classification accuracy was improved approximately by 4% while f_1 was improved by 0.06% which was not statistically significant. Interestingly, according to sim_{sem+wo}'s performance, the combination of semantic and syntactic information alone was almost as effective in identifying paraphrases. We believe this outcome was not entirely surprising. One explanation is that the sentence similarity measure was able to implicitly infer the categorical information of questions to some extent based on the combination of word-level semantic

similarity and syntactic similarity components. Secondly, the performance of the hybrid measure depended on the effectiveness of the question classifier. In this case, our trained classifier performed at 89% accuracy which was reasonably high. Still, there might be some cases where questions which shared very few common words, were wrongly classified into the same category. In such cases, the overall question similarity judgment could be biased toward not rejecting the negative pairs.

Additionally, the two naïve measures, Jaccard coefficient and cosine similarity, also produced an extremely strong rejection rate. That is, they were always reject unrelated questions. This is explained by the fact that the negative pairs in TREC9 contain a relatively small number of lexical overlaps. Finally, the syntactic-only similarity measure produced the least accurate result. This is not surprising since paraphrases are mainly judged based on their common meanings.

3.6 Conclusions

In this chapter, we demonstrated that the proposed question similarity measure is effective in identifying paraphrased questions. Semantic and syntactic measures were helpful in handling synonyms, related words, and different sentence compositions. The addition of question category information has significantly improved the performance of the similarity measure by providing a discriminative power from the specific words in the interrogative sentences. We recognized certain shortcomings in the use of TREC-9 data set since it is partially artificial. Some of interrogative sentences used as the paraphrased samples in the experiment were created by human experts, therefore, they might not fully reflect the complexities and variations of the questions being formulated by the real-world information

seekers. Moreover, most questions in TREC-9 data set are short factoid questions which only cover a subset of those being queried in the real-world settings.

This chapter answers research question 1 by presenting a hybrid question similarity measure that utilizes the semantic, syntactic, and categorical information for identifying paraphrased question pairs. It demonstrated that, in the context of interrogative sentences, these components were very effective in improving the accuracy of paraphrase recognition task. The semantic component computed the word-level semantic similarity score between interrogative questions using a WordNet-based semantic similarity measure. Next, the syntactic component calculated the syntactic similarity between questions according to the differences in their compositional orders. Lastly, the question category component computed the similarity between questions based on the cosine similarity of the category vectors. The optimal performance was achieved by weighing the semantic component more than syntactic component and weighting sentence similarity component equally to the question category component.

CHAPTER 4: IMPROVING THE SIMILARITY JUDGMENT THROUGH SENTENCE SEMANTIC STRUCTURE

"Basic research is what I am doing when I don't know what I am doing."

--Ex-Nazi rocket scientist

who helped land the first men on the Moon

4.1 Introduction

A major issue that many text mining applications have to deal with is the variability of natural language expression. Due to flexibility of human language, the same information can be expressed in numerous ways. For instance, in Monty Python's humorous dead parrot sketch , more than fifteen variations of an expression "the parrot is dead" have been uttered as euphemisms. These include "the parrot has ceased to be," "the parrot is no more," and "this is the dead parrot." This issue has a great implication in many sentence extraction applications, such as text summarization and question answering , where the identification of novel or redundant information is crucial to system performance. In text summarization context, a good sentence extraction module should be able to recognize such variations; otherwise, a lot of redundancy might occur in the extracted summary and a lot of relevant information will be left out. In redundancy-based question answering application, the candidate answers may be expressed in sentences with different vocabulary and/or a syntactic construction. Without an effective measure, QA systems are unable to effectively make use of redundant information.

Another important issue is about the notions of sentence similarity. In the past, many text mining and information retrieval applications, e.g. text classification and clustering, rely on common word occurrences to measure the similarity between text units. This approach might not work well in sentence-level applications because of limited context imposed by sentence. As previously described in the literature review, many research communities have focused on defining the notion of sentence similarity over the past few years. For instance, in information retrieval, the levels of topical similarity between sentences are proposed (Metzler et al 2005; Murdock 2006). Relevant sentences either address the same specific topics or they might talk about the similar general topics. In the natural language processing (NLP) community, two notions of sentence similarity are under studied, semantic equivalence and entailment (Dolan et al 2004). These two notions are strongly related, and a clear distinction is difficult to define. Two sentences are considered to be semantically equivalent if they share the exact same meaning. That is, if we can make a bidirectional inference between them. Following this definition, paraphrase is considered the most common form of semantic equivalent sentence. On the other hand, entailment focuses on unidirectional inference. If the meaning of one sentence can be inferred from the other sentence, the two sentences are said to be an entailment pair.

This chapter introduces the approaches that employ semantic structure of sentences to improve the accuracy of sentence similarity judgment. Traditionally, sentences are represented as an unstructured bag of words in similarity computation. This results in an information loss because syntactic and semantic information of the sentences is ignored. This has a particularly crucial consequence to the identification of semantic equivalence or entailment sentences in which a very

specific inference has to be made. For example, two sentences, "John bought a car from Bill" and "Bill bought a car for John," share the same bag of words representation {John, bought, car, Bill}, thus they are judged to be identical by naïve measures. However, it can be seen that each sentence uniquely describes an event because of differences in subjects and objects. Our proposed method copes with the issue by utilizing the information about semantic roles of constituents in the sentences and computing sentence similarity at sentence semantic level. We believe a more accurate similarity judgment can be made between semantically related components.

The rest of the chapter is organized as follows. First, we introduce the research question being tested in this chapter in section 4.2 and describe the proposed method in section 4.3. Then, in section 4.4, we outline the experimental evaluation, including data sets and evaluation metrics used in this study. Lastly, we discuss about the results and conclude the chapter in section 4.5 and 4.6, respectively.

4.2 Research Question Tested

This chapter focuses on answering the first research question described as follows:

RQ1: What are the useful resources that helps improve the similarity judgment at sentence level? How can we incorporate them into the similarity function? Specifically, we are interested in investigating how semantic knowledge of sentences can be integrated into the sentence similarity measures? What is their effectiveness on the similarity judgments of the generic sentences? How effective are they in dealing with two specific semantic similarity notions: semantic equivalence and textual entailment? Particularly, we anticipate that the structural similarity measure, described in the later part of this chapter, is significantly more effective

than most occurrence-based similarity measures in handling the complex judgments involved in textual entailment classification task.

4.3 Sentence Similarity Measures

First, we briefly introduce the existing similarity measures that can be used for identifying the similarity between sentences. We categorize these measures into three different approaches: word overlap, TFIDF-based, and knowledge-based measures. Note that all similarity score produces by these measures have a real-number value ranging from 0 (unrelated sentences) to 1 (identical sentences).

4.3.1 Word Overlap Measures

Word overlap measures is a family of combinatorial similarity measure that compute similarity score based on a number of words shared by two sentences. In this work, we consider four word overlap measures: Jaccard similarity coefficient, simple word overlap, IDF overlap, and phrasal overlap.

Jaccard similarity coefficient Jaccard coefficient is a similarity measure that compares the similarity between two feature sets. When applying to sentence similarity task, it is defined as the size of the intersection of the words in the two sentences compared to the size of the union of the words in the two sentences.

$$sim_{jaccard}(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} \tag{4.1}$$

Word overlap and IDF overlap measures Metzler et al. (2005) defined two baseline word overlap measures to compute the similarity between sentence pairs. Simple word overlap fraction is defined as the proportion of words that appear in both sentences (equation 4.2), while IDF overlap is defined as the proportion of

words that appear in both sentences weighted by their inverse document frequency (equation 4.3).

$$sim_{overlap}(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1|} \tag{4.2}$$

$$sim_{overlap,IDF}(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1|}(\sum_{w \in s_1 \cap s_2} \log \frac{N}{df_w}) \tag{4.3}$$

where N is a total number of sentences in a text collection. df_w is a number of documents that contain the word w.

N-gram phrasal overlap measure Banerjee and Pedersen (2003) introduced the overlap measure based on the Zipfian relationship between the length of phrases and their frequencies in a text collection. According to Zipf's law, longer phrases tend to occur fewer times than shorter phrases. Their motivation stems from the fact that a traditional word overlap measure simply treats sentences as a bag of words and does not take into account the differences between single words and multi-word phrases. Since a phrasal n-word overlap is much rarer to find than a single word overlap, thus a n-gram phrasal overlap calculation for m phrasal n-word overlaps is defined as a non-linear function displayed in the equation below.

$$overlap_{NGRAM}(s_1, s_2) = \sum_{i=1}^{n} \sum_{m} i^2 \tag{4.4}$$

where m is a number of i-word phrases that appear in sentence pairs. For example, given two sentences "a cock is an adult male chicken" and "a rooster is an adult male chicken", a phrasal overlap between the two sentences is calculated from a sum of three one-word overlaps (adult, male, and chicken), two two-word overlaps (adult male and male chicken), and one three-word overlap (adult male chicken),

which is equal to 3 + 8 + 9 = 20. Ponzetto and Strube (2007) perform the normalization on equation 4.4 by the sum of sentence length and apply the hyperbolic tangent function to minimize the effect of the outliers. The normalized phrasal overlap similarity measure is defined in the following equation.

$$sim_{NGRAM}(s_1, s_2) = tanh(\frac{overlap_{NGRAM}(s_1, s_2)}{|s_1| + |s_2|}) \tag{4.5}$$

4.3.2 TF-IDF Measures

Three variations of measures that compute sentence similarity based on term frequency-inverse document frequency (TFIDF) are considered in this study.

TFIDF cosine similarity First, standard vector-space model represents a document as a vector having indexing words as a feature set and TFIDF as term weights. For sentence similarity task, we adopt the standard vector-space approach to compare the similarity between sentence pairs by computing a cosine similarity between the vector representations of the two sentences. A slight modification is made for sentence representation. Instead of using indexing words from a text collection, a set of words that appear in the sentence pair is used as a feature set. This is done to reduce the degree of data sparseness in sentence representation. The standard TFIDF cosine similarity is defined as follow

$$sim_{TFIDFcosine}(s_1, s_2) = \frac{sv_1 \cdot sv_2}{|sv_1||sv_2|} \tag{4.6}$$

where sv_1 and sv_2 is a vector representation of sentence s_1 and s_2, respectively.

Novelty Detection Measure Allan et al. (2003) proposed TFIDF measure for detecting topically similar sentences in TREC novelty track experiment. The

formulation is based on the sum of the product of term frequency and inverse document frequency of words that appear in both sentences.

$$sim_{TFIDFnovelty}(s_1, s_2) = \sum_{w \in s_1 \cap s_2} \log(tf_{w,s_1} + 1) \log(tf_{w,s_2} + 1) \log(\frac{N+1}{df_w + 0.5}) \quad (4.7)$$

Where $tf_{w,s1}$ is a number of occurrences of w in s_1 while $tf_{w,s2}$ is a number occurrences of w in s_2. df_w is a number of documents that contain w and N is a total number of sentences in a text collection.

Identity Measure Identity measure (Hoad and Zobel 2003) is another variation of TFIDF similarity measure. It was originally proposed as a measure for identifying plagiarized documents or co-derivation and has been shown to perform effectively for such application. The motivation underlying this measure is that a similarity measure should consider differences in the number of word occurrences between two documents. The simplified variation of the formulation used in (Metzler et al. 2005) is displayed in the equation below.

$$sim_{TFIDFidentity}(s_1, s_2) = \frac{1}{1 + \frac{\max(|s_1|,|s_2|)}{\min(|s_1|,|s_2|)}} \sum_{w \in s_1 \cap s_2} \frac{\log \frac{N}{df_w}}{1 + |tf_{w,s_1} - tf_{w,s_2}|} \quad (4.8)$$

where N is a total number of sentences in a text collection. df_w is a number of documents that contain the word w. $tf_{w,s1}$ is a number of occurrences of w in s_1 while $tf_{w,s2}$ is a number occurrences of w in s_2. It can be seen that the identity measure is derived from the sum of inverse document frequency of the words that appear in both sentences normalized by the overall lengths of the sentences and the relative frequency of a word between the two sentences.

4.3.3 Knowledge-Based Measure

Knowledge-based measure generally refers to those that employ external knowledge bases, such as WordNet, to compute similarity scores. Mihalcea et al. (2005) proposed sentence similarity method that combines word semantic similarity scores with word specificity scores. Given two sentences s_1 and s_2, the sentence similarity calculation begins by finding the maximum word similarity score for each word in s_1 with words in the same part of speech class in s_2. Then, apply the same procedure for each word in s_2 with words in the same part of speech class in s_1. The derived word similarity scores are weighted with *idf* scores that belong to the corresponding word. Finally, the sentence similarity formulation is defined in the following equation.

$$sim_{sem,IDF}(s_1, s_2) = \frac{1}{2}\left(\frac{\sum_{w\in\{s_1\}}(\max sim(w, s_2) \times idf(w))}{\sum_{w\in\{s_1\}} idf(w)} + \frac{\sum_{w\in\{s_2\}}(\max sim(w, s_1) \times idf(w))}{\sum_{w\in\{s_2\}} idf(w)}\right) \quad (4.9)$$

where $max\ sim(w,s_i)$ is the maximum semantic similarity score of w and words in s_i that belong to the same part-of-speech as w while $idf(w)$ is an inverse document frequency of w. The reason for computing the semantic similarity scores only between words in the same part of speech class is that most WordNet-based measures are unable to compute semantic similarity of cross-part-of-speech words. Since there are no explicit relations between different concept hierarchies in WordNet, most path-based similarity measures will judge a cross part-of-speech word pair as unrelated.

4.4 Utilizing Semantic Structure to Measure Sentence Similarity

One major drawback of the existing approaches described in the previous section is that they compute the similarity scores based on common word occurrences between

sentences, ignoring semantic and syntactic construction of sentences. Thus, in this section, we propose the approaches that incorporate the underlying semantic structure of sentences to measure sentence similarity. The semantic structure of sentences, referred to as verb-argument structure, encodes the relations between individual components and their semantic roles with respect to a given verb in a sentence. The labels of semantic roles are varied depending on the annotation scheme (Baker et al. 1998; Palmer et al. 2005). Generally, Arg0 denotes a prototypical agent, Arg1 indicates a prototypical patient or theme of a given verb, and ArgM represents adjunctive argument (e.g. ArgM-LOC specifies location-related argument). For instance, "*Data mining identifies trends within data that go beyond simple data analysis*" consists of two following verb-argument structures:

[$_{Arg0}$ Data mining][$_{rel}$ identifies][$_{Arg1}$ trends][$_{Arg2}$ within…analysis] and
[$_{Arg1}$ data][$_{rel}$ go][$_{Arg4}$ beyond simple data analysis]

Our motivation arises from the assumption that semantically similar sentences contain more similar verb arguments between each other than dissimilar sentences. By measuring semantic similarity of verb-argument structures, we can improve the effectiveness of sentence similarity measures despite the syntactic variability of language expression. Consider another simple sentence "a glass is broken". A verb-argument structure of this sentence is [$_{Arg1}$ a glass] is [$_{rel}$ broken]. Apparently, it describes a similar event as "John broke a glass" ([$_{Arg0}$ John] [$_{rel}$ broke] [$_{Arg1}$ a glass]) if we consider each matching component.

Two different approaches that employ semantic information in verb-argument structure are investigated. First, we describe a vector-space based

approach which computes semantic similarity scores of sentences based on conceptual term frequency weights in section 4.2.1. Then, we define a structural similarity approach which measures semantic similarity of two sentences by comparing the similarities of verbs and arguments between the two in section 4.2.2.

4.4.1 Conceptual Term Frequency Vector Approach

Conceptual term frequency (*ctf*) is defined as a number of occurrences of a concept in verb argument structures of a sentence (Shehata et al. 2007). It is based on the assumption that words or phrases that appear in a greater number of verb argument structures contribute more to sentence semantics than those that appear in a lesser number of verb argument structures. Shehata et al. (2007) define a concept-based weight (henceforth *CTF*) of term i in sentence j as a linear combination of its normalized term frequency (*tf*) and normalized conceptual term frequency (*ctf*).

$$CTF_i = \frac{tf_i}{||s_j||} + \frac{ctf_i}{||t_j||} \tag{4.10}$$

where tfi is a frequency of term i, ctf_i is a conceptual term frequency of i, s_j is a term-frequency vector of sentence j, and t_j is a conceptual term-frequency vector of sentence j.

In general, document frequency plays a role in determining the importance of terms and subsequently the values of term weights in document vectors. We apply the notion of inverse document frequency (*idf*) from information retrieval and scale it to sentence and verb-argument structure levels and define inverse sentence frequency (*isf*) as a function of sentence frequency and inverse verb-argument structure frequency (*ivf*) as a function of verb-argument structure frequency,

respectively. Combining *CTF* with the term importance measures, we derive the following equations:

$$CTFisf_i = CTF_i \times \log \frac{|S|}{sf_i} \tag{4.11}$$

$$CTFivf_i = CTF_i \times \log \frac{|S|}{vf_i} \tag{4.12}$$

where $|S|$ is a total number of sentences in the corpus, sf_i is a number of sentences where i appears, $|V|$ is a total number of verb-argument structures in the corpus, vf_i is a number of verb-argument structures where i appears.

Based on the aforementioned term weight formulae, we construct term-document matrix and measure the similarity between sentences according to cosine similarity of sentence vector representation. Two lexical units are employed to extract conceptual term features of sentence vectors: single words and multi-word phrases. To extract single word tokens, we remove functional words from the sentence but keep the cardinal numbers. Then, we stem word tokens using Porter Stemmer. To extract multi-word phrases, we perform a part-of-speech tagging and stem the sentences. Then, a syntactic rule similar to the one in (Park et al. 2002) is applied to extract noun phrases from sentences. In addition, phrase length is to 8-word limit. After preprocessing stage, we consider the generated single words and multi-word phrases as conceptual term features of sentence vectors. Once conceptual term features are extracted, *ctf*, *isf*, and *ivf* are determined by counting the occurrences of conceptual terms in verb-argument structures. Over the years, a number of tools and techniques have been developed to perform automatic semantic role tagging (Pradhan et al. 2004; Collobert and Weston 2007). In this work, we

employ SENNA (Collobert and Weston 2007), a neural-network based semantic role labeler, because of its efficient computation and high accuracy on the most frequent argument types (Arg0 and Arg1).

4.4.2 Structural Similarity Approach

The motivation behind this approach is that sentences that express the same event or idea should share the similar underlying semantic structure or verb-argument structures. Therefore, we represent sentences as a set of verb-argument structures instead of representing sentences as unstructured text. We define the structural similarity measure as follows. First, each sentence can be broken down into m verb-argument structures. Each verb-argument structure consists of verb r and n number of argument components. Each argument component is composed of text segment t. Then, given sentence i and j, the similarity score between verb-argument structures v_i and verb-argument structure v_j is determined by two similarity components: the verb similarity $V(r_i,r_j)$ and the argument similarity $A_k(t_i,t_j)$.

$$S(v_i, v_j) = \alpha \cdot V(r_i, r_j) + \frac{(1-\alpha)}{n} \cdot \sum_{k=0}^{n} A_k(t_i, t_j) \tag{4.13}$$

where α is a coefficient that controls the weight between verb similarity component and argument similarity component while n is a total number of argument components.

Verb similarity. We use a gloss-overlap similarity measure to compute the verb similarity $V(v_i,v_j)$. Essentially, two verbs are semantically similar if they share the same meaning measured by the textual overlap between their dictionary definitions. As each word (dictionary form) can carry multiple meanings (word

senses), the most similar senses are used to represent their corresponding lexical similarity. The following equations describe the similarity measure.

$$sim_{k,l}(r_i, r_j) = \frac{|g(k_i) \cap g(l_j)|}{|g(k_i) \cup g(l_j)|} \tag{4.14}$$

$$V(r_i, r_j) = \max_{k,l}[sim_{k,l}(r_i, r_j)] \tag{4.15}$$

where $sim_{k,l}(r_i,r_j)$ is the gloss-overlap similarity between a word sense k of verb r_i and a word sense l of verb r_j, $g(k_i)$ is a gloss (dictionary definition) of the word sense k of r_i and $g(l_j)$ is gloss of the sense l of r_j. Gloss is treated as a bag of words in the calculation. Then, the verb similarity $V(r_i,r_j)$ is obtained from gloss pair that gives the maximum gloss-overlap score. To obtain glosses, we search WordNet lexical taxonomy.

Intra-argument similarity. To compute the similarity of the matching argument classes, we consider argument texts as multi-word phrases and compute the similarity between text segments of the corresponding components based on their n-gram phrasal overlap score (Banerjee and Pedersen 2003; Ponzetto and Strube 2007). The formulas are defined in the equations below.

$$overlap(t_i, t_j) = \sum_{k=1}^{n} \sum_{m} k^2 \tag{4.16}$$

$$A_k(t_i, t_j) = tanh(\frac{overlap(t_i, t_j)}{|t_i| + |t_j|}) \tag{4.17}$$

where m is a number of k-word phrases that appear in text segments. Equation 4.17 is a normalized form of equation 4.16 via the hyperbolic tangent function to minimize the effect of the outliers .

Inter-argument similarity If an adjunctive argument ArgM is presence, we treat all of its subclasses, e.g. ArgM-LOC, ArgM-TMP, etc., as a single class ArgM. Then, we calculate its A_k score from all possible inter-argument comparison, such as ArgM vs. Arg0, ArgM vs. Arg1, etc. The maximum A_k score is chosen as the final score for ArgM. After that, the final ArgM score is added to the inter-argument similarity scores. Finally, the similarity of sentence i and j is derived from the verb-argument structure pair which produces the maximum $S(v_i,v_j)$ score.

$$sim(i,j) = \max_{v_i,v_j}[S(v_i,v_j)] \tag{4.18}$$

Lexical Expansion and Simplification We apply a set of syntactic rules to expand a single verb into a verb phrase. In addition, we remove any words that are not part of the longest noun phrases in argument components to simplify the argument text. For example, given a verb-argument structure:

[Arg1 BBC] [rel stands] [Arg2 for British Broadcasting Corporation]

A single verb "stands" will be expanded into a verb phrase "stands for". Arg1 text contains "BBC" and Arg2 text contains "British Broadcasting Corporation". This results in the following:

[Arg1 BBC] [rel stands for] [Arg2 British Broadcasting Corporation]

Moreover, we perform noun denominalization on those that contain an auxiliary verb. The auxiliary verb is replaced with a verb form of its adjacent noun.

After that, the denominalized noun is removed from the corresponding argument. For instance, given a verb-argument:

[Arg1 BBC] [rel is] [Arg2 the abbreviation of British Broadcasting Corporation]

The above verb-argument structure will be transformed into:

[Arg1 BBC] [rel abbreviates of] [Arg2 British Broadcasting Corporation]

4.5 Experimental Evaluation

4.5.1 Data Sets

We use two publicly-available sentence pair data sets, Microsoft Research paraphrase corpus (MSRP) (Dolan et al. 2004) and the third PASCAL recognizing textual entailment challenge (RTE3) data set (Dagan et al 2005), to evaluate the performance of the similarity measures.

MSRP contains 5,801 sentence pairs (4,076 training pairs and 1,725 test pairs) automatically constructed from various web new sources. Each sentence pair is judged by two human assessors whether they are semantically equivalent or not. Positive examples comprise 67% of the total sentence pairs. Semantically equivalent sentences may contain either identical information or the same information with minor differences in detail according to the principal agents and the associated actions in the sentences. In contrast, non-paraphrased sentences may contain several word overlaps, but they are judged to be not equivalent if they do not the same key information, i.e. principal agents and actions. In addition, sentence that describes the same event but is a superset of the other is considered to be a

dissimilar pair. Note that the latter rule is similar to the one used in text entailment task.

RTE3 consists of 800 sentence pairs from the development set and 800 sentence pairs from the test set. Each pair comprises two small text segments, which are referred to as *text* and *hypothesis*. The text-hypothesis pairs are collected by human assessors from four subsets of application domains: information retrieval, multi-document summarization, question answering, and information extraction. Similarity judgment between sentence pairs is based on directional inference between text and hypothesis. If the hypothesis can be entailed by the text, then that pair is considered to be a positive example. On the other hand, a negative example indicates that the hypothesis cannot be inferred from the text. Although the sentence judgment in RTE3 is different than the other sentence similarity judgments such as paraphrase recognition, topical relevance, etc., and many textual entailment methods often involve performing logical inference operations between text and hypothesis, we believe the comparison of sentence similarity measures on RTE3 data set offer an interesting insight into how well these classes of measures perform on an entailment task. We summarize the basic characteristics of the two test sets in table 4.1.

Table 4.1. Summary of two sentence pair data sets used in the experiment.

Summary	MSRP	RTE3
Major class of semantic similarity notion	Equivalence	Entailment
Number of sentence pairs	1,725	800
Number of unique words	8,256	5,700
Average sentence length (in characters)	115.30	227.87
Degree of symmetry: average difference in length between two comparing sentences (in characters)	9.68	132.81

4.5.2 Evaluation Metrics

We employ the standard definitions of recall, precision, and F1 metrics used in information retrieval and text classification evaluation to measure the effectiveness of each sentence similarity method. *Recall* is a proportion of correctly predicted similar sentences compared to all similar sentences. *Precision* is a proportion of correctly predicted similar sentences compared to all predicted similar sentences. F_1 is a uniform harmonic mean of precision and recall. A scoring threshold for positive pairs is defined at 0.5 as it is typically used in the literature (Mihalcea et al. 2006).

$$\textit{Recall} = \text{Number of correctly predicted pairs / Number of all positive pairs} \quad (4.19)$$

$$\textit{Precision} = \text{Number of correctly predicted pairs / Number of predicted positive pairs} \quad (4.20)$$

$$F_1 = (2 \times \textit{Recall} \times \textit{Precision}) / (\textit{Recall} + \textit{Precision}) \quad (4.21)$$

4.5.3 Evaluation Settings

The main objective of the experiment is to evaluate the effect of integrating sentence semantic structure in sentence similarity methods to cope with the variability of natural language expression. In particular, we investigate the effectiveness of the proposed approaches on paraphrase recognition and textual entailment tasks.

For conceptual term frequency approach, we compute CTF-weighted cosine similarity between sentence vectors with different weighting schemes. Furthermore, two lexical units, single words and multi-word collocations, are extracted as features of sentence vector. Thus, we experiment with 3 types of term weights for single-word feature vector: *CTF*, *CTFisf*, and *CTFivf*, and 3 different term weights for multi-word phrase feature vector: CTF_{phrase}, $CTFisf_{phrase}$, and $CTFivf_{phrase}$.

4.6 Results and Discussion

4.6.1 Paraphrase Recognition

Table 4.2 presents a performance comparison of the proposed sentence similarity approaches and other baseline measure. In this experiment, the best performance is achieved when *CTFisf* is employed as term weight of sentence vector (F_1 = 0.7997). However, the result does not differ significantly compared to the structural similarity or the best baseline methods. That is, F_1 scores of the structural similarity and TFIDF-identity are 0.7982 and 0.797, respectively.

Table 4.2. The performance of sentence similarity measures on paraphrase recognition task.

Measure	Recall	Precision	F_1
CTF	0.8875	0.7154	0.7922
CTFisf	**0.8806**	**0.7324**	**0.7997**
CTFivf	0.8823	0.7281	0.7978
CTF_{phrase}	0.8108	0.7405	0.7740
$CTFisf_{phrase}$	0.6905	0.7689	0.7276
$CTFivf_{phrase}$	0.7036	0.7657	0.7333
structure	0.9758	0.6753	0.7982
jaccard	0.6033	0.8347	0.7000
word-overlap	0.678	0.76	0.717
IDF-overlap	0.325	0.829	0.467
NGRAM-overlap	0.8919	0.7001	0.7848
TFIDF-cosine	0.881	0.713	0.789
TFIDF-novelty	0.283	0.858	0.426
TFIDF-identity	1	0.665	0.797
semantic-IDF	0.835	0.714	0.77

4.6.2 Textual Entailment Recognition

According to table 4.3, the best performance is attained by the proposed structural similarity measure at F_1 score of 0.6555. In contrast to the result in paraphrase recognition experiment, the use of verb-argument structure has significantly improved the performance of textual entailment recognition task over CTF-weighted cosine similarity and other baseline measures. In particular, most baseline measures, apart from semantic-IDF, perform very poorly on this task.

According to the result, we conclude that structural approach offers a greater performance gain to the similarity judgment of highly asymmetric sentences (textual entailment) than those which are more symmetric in length.

Table 4.3. The performance of sentence similarity measures on textual entailment recognition task.

Measure	Recall	Precision	F_1
CTF	0.4268	0.6341	0.5102
CTFisf	0.4000	0.6142	0.4845
CTFivf	0.4293	0.6197	0.5072
CTF_{phrase}	0.2927	0.6154	0.3967
$CTFisf_{phrase}$	0.1976	0.6045	0.2978
$CTFivf_{phrase}$	0.2171	0.6138	0.3207
structure	**0.7734**	**0.5688**	**0.6555**
jaccard	0.0512	0.6363	0.0948
word-overlap	0.032	0.565	0.06
IDF-overlap	0.007	0.6	0.014
NGRAM-overlap	0.4561	0.6493	0.5358
TFIDF-cosine	0.283	0.644	0.393
TFIDF-novelty	0.141	0.69	0.235
TFIDF-identity	0.471	0.539	0.503
semantic-IDF	0.585	0.602	0.593

4.6.3 The Impact of Semantic Role Labeler on the Overall Effectiveness

The annotation accuracy of our semantic role labeler, SENNA, is a contributing factor to the overall performance of the proposed approaches. In both sentence pair

data sets, the semantic role labeler produces 20% miss rate by which both Arg0 and Arg1 are not found in the extracted verb-argument structures. Additionally, 20% of verb-argument structures do not contain either Arg0 or Arg1. Since SENNA's model focuses exclusively on the accuracy of Arg0 and Arg1 classification, there are many cases in which semantic roles are poorly annotated for the other argument classes. Consequently, this adversely affects the precision of structural similarity approaches.

4.6.4 Shallow vs. Deep Semantic Parsing

The overall result differs from that of text categorization task (Shehata et al. 2007) where concept-based weighting has significantly improved classification performance over the traditional TFIDF scheme. One reason is that conceptual term frequency aims to capture the importance of a given concept in a document by leveraging the frequency of a concept in verb-argument structures. This approach is more compatible with text categorization mechanism in which documents are classified according to their distinct topics represented by terms or concepts in documents. On the other hand, the task of identifying semantic equivalence or entailment pairs requires a deeper semantic processing of constituents in a sentence. Deeper semantic measures are able to recognize at least the same or greater number of positive pairs according to F_1 scores than those of vector space approach. The magnitude of improvement is even more apparent in entailment task in which specific relations between constituents have to be identified.

4.6.5 Structural Approach vs. Knowledge-Based Measures

In the previous study by Mihalcea et al. (2006), knowledge-based similarity measure has been proven to be quite effective in sentence similarity task. Our experiments

also confirm their result. However, a major drawback of such approach is the lack of efficiency due to the exhaustive calculation of semantic similarity between word pairs. Therefore, they might not be as robust to employ in the real-world text mining applications as most naïve measures. In this regard, our approach offers a greater benefit over the knowledge-based measure as it greatly improves the effectiveness of naïve measures while maintaining their computational efficiency, particularly at sentence processing time.

4.7 Conclusion

In this chapter, we presented the approaches that integrate semantic structure of the sentences to handle variability of natural language expression in sentence similarity. Traditional similarity measures, which represent sentences as a bag of words, simply judge similarity between sentences according to their common word occurrences. However, Due to the complexity of many text mining applications where the similarity judgment at semantic level is expected, the performance of naïve measures are likely to degrade because of their disregard of sentence structure. Our proposed approaches aim to address the issue by computing sentence similarity at verb-argument structure level. By annotating sentences with semantic roles, we can better perform similarity calculation between semantically related components. The evaluation results confirm that the inclusion of sentence semantics significantly improves the effectiveness of sentence similarity tasks, especially on textual entailment recognition.

This chapter answers research question 1 by demonstrating that the semantic structure of sentences is helpful in improving the similarity judgment. We introduced two approaches to incorporate sentence semantic structure into the

similarity function. The first approach is based on conceptual term weighting scheme, first proposed by Shehata et al. (2007). The second approach used the structural information to deconstruct sentences into verb-argument structures. To compute the similarity score, the calculation was carried out between the corresponding semantic constituents. The overall results suggested that both approaches were more effective than most baseline measures in identifying similar sentences. In particular, the structural similarity measure significantly outperformed other measures on textual entailment recognition task.

CHAPTER 5: THE EFFECTIVENESS OF NEGATIVE ENDORSEMENTS AND SENTENCE SEMANTIC STRUCTURE ON FINDING NOVEL SENTENCES

"Research is the act of going up alleys to see if they are blind."

--Greek historian

& author of Parallel Lives

5.1 Introduction

This chapter addresses the novelty, redundancy, and diversity issues in sentence extraction tasks. Since, their notions were sometimes defined differently depending on the domains, we formally describe our definitions of novelty, redundancy, and diversity as follows. We first assume that each selected sentence needs to be topically related to a given task or information need. For example, given the task "*what effect does steroid use have on athlete's performance?",* we consider "*steroids enhance athletic performance*" to be a topically related sentence while "*steroid is an organic compound*" to be an unrelated sentence. Then, we define **novelty** as a property indicating the degree of new or novel information being expressed in one sentence relative to another selected sentence. Using the same example, "*steroids act like testosterone in building muscle mass*" has higher novelty than "*steroids help boost athlete's performance*", compared to the anchoring sentence "*steroids enhance athletic performance.*" Next, we define **redundancy** as an opposite property of novelty. That is, two selected sentences are highly redundant if they contain identical information. In the previous example, "*steroids help boost athlete's performance*" has higher redundancy than "*steroids act like testosterone in building*

muscle mass", compared to "*steroids enhance athletic performance.*" Lastly, we define **diversity** as an intrinsic property of a set or collection. That is, novelty is a unit of diversity. In other words, a high diversity set contains many distinct sentences. For example, consider two sets of sentences, A and B, below.

Set A	Set B
• Steroids help boost athletic performance by improving muscle mass. • Steroids can cause many adverse effects.	• Steroids enhance athletic performance. • Athletes use steroids to improve their performance.

According to our definitions, A is more diverse than B because sentences in A contain more distinct information than those in B. In most sentence extraction tasks, a high-diversity set is preferred as it means more distinct information are included in the extracted set. In this regard, our definitions of novelty and diversity are similar to Clarke et al.'s definitions (2008). In their retrieval evaluation framework, novelty represented the need to avoid redundancy while diversity represented the need to resolve ambiguity, which was achieved by maximizing the distinct nuggets returned in the result set.

We explore the method to promote diversity in sentence extraction tasks in this chapter. Specifically, our focus is on applying a graph-based ranking model to find novel sentences. Next, we also examine how sentence semantic structure affects the overall performance of sentence extraction tasks. Two extraction tasks are employed to evaluate the performance of the proposed method: focused summarization and question answering.

The rest of the chapter is organized as follows. Section 5.2 introduces the two research questions being tested in this chapter. Section 5.3 describes the proposed method that promotes diversity by employing the negative endorsement principle to extract highly novel sentences. Next, section 5.4 outlines the overall sentence extraction process. Then, the experimental evaluation is described in section 5.5. Section 5.6 discusses the evaluation results. Finally, the chapter is concluded in section 5.7.

5.2 Research Question Tested

This chapter focuses on answering the second and the third research questions. They are described as follows:

RQ2: What are the effectiveness of the proposed similarity measure in different application contexts? How can we incorporate the proposed similarity method into sentence extraction methods? Sentence similarity measures play a crucial role in many text mining applications, e.g., text summarization and question answering. These applications typically employ several similarity functions as part of the sentence extraction process. Most similarity functions compute the similarity scores based on co-occurrences or distributional similarity of words between two sentences. These functions include the Jaccard coefficient, cosine similarity, etc. We are interested in the effectiveness of these methods in the specific application contexts. Furthermore, we examine whether the measures that perform well in sentence similarity evaluations improve the overall performance of the sentence extraction tasks. Overall, we expect that the proposed sentence similarity measure should significantly contribute to the effectiveness of sentence extraction tasks as it has

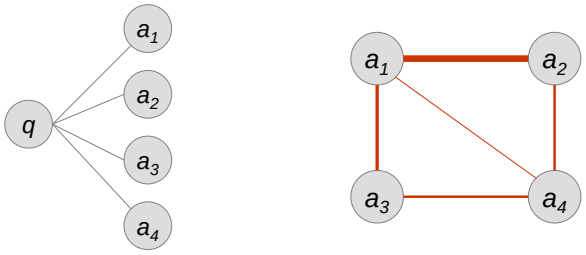
been proved in the previous chapter that it was able to deal with the problem of natural language variation quite well.

RQ3: How can we apply a graph-based ranking model to intrinsically promote diversity of a set of sentences? This question specifically focuses on intrinsic diversity of the extracted sets of sentences. That is, each sentence in a diverse set should collectively contain novel information with respect to others. To date, very few methods have considered a graphical model to promote diversity. The previous works have demonstrated the effectiveness of the graphical models, such as random walks, in finding the salient items from a sentence graph. Drawing upon research in the graphical models and diversity in ranking, we focus on answering the following questions: How can we incorporate novelty, the opposite of redundancy, into a sentence graph? How effective is the proposed graphical representation, compared to the traditional graph-based models? How effective is the proposed graphical model in focused summarization and question answering? Additionally, we expect that the proposed method should produce better results since it employs the graph-based ranking model to balance the initial relevance scores with the novelty.

5.3 The Proposed Method

In this section, we describe the proposed graph-based ranking model that focuses on finding representative and novel sentences. Specifically, it is motivated by two key attributes of a good query-focused summary. First, the focused summary should contain many relevant facts pertaining to an information need. This means representative sentences should be ranked according to their relevance score given the query. Second, the good focused summary should contain as many novel (or few redundant) facts as possible. In other words, the focused summary should be diverse

in its coverage. With the key attributes in mind, we present a simple illustration of the NegativeRank model, shown in figure 5.1, given a question q and a simple graph with four answer nodes. As displayed in the figure, two set of relations are represented by two types of edges. First, the relevance relation is denoted by the positive edges (1a), while the redundancy relation is represented by the negative edges (1b). A stronger link indicates higher relevance or redundancy. The negative sign can be interpreted as a disapproval vote between sentence nodes in contrast to a recommendation vote of the positive link. The absolute value of negative edge weight represents the degree of similarity of sentences. Thus, from figure 5.1, the relevance relation satisfies the first condition that the representative sentences should be focused to the specific information need. Next, the redundancy relation satisfies the condition that the content of the summary should be diverse. To incorporate negative edges into the ranking, we adapt the random walk over the graph structure to find a long-term negative endorsement of each sentence node.



1a. Relevance **1b. Redundancy**

Figure 5.1. An illustration of NegativeRank model.

Starting from random walk on a regular graph, we define G= (V,E) as an undirected graph where V is a set of vertices representing n sentences, E is a set of edges representing the similarity between vertices where$E \subset V \times V$. We can represent the sentence graph G as an n x n weighted matrix S where S_{ij} is a

similarity score sim(*i*,*j*) of node *i* and *j* and sim(*i*,*j*) is non-negative. If *i* and *j* are unrelated, then $S_{ij} = 0$. Given *S*, we can derive an *n* x *n* adjacency matrix A such that each element A_{ij} in *A* is the normalized value of S_{ij} such that $A_{ij} = S_{ij} / \sum_{k=1}^{n} S_{ik}$ and $\sum_{j=1}^{n} A_{ij} = 1$.

Next, given a specific query *q*, we define a vector *r* where each element r_i is the relevance score *rel*(*i*,*q*) of *i* and *q*. Then, we transform *r* into an *n* x *n* matrix *B* from the outer product of an all-1 vector and r^{T} such that each element $B_{ij} = r_i / \sum_{k=1}^{n} r_k$ and $\sum_{i=1}^{n} B_{ij} = 1$. From two probability distributions, the transition matrix *P* can be defined as:

$$P = dA + (1 - d)B \tag{5.1}$$

where *d* is a damping factor with a real value from [0,1], *A* is the initial adjacency matrix, *B* is the query-sentence relevance matrix. Since all rows in *P* have non-zero probabilities which add up to 1, *P* is a stochastic matrix where each element P_{ij} corresponds to the transition probability from state *i* to *j* in the Markov chain. Thus, *P* satisfies ergodicity properties and has a unique stationary distribution $\vec{\pi}P = \vec{\pi}$. Notice that equation 3.1 is the random walk over regular sentence graph which does not address the redundancy issue.

To exploit the negative edges for redundancy reduction problem, we modify the sentence graph *G* such that all edge weights in *G* have a negative sign. As such, we define $G^{-} = (V,E^{-})$ as an undirected graph where *V* is a set of *n* sentence vertices, E^{-} is a set of negative edges where $E^{-} \subset V \times V$. Intuitively, the negative edges in G^{-} represent the penalty of redundancy between nodes. Then, an adjacency matrix *M,*

corresponding to edge weights in G^-, is defined as an all-negative matrix of S where $M_{ij} = -S_{ij} \sum_{j=1}^{n} S_{ij}$ and $\sum_{i=1}^{n} M_{ij} = -1$.

Next, we define a new transition matrix Q to incorporate the negative edges. To ensure that Q is still ergodic, we multiply matrix B with a scaling factor c. The value of c is determined by the conditions that all elements in Q should be non-negative and each i-th row of Q should add up to 1. That is, $\sum_{j=1}^{n} Q_{ij} = 1$. Since all rows of M sum to -1 and all rows of B add up to 1, c is a function of d where $c = 1 + d/1 - d$.

$$Q = dM + (1 - d)cB \tag{5.2}$$

where M is an all-negative adjacency matrix. Since ergodicity properties still hold, the modified transition matrix Q has a unique stationary distribution $\vec{\pi}Q = \vec{\pi}$. Finally, we rank each node i according to its stationary probability $\vec{\pi}_i$. Following the matrix notation, the simplified NegativeRank equation can be written as follow:

$$DR^t(i) = (1+d) \cdot \frac{rel(i,q)}{\sum_{i=1}^{n} rel(i,q)} - d \sum_{j \in adj(i)} \frac{sim(i,j)}{\sum_{k=1}^{n} sim(j,k)} DR^{t-1}(j) \tag{5.3}$$

Where d is a damping factor with a real value in [0,1] range. Additionally, d serves as a penalty factor of redundancy. *rel*(*i*,*q*) is the relevance score of a sentence i given a query q. And sim(*j*,*i*) is a similarity score of sentence j and i.

To estimate the value of *rel*(*i*,*q*), we employ a sentence weighting function described in Allen et al. (2003) as it is shown to consistently outperform other relevance models at the sentence level. It defines the relevance score of sentence s given query q as a dot product between TFISF (term frequency times inverse sentence frequency) sentence vector and TF-weighted query vector.

$$R(s|q) = \sum_{t \in q} \log(tf_{t,q} + 1) \log(tf_{t,s} + 1) \log(\frac{n+1}{0.5 + sf_t}) \quad (5.4)$$

where $tf_{t,q}$ and $tf_{t,s}$ are the number of times term t appears in the query and sentence, respectively, sf_t is the number of sentences in which terms t occurs, and n is the number of sentences being scored.

In the case where relevance score is ignored, e.g. generic summarization, we simply assign the uniform distribution $1/n$ to all nodes. Thus, equation 5.5 specifies the novelty-only variant of NegativeRank.

$$DR^t(i) = \frac{(1+d)}{n} - d \sum_{j \in adj(i)} \frac{sim(i,j)}{\sum_{k=1}^{n} sim(j,k)} DR^{t-1}(j) \quad (5.5)$$

We expect this variant to perform worse in the focused summarization task than an inclusive approach which takes both relevance and novelty into consideration.

Integration with other methods. Apart from the relevance function $R(s \mid q)$, other methods can be used to supply the alternative initial ranking distribution, e.g., topic model (Blei et al. 2003), topic-sensitive graph centrality (Otterbacher et al. 2005), query-likelihood language model, etc. In addition, we can adapt NegativeRank to generic summarization by replacing the relevance function with other saliency functions, e.g. word probability(Nenkova and Vanderwende 2005), lead-based scoring (Brandow et al. 1995), and graph centrality (Erkan and Radev 2004; Mihalcea and Tarau 2004).

Convergence. To determine a stopping point of NegativeRank iteration, we find rank convergence using Kendall tau distance as it has been proved that rank

convergence tends to reach its saturation at certain point while L_1 convergence improves monotonically (Berkhin 2005). Therefore, it takes lesser time to find the stopping point through rank convergence. Kendall tau measures the dissimilarity between ranking order at t iteration and t-1 iteration. Suppose $\sigma(i)$ denotes the rank of sentence i, Kendall tau K distance between the two ranking orders σ^t and σ^{t-1} is defined as follow:

$$K(\sigma^t, \sigma^{t-1}) = \frac{\#\{(i<j)|sgn((\sigma^t(i)-\sigma^t(j))/(\sigma^{t-1}(i)-\sigma^{t-1}(j))=-1\}}{n(n-1)/2} \quad (5.6)$$

According to the above equation, K is defined as the number of discordant pairs normalized into a range [0,1]. If two ranking orders are identical, K=0. In contrast, if they are completely different, K=1. In this work, we set K-threshold to 0.1. At the end of t-th iteration, if $K < 0.1$, we choose t as the stopping point.

5.4 Sentence Extraction Process

In order to generate the focused summaries, we employ the two-stage architecture as shown in figure 5.2. The first stage involves preprocessing and retrieving the relevant sentences for a given query topic from a document collection. Then, we perform sentence re-ranking by applying the ranking algorithms in the second stage.

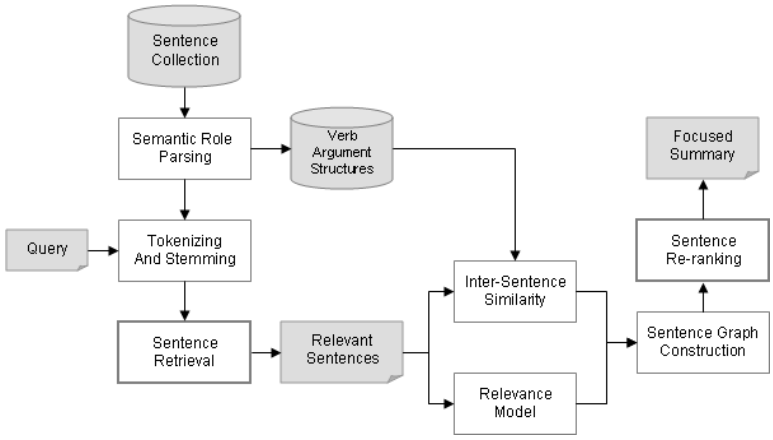


Figure 5.2. The overall two-stage extraction process.

5.4.1 Sentence Retrieval

Starting from the preprocessing step, we first assign a semantic role to each sentence constituent using a semantic role labeler (Collobert and Weston 2007). Next, we derive a set of verb-argument structures for each sentence based on the semantic role information. Then, we extract word-level features from the sentence collection by tokenizing sentences into single words, removing non content-bearing words, e.g., articles, conjunctions, prepositions, etc., and stemming the tokens using Porter Stemmer.

After preprocessing step, we use a vector-space model to retrieve the relevant sentences. Free-form narrative field associated with each topic/question is used as a query. The relevance score between the sentence and query is derived from a cosine similarity between *concept-based weighted vectors* (henceforth *CTF*-weighted vectors) of a sentence and CTF-weighted vector of a given query. We construct conceptual term-sentence matrix. Single-word tokens are used as the conceptual term features. Next, CTF_i weight is computed for each conceptual term feature i. Finally, the relevance score of a sentence is calculated from a cosine similarity

between *CTF*-weighted sentence vector and *CTF*-weighted query vector. A list of top-500 relevant sentences is selected from the retrieved set.

5.4.2 Sentence Re-ranking

The next step is to re-rank the list of relevant sentences, obtained from the previous stage, using the sentence ranking model. First, we represent the list of relevant sentences as an undirected graph with negative edges. Different edge weighting schemes based on inter-sentence similarity measures are considered. In a case of sentence semantic structure similarity, a set of verb-argument structures of sentences are used as an additional input. The relevance models of the retrieved sentences and a query are formulated. The relevance sub-graph is represented by the positive edge between the query node and sentence nodes. After the ranking scores are calculated, the top-k sentences with a cut-off level of 250 words are selected for focused summary experiment. Next, for question answering experiment, the default cut-off level for the answer set is 7,000 characters.

5.5 Experimental Evaluation

The goal of the experiment is to evaluate the contributions of the proposed ranking model and sentence similarity measure in promoting diversity in two sentence extraction tasks: focused summary and question answering. The comparison is done with regard to several well-known baseline methods.

5.5.1 Data Sets

Focused Summarization Data Sets. We conduct a query-focused summarization evaluation using the DUC 2006 (DUC06) and DUC 2007 (DUC07) data sets. These publicly-available data sets are prepared by human experts at

National Institute of Standards and Technology (NIST) to be used in Document Understanding Conferences for evaluating document summarization systems. Each data set comprises a set of topics (50 topics for DUC06 and 45 topics for DUC07), a set of 25 relevant news articles, and a set of human-extracted summaries for each topic to be used as the reference. Each topic contains title and a brief narrative. The main task is to generate a 250-word summary corresponding to each summary topic description.

Question Answering Data Sets. Two question answering data sets are used in the evaluation: a subset of Yahoo! Answers data set (YahooQA) used in Liu et al.'s work (2008) and a complex interactive question answering test set (ciQA) used in TREC 2006 (Kelly and Lin 2007). YahooQA data comprises subjective and ill-defined information needs formulated by the community members. The subjects of interests span widely from mathematics, general health, to wrestling. In contrast, ciQA data largely focus on the complex entity-relationship questions. Their information needs reflect those posed by intelligence analysts. From data quality perspective, YahooQA data are much noisier than ciQA data as they contain mostly informal linguistic expressions.

To prepare YahooQA data set, we randomly select 100 questions and 10,546 answers from the top 20 most frequent categories (measured in terms of a number of responded answers) to use as a test set. A set of information nuggets for YahooQA is automatically created by matching relevant answers with the corresponding questions. The best answer chosen by askers for each question is marked as a vital nugget while the other answers are marked as an okay nugget. In the case of ciQA data set, 30 question topics and their free-form description are prepared by human assessors at NIST. Documents containing relevant answers are selected from the

AQUAINT corpus – the standard text collection consisting of newswire text data from the Xinhua News Services, the New York Times News Services, and the Associated Press News Services. Moreover, NIST assessors also create the benchmark nuggets for each question.

Table 5.1 summarizes the two focused summarization data sets while table 5.2 summarizes the two question answering data sets used in the experiments.

Table 5.1. Summary of focused summarization data sets

Summary	DUC06	DUC07
Number of topics	50	45
Number of relevant documents per topic	25	25
Number of reference summaries per topic	10	10
Number of candidate sentences per topic	680	527.62
Avg. sentence length (in words)	22.16	22.20

Table 5.2. Summary of question answering data sets

Summary	ciQA	YahooQA
Number of questions	30	100
Number of total candidate answers	69,626	10,546
Average answer sentence length (in characters)	144.53	295.67
Avg. nuggets per question	16	10

5.5.2 Evaluation Metrics

Focused Summarization. We adopt three evaluation metrics normally employed in document summarization evaluation. These are ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4). Basically, ROUGE score is computed from a lexical n-gram recall between system-extracted summaries and human-constructed reference summaries.

$$ROUGE - N = \frac{\sum_{S \in \{ref\}} \sum_{gram_n} Count_{match}(gram_n)}{\sum_{S \in \{ref\}} \sum_{gram_n} Count_{ground}(gram_n)} \tag{5.7}$$

Where n is the length of the n-gram, $gram_n$ and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, and $Count(gram_n)$ is the number of n-grams in the reference summaries. Based on this definition, R-2 is computed for word bigram and R-SU4 is computed for skip-4 bigram.

Question Answering. To assess the performance of the proposed method, we employ the nugget pyramid procedures to evaluate the quality of the extracted set of answers. It has been shown that the pyramid scores correlate well with human judgments and can be used as a proxy for manual evaluation (Lin and Demner-Fushman 2005). Generally, we assume that the factual diversity of the extracted answers can be measured in terms of a number of information nuggets the extracted sets have in common with the benchmark nuggets. Information nugget is a small text fragment that describes a certain fact about a given question. In an automatic evaluation setting, a benchmark set of information nuggets has to be prepared beforehand by human experts. Each nugget can be categorized into two binary

classes: *vital* and *okay*. Vital nuggets are those that must be contained in a good answer while okay nuggets are useful but not essential information and has no adverse effect on the overall score.

The formulas to compute the pyramid F-score are described in Lin and Demner-Fushman (2005). In summary, the pyramid F-score is computed as a weighted harmonic mean (F-score) between nugget recall (NR) and nugget precision (NP). NR and NP are derived from summing the unigram co-occurrences between terms in each information nugget and terms from each extracted answer set. NR is computed on vital nuggets only while NP is estimated from a length allowance based on the number of both vital and okay nuggets returned. This is done to penalize verbosity in NP approximation.

$$NR = \frac{r}{R} \tag{5.8}$$

$$\alpha = 100 \times (r + a) \tag{5.9}$$

$$NP = \begin{cases} 1 & \text{if } l < \alpha \\ 1 - \frac{l-\alpha}{l} & \text{otherwise} \end{cases} \tag{5.10}$$

$$F_\beta = \frac{(\beta^2 + 1) \times NR \times NP}{\beta^2 \times NP + NR} \tag{5.11}$$

where r is a number of vital nuggets returned in a system response, R is a number of vital nuggets in the answer key, a is a number of okay nuggets returned in a system response, and l is a number of non-whitespace characters in the entire answer string. Following the standard procedure in TREC 2006, we set the evaluation parameters to $\beta = 3$ and l = 7,000 and use Pourpre (Lin and Demner-Fushman 2005) script version 1.1c to automatically compute the scores.

Additionally, we also perform additional analysis to measure the performance of each method at varying answer lengths via recall-by-length performance curve (Lin 2007). The curve offers a more fine-grained analysis of system performance by quantifying the linear rate in which a particular system outputs relevant nuggets. Ideally, better systems will produce curves that rise faster. To plot the recall-by-length curve, we increment all lengths by a hundred characters, from 100, 200, and so on. Then, recall is averaged across all questions at each length increment.

5.5.3 Methods to Compare

We compare the effectiveness of several baseline methods and NegativeRank variants in selecting representative sentences. They are described as follows:

Maximal marginal relevance (MMR). A redundancy reduction technique commonly used in information retrieval and text summarization (Carbonell and Goldstein 1998). It consists of two major components. The relevance component of each sentence is calculated as cosine similarity between the query and sentences while the redundancy component employs cosine similarity between each relevant sentence and the selected sentences in the summary.

SumBasic. A sentence scoring method based on the probability of words in a document collection. We made a slight modification to the formula in (Nenkova and Vanderwende 2005) to integrate query terms into sentence scoring. Each sentence is scored by the sum of the average probability of the words described in equation 5.12. After the best scoring sentence is chosen, the probability of each word is updated according to equation 5.13. This ensures that term redundancy is being penalized after each representative sentence is selected.

$$Weight(S_j|q) = \sum_{w_i \in \{S_j \cap q\}} \frac{p(w_i)}{|\{w_k|w_k \in S_j\}|} \tag{5.12}$$

$$p_k(w_i) = p_{k-1}(w_i)^2 \tag{5.13}$$

Topic-Sensitive LexRank. An eigenvector-centrality ranking model based on the random walk over sentence graph (Otterbacher et al. 2005). It extends the generic LexRank method by defining the mixture model between sentence *s* and a summary topic *q* as the sum of its relevance to the topic query and the TFIDF-weighted cosine similarity to the other sentence.

$$p(s|q) = d\frac{rel(s|q)}{\sum_{z \in C} rel(z|q)} + (1-d)\sum_{v \in C} \frac{sim(s,v)}{\sum_{z \in C} sim(z,v)} p(v|q) \tag{5.14}$$

where *C* is the set of all sentences in the cluster and *d* is a trade-off between the influence of the relevance component and inter-sentence similarity component. In this work, we use the optimal values of $d = 0.9$ and sentence similarity threshold = 0.2 obtained from (Otterbacher et al. 2005). Since the representative sentences are ranked via graph-based centrality, the novelty of sentences is accounted for by the principal of information subsumption between nodes. We anticipate this method to be a competitive baseline.

Inverse LexRank. We specifically define this method as a comparison to the key idea in the proposed method. Hypothetically, if diversity can be improved by the negative-edge penalty as defined in NegativeRank, a simple backward ranking of LexRank scores (inverse LexRank) should be as effective as NegativeRank in generating a diversified summary. To test the hypothesis, we run the topic-sensitive LexRank algorithm to find the stationary distribution for each sentence node.

However, representative sentences are ranked in ascending order according to its stationary distribution instead of descending order.

Next, several NegativeRank variants are defined based on the combinations of the initial ranking distribution: SumBasic (*SB*), TFISF-weighted relevance function (*REL*), and a uniform distribution 1/*n*, and inter-sentence similarity measure: sentence semantic similarity(*SS*) and TFISF-weighted cosine similarity (*TFISF*). The summary of NegativeRank variants is shown in table 5.3.

NegativeRank variants. Different variants are defined based on the initial ranking distributions and sentence similarity measures. We also test two novelty-only variants, abbreviated as 1+SS and 1+TFISF, which do not assign specific the initial ranking distributions to the sentences. That is, all sentences are assigned a uniform distribution 1/n instead. We expect the performance of the novelty-only variants to be poor since they ignore topicality when extracting sentences.

In focused summarization experiment, the top-k representative sentences are selected as the summary. The summary length is cut off at 250 words. Next, for question answering experiment, we select top-*k* answer sentences to form an answer set for each question. The default cut-off level for the answer set is 7,000 characters.

Table 5.3. Summary of the variants

Abbreviation	Initial Ranking Distribution	Inter-Sentence Similarity
SB+SS	SumBasic	Sentence-level structural similarity
SB+TFISF	SumBasic	TFISF-weighted cosine similarity
REL+SS	R($s \mid q$)	Sentence-level structural similarity
REL+TFISF	R($s \mid q$)	TFIDF-weighted cosine similarity
1+SS	$1/n$	Sentence-level structural similarity
1+TFISF	$1/n$	TFIDF-weighted cosine similarity

5.5.4 Parameter Tuning

We estimate the parameters of the NegativeRank model on DUC06's task 1 through 5 (10% of DUC06 tasks). The training set contains 125 documents and approximately 3,400 sentences. The optimal parameter settings for NegativeRank are d = 0.8 and c = 9. The thresholds for inter-sentence similarity score for *SS* and *TFISF* are set to 0.4 and 0.2, respectively.

5.6 Results and Discussion

5.6.1 Focused Summarization Experiment

The performance of various NegativeRank variants on DUC06 and DUC07 are shown in table 5.4. Here, variants which employ the sentence semantic similarity (SS) as an inter-sentence similarity measure performs significantly better than those which employ the cosine similarity, $p<0.05$. For instance, in DUC06 case, SB+SS performs 16.83% and 8.68% better than SB+TFISF on R-2 and R-SU4,

respectively. Next, in DUC07 case, SB+SS also significantly outperforms SB+TFISF by 9.58% and 5.34% on R-2 and R-SU4, respectively. Similarly, REL+SS and 1+SS also outperform their counterparts across all metrics and data sets. Overall, REL+SS is the best variant on both DUC06 and DUC07. This confirms our expectation that the sentence semantic structure is helpful in handling the variability of natural language expression. Particularly, a sentence-level text segment is more sensitive to this problem because it expresses a more specific meaning. The slight changes in a composition of sentence might result in two different meanings. As a result, the application of sentence semantic structure in edge weighting provides a significant contribution to redundancy reduction among nodes in the sentence graph.

Table 5.4. The average R-2 and R-SU4 scores of the NegativeRank variants. The best results are in bold.

Variant	DUC06		DUC07	
	R-2	R-SU4	R-2	R-SU4
SB+SS	0.0729	0.1302	0.0904	0.1441
SB+TFISF	0.0624	0.1198	0.0825	0.1368
REL+SS	**0.0789**	**0.1341**	**0.1017**	**0.1535**
REL+TFISF	0.0781	0.1336	0.0973	0.1533
1+SS	0.0728	0.1298	0.0950	0.1500
1+TFISF	0.0677	0.1240	0.0883	0.1413

Table 5.5. The comparison between variants with different sentence similarity measure.

Variant	DUC06		DUC07	
	R-2	R-SU4	R-2	R-SU4
SB+SS vs. SB+TFISF	+16.83%	+8.68%	+9.58%	+5.34%
REL+SS vs. REL+TFISF	+1.02%	+0.37%	+4.52%	+0.13%
1+SS vs. 1+TFISF	+7.53%	+4.68%	+7.59%	+6.16%

Next, the results also confirm our initial expectation that the methods which select the sentences based on the balance between relevance and novelty should produce a better focused summary than the novelty-centric counterpart. By supplying the relevance scores as the initial ranking probabilities, the sentence ranking model performs the best. For example, the performance of the best NegativeRank variant *REL+SS* are significantly higher than those of *1+SS*, $p<0.05$.

Next, we compare the performance of the best NegativeRank variant with respect to that of the baseline methods. Table 5.6 displays the average R-2 and R-SU4 the baselines methods and the best NegativeRank variant on DUC06 and DUC07 data sets. Overall, the best NegativeRank variant outperforms most baselines on most evaluation metrics. First, when DUC06 is the test set, NegativeRank significantly outperforms MMR, by 4.23% on R-2 and 2.52% on R-SU4, $p<0.05$. Next, when DUC07 is the test set, NegativeRank performs 11.15% and 8.10% better than MMR on R-2 and R-SU4, respectively. Moreover, it also outperforms SumBasic across both data sets. However, when comparing with LexRank, NegativeRank performs slightly better but the differences are not statistically significant. In addition, inverse LexRank produces significantly inferior

R-2 and R-SU4 scores than NegativeRank despite the similar key ranking idea. This suggests that our proposed method is not merely a backward ranking of the regular graph centrality.

Table 5.6. The average R-2 and R-SU4 scores of the baseline and NegativeRank methods. The best results are in bold.

Baseline Method	DUC06		DUC07	
	R-2	R-SU4	R-2	R-SU4
Human Average	*0.1125*	*0.1710*	*0.1410*	*0.1916*
MMR	0.0757	0.1308	0.0915	0.1420
SumBasic	0.0659	0.1225	0.0852	0.1389
LexRank	0.0785	**0.1394**	0.0967	0.1528
LexRankInv	0.0555	0.1126	0.0699	0.1260
NegativeRank	**0.0789**	0.1341	**0.1017**	**0.1535**

Table 5.7. The performance differences of NegativeRank compared to the baseline methods

Baseline Method	DUC06		DUC07	
	R-2	R-SU4	R-2	R-SU4
MMR	+4.23%	+2.52%	+11.15%	+8.10%
SumBasic	+19.73%	+9.47%	+19.37%	+10.51%
LexRank	+0.51%	-3.80%	+5.17%	+0.46%
LexRankInv	+42.16%	+19.09%	+45.49%	+21.83%

The focused summaries obtained from NegativeRank, LexRank, and inverse LexRank shown in figure 5.3 suggest the effectiveness of our method. Task#D0706B requires the summary to focus on the main events and important personalities in Myanmar surrounding the government changed in 1988. The reference summary created by the human expert contains six unique facts. In this instance, the focused summary obtained from NegativeRank only misses one fact while summary generated by LexRank misses two facts. Moreover, the first two sentences in LexRank summary are redundant while summary obtained from inverse LexRank does not contain any relevant facts.

Reference:

- Myanmar has been ruled by the military in various guises since 1962.
- After crushing a nationwide democracy movement, the State Law and Order Restoration Council took over Burma in 1988 and changed its name to Myanmar.
- Nobel laureate Aung San Suu Kyi heads the popular opposition political party, the National League for Democracy.
- The military government has maintained a campaign to harass and imprison her and members of the NLD, anti-government ethnic armed groups, and student organizations.
- The government has used forced labor and torture in its war against stubborn resistance by ethnic minorities.
- Myanmar has demanded that the Thai government strictly control refugee camps on the Thai side of the border between the two countries.

NegativeRank:

- He said there are 24 refugee camps along the Myanmar-Thai border where members and their families of different anti-Myanmar government armed groups such as the All Burma Students' Democratic Front (ABSDF), Kayin National Union (KNU) and Democratic Alliance of Burma (DAB) are living and conducting military and "terrorist" training there involving foreigners.
- Suu Kyi won the Nobel Peace Prize in 1991 for her peaceful struggle for democracy against the military regime in Myanmar, also known as Burma.
- There are 42 NLD members of parliament in Myanmar's prisons, according to the All Burma Students Democratic Front, an exile group.
- The vice chairman of Myanmar opposition leader Aung San Suu Kyi's political party was threatened with arrest in a commentary in a government-run newspaper Sunday.

LexRank:

- The military has ruled Myanmar, also known as Burma, since 1962.
- Myanmar, also known as Burma, has been ruled by the military since 1962.
- The current military government came to power on Sept. 18, 1988 after brutally crushing a nationwide democracy movement.
- Suu Kyi won the Nobel Peace Prize in 1991 for her peaceful struggle for democracy against the military regime in Myanmar, also known as Burma.

Inverse LexRank:

- A high-ranking Myanmar military official said Sunday that the authorities made timely arrest of 40 persons in January, who allegedly attempted to commit terrorist acts in the country.
- Citing her personal physicians, who have visited her twice in her van outside Yangon, her eyes are turning yellow and she has low blood pressure, the party statement said.
- On Thursday, the Burma Lawyers Council, composed of exiles, called on the country's lawyers to endorse the convening of parliament.
- In the spirit of this philosophy, I present today in my capacity as chairman of the billion-dollar multinational Make a Buck at Any Cost Corp. my special report on American Business Sentiment toward Burma.

Figure 5.3. Examples of the summaries for DUC07's task #D0706B.

5.6.2 Question Answering Experiment

Table 5.8. The average F-Scores of the NegativeRank variants. The best results are in bold.

Method	YahooQA	ciQA
SB+SS	0.3094	0.3542
SB+TFISF	0.2725	0.3454
REL+SS	**0.3353**	**0.3746**
REL+TFISF	0.2501	0.3686
1+SS	0.2913	0.3471
1+TFISF	0.2740	0.3439

Table 5.9. The comparison between variants with different sentence similarity measure.

Method	YahooQA	ciQA
SB+SS vs. SB+TFISF	+13.54%	+2.55%
REL+SS vs. REL+TFISF	+34.07%	+1.63%
1+SS vs. 1+TFISF	+6.31%	+0.93%

Table 5.8 displays the performance of NegativeRank variants according to the combinations of the initial ranking distribution and sentence similarity measures. The best pyramid F-scores among variants on YahooQA and ciQA data sets are 0.3353 and 0.3746, respectively. In both data sets, the best performance is achieved by employing TFISF-based relevance function (REL) as the initial ranking distribution and the structural similarity measure (SS) as the inter-sentence similarity function. Furthermore, the result confirms our expectation that the use of sentence similarity improves the overall performance across all variants.

Table 5.10 shows the average pyramid F-scores of the baseline methods and the best NegativeRank variant. In both data sets, the proposed method significantly outperforms all baseline methods, $p<0.05$. When YahooQA is the test set, NegativeRank significantly outperforms MMR by 13.82%, $p<0.05$. Next, when ciQA is the test set, NegativeRank performs 50.68% better than MMR. Moreover, it also outperforms SumBasic by 15.82% and 26.73% on YahooQA and ciQA, respectively. Considering the performance between random-walk based methods (LexRank vs. NegativeRank), NegativeRank also outperforms LexRank in both data sets although the improvements are relatively minor (6.01% and 4.35%), compared to those of other baselines. Furthermore, inverse LexRank produces inferior scores to

NegativeRank in both data sets. This result is consistent with the one in the focused summarization experiment.

Table 5.10. The average F-Scores of the baseline and NegativeRank methods. The best results are in bold.

Method	YahooQA		ciQA	
	F-Score	% change compared to NegativeRank	F-Score	% change compared to NegativeRank
MMR	0.2946	+13.82%	0.2486	+50.68%
SumBasic	0.2895	+15.82%	0.2956	+26.73%
LexRank	0.3163	+6.01%	0.3590	+4.35%
LexRankInv	0.2391	+40.23%	0.3516	+6.54%
NegativeRank	**0.3353**	-	**0.3746**	-

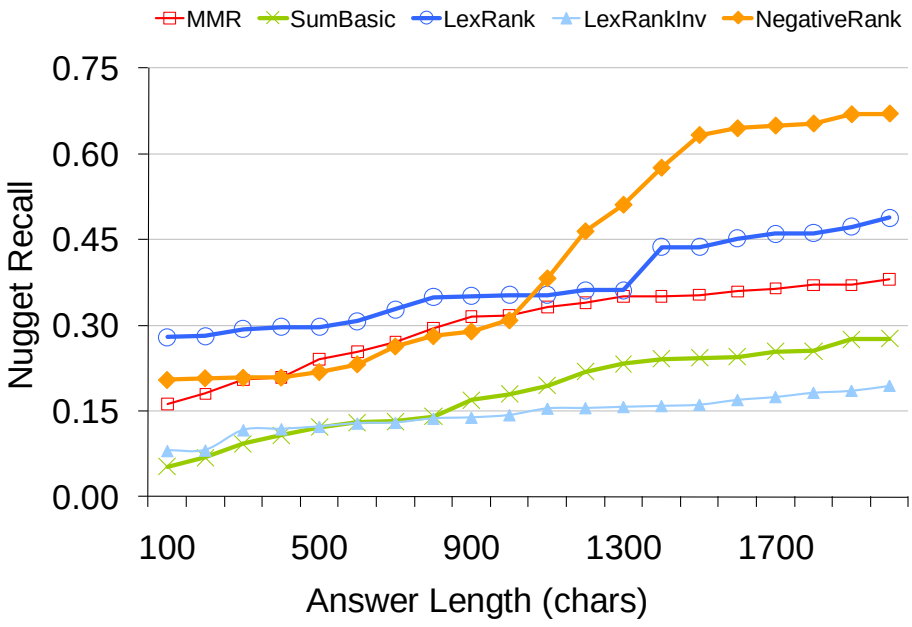


Figure 5.4. The recall-by-length performance curves on YahooQA data set.

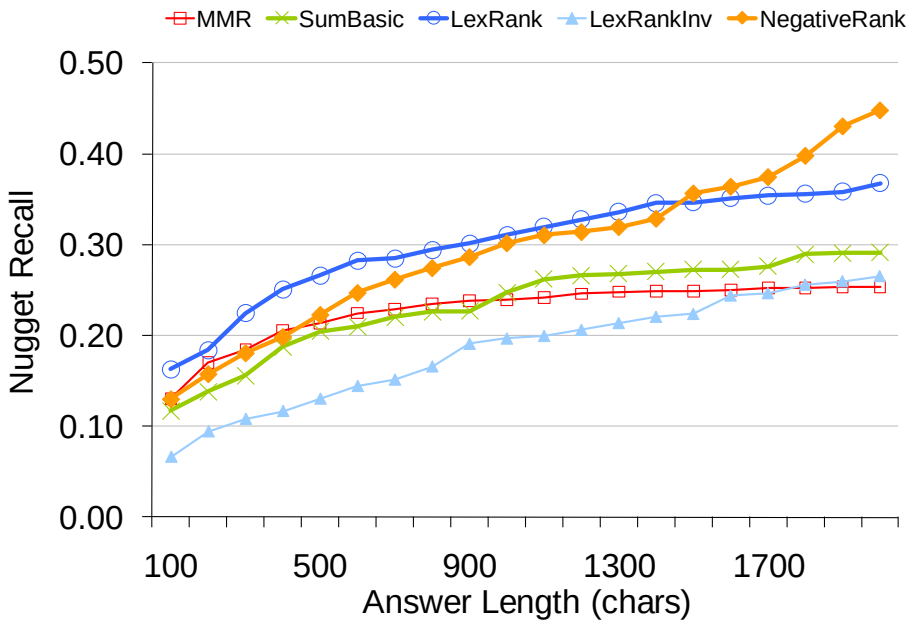


Figure 5.5. The recall-by-length performance on ciQA data set.

Figure 5.4 shows recall-by-length curves of the baselines and the proposed method in each data set. In YahooQA case illustrated in figure 2a, NegativeRank starts to perform significantly better than LexRank, $p<0.05$, which is the best baseline method, at the length increment of 1,000 characters. However, the proposed method does not perform quite well in the ciQA case. As shown in figure 5.5, NegativeRank does not outperform LexRank until after the incremental length of 1,500 characters. As a smaller answer set contains a fewer number of information nuggets, therefore there are fewer items to be diversified. As the answer set continues to grow, NegativeRank eventually outperforms LexRank. Moreover, LexRank has been shown to perform well in the sentence retrieval of online news articles (Otterbacher et al. 2005) whose contents share similar characteristics with ciQA data. This further explains a greater gap between NegativeRank's performance

and LexRank's performance on ciQA data set with respect to YahooQA. Furthermore, the size of the answer set cannot be known a priori in most cases. Thus the effectiveness of answer ranking methods tends to vary by the nature of questions. Note that our ciQA experimental results illustrate the similar trend to that of TREC 2006 ciQA task (Kelly and Lin 2007) by which a method that produces the best F-score at a predefined answer length does not necessarily perform effectively across all incremental lengths.

5.7 Conclusion

In this chapter, we examined the effects the proposed graph-based ranking model and the structural similarity measure on finding the novel sentences . The proposed ranking model employed random walks over a negative-edge graph to promote diversity by lowering redundant sentences' rank. To mitigate the problem of natural language variation, we utilized the structural similarity measure to weigh edges during the construction of a sentence graph. To evaluate their effectiveness in various sentence extraction contexts, we performed a comprehensive evaluation on two sets of experiments, focused summarization and question answering. The focused summarization experiment was tested on Document Understanding Conferences data sets while the question answering experiment was performed on Yahoo! Answers and TREC 2006's complex question answering data sets. The results showed that both the negative-edge random walk and the structural similarity measure significantly improved the results of focused summarization and question answering tasks, compared to most baseline methods at $p<0.05$.

This chapter answers research question 2 by demonstrating that the proposed similarity measure, which computes the similarity at sentence structure

level, significantly improved the effectiveness of both sentence extraction tasks being tested: focused summarization and question answering, $p<0.05$. Since the proposed sentence similarity measure was more effective than the baseline similarity measure in dealing with natural language variations, it was more effective in representing the relations between vertices in sentence graph representation. By employing the structural similarity measure as the edge weighting function in the proposed graph-based ranking model, the performance of the sentence extraction method improved significantly, compared to those which employed the baseline TFIDF-weighted cosine similarity measure.

This chapter answers research question 3 by introducing the notion of negative endorsements and applying it to model the redundancy relation in a sentence graph. As a result, the proposed NegativeRank model promoted diversity by lowering the redundancy sentences' rank based on their long-term negative endorsements. According to the experimental results, NegativeRank significantly outperformed the traditional graph-based models, e.g. topic-sensitive LexRank, in most focused summarization and question answering tasks, $p<0.05$.

CHAPTER 6: TOWARD A UNIFIED MODEL OF CENTRALITY AND DIVERSITY RANKING FOR SENTENCE EXTRACTION

"That's the nature of research--you don't know what in hell you're doing."

--Papa Flash

6.1 Introduction

This chapter further investigates the issue of diversity in ranking for sentence extraction. In the early days of information retrieval research, Boyce (1964) and Goffman (1982) identified the importance of diversity in document retrieval. The basic idea is that the relevance of a document in the retrieved set is dependent on the other retrieved documents. Several research has been done to address the diversity issue. The earliest and the most well-known work is Maximal Marginal Relevance (MMR) (Carbonell and Goldstein 1998) in which diversity is achieved by minimizing redundancy between the retrieved documents. Subsequent works considered query disambiguation as an objective of diversity. Zhai et al. (2003; 2006) suggested that a query typically contains more than one interpretation. They proposed a subtopic retrieval methods based on a risk minimization framework. Similar to Zhai et al.'s works, Chen and Karger (2006) presented a probabilistic retrieval method that incorporates negative feedback from irrelevant documents to maximize diversity. Agrawal et al. (2009) proposed a taxonomy-based classification for query and documents. They modeled user intents as topics and diversified the search results according to an objective function for minimizing user dissatisfaction. Similarly, Carterette and Chandra (2009) defined the probabilistic models of novel document rankings based on faceted topic retrieval. They assumed that information

needs are multi-facet and the goal of the faceted retrieval model was to maximize facet coverage with the smallest retrieved set.

In extractive summarization, diversity is also considered a key requirement of an effective summarization method. Zhu et al. (2007) suggested that a good sentence should be representative such that it reflects one of the core meanings of a document. In addition, the top sentences should be diverse collectively. They incorporated centrality, diversity, and prior ranking distributions into a unified framework of absorbing Mark chain random walk. Next, Li et al. (2009) defined diversity, coverage, and balance as three important aspects of extractive summary. Diversity is enforced by reducing redundancy among the sentences. Coverage focuses on minimizing the information loss. Lastly, balance emphasizes on giving an equal importance of different aspects of the document in the summary. Arguably, diversity was also implicitly encouraged in a random walk summarization model proposed by Erkan and Radev (2004). Since their method was based on the cross-sentence information subsumption principle, the representative sentences were ideally selected from the distinct centers of the sentence graph.

While much research has addressed diversity in ranking from various approaches, they did not consider centrality and diversity together. Generally, diversity ranking was performed at a post-processing stage. Recently, a few methods which explicitly focus on diversity promotion have been proposed. In this chapter, we explore the effectiveness of the unified models with respect to other diversity ranking methods. These methods represented various ranking principles.

The outline of this chapter is described as follows. First, the research question being tested in this chapter is described in section 6.2. Next, we propose the unified centrality and diversity ranking model based on the negative state random walk

principle in section 6.3. Then, we describe the experimental evaluation in section 6.4 and discuss the results in section 6.5. Finally, we conclude the chapter in section 6.7.

6.2 Research Question Tested

This chapter focuses on answering the fourth research question described as follows:

RQ4: What is the best way to incorporate diversity ranking into the graph-based ranking model while retaining the advantage of centrality ranking? How effective is the proposed diversity ranking model, compared to the similar state-of-the-art methods? Specifically, we investigate the performance of the graph-based ranking models which consider centrality ranking and diversity ranking in one unified process. To that end, we will conduct a comprehensive evaluation on the standard focused summarization and question answering tasks. In particular, we focus on answering the following questions: What are the performance improvements, if any, of the unified centrality and diversity ranking models, compared to the models that consider diversity implicitly? What is the effectiveness of different diversity ranking principles in extracting a diverse set of sentences? What performance metrics should be used to evaluate the diversity of the sets of sentences? What are the agreements among different evaluation metrics? We anticipate that the unified centrality and diversity ranking principles should be highly effective in extracting the diverse sets of sentences, compared to other diversity ranking principles.

6.3 The Proposed Method

The basic NegativeRank model introduced in chapter 5 promotes the diversity of an extracted set by iteratively lowering the prior distributions or the ranking scores of sentences by their negative endorsements. However, in many application contexts,

the model may not performs adequately as a standalone ranking model as it focuses solely on redundancy reduction. That is, its behavior can be best regarded as a re-ranking model or redundancy reduction method. Our goal is to extend the NegativeRank model by combining centrality ranking and diversity ranking into one unified process. We propose Multi-stage NegativeRank, an eigenvector centrality and diversity ranking model that encourages diversity through negative endorsements. Our key idea is similar to the GRASSHOPPER model (Zhu et al. 2007) in that the extended model requires multiple stages or iterations to rank all sentences. In each iteration, the sentence with the largest stationary probability is selected. This stage reflects centrality ranking employed by the typical random walk. Next, the key to promote diversity or increase the novelty of the selected item is to heuristically modify the sentence graph to discourage the selection of similar items.

To achieve that, we incorporate the notion of negative endorsements in the extended model. We define a negative-transition state as a mean to propagate the negative endorsements from the ranked sentence to its adjacent nodes. After the central item has been found, it is transformed into a negative-transition state where any vertices with an edge connecting to the ranked sentence will have its edge weight convert to a negative value. Then, to find the next ranked sentence, a random walk is defined over the modified transition matrix. Sentence with the largest stationary distribution after the transformation is selected. According to figure 6.1, we find the first central item s_1 through random walk over a regular sentence graph in stage one, shown in 6.1a. After that, the top-ranked item s_1 is transformed into a negative state in order to penalize adjacent redundant items, shown in 6.1b. Because of the negative endorsements propagating from s_1, the importance of the connected

nodes to s_1 will be lowered as the walk progresses. The second central item s_2 can be found in stage two through random walk over the negative-state sentence graph. Then process continues until all sentences are ranked.

Next, we formally describe the proposed model as followings. Starting with an undirected sentence graph $G(V,E)$, we construct the modified transition probability matrix Q from the prior distribution matrix R where $0 \leq R_{ij} \leq 1$ and all row sums of R add up to 1, the transition probability matrix P, and a coefficient d. Then, we find the state with the largest stationary probability $s_1 = argmax_{i=1}^{n} \vec{\pi}_i$ to be the first ranked sentence. In the next stage, we convert s_1 into a negative-transition state P^- where $P_{si}^- = -P_{si}$ and $P_{is}^- = -P_{is}$. To preserve ergodicity properties, we scale up all negative transitions and find the modified transition probability matrix Q_s of the ranked sentences such that each element in Q_s is non-negative and all rows of Q_s and Q_s^T add up to 1.

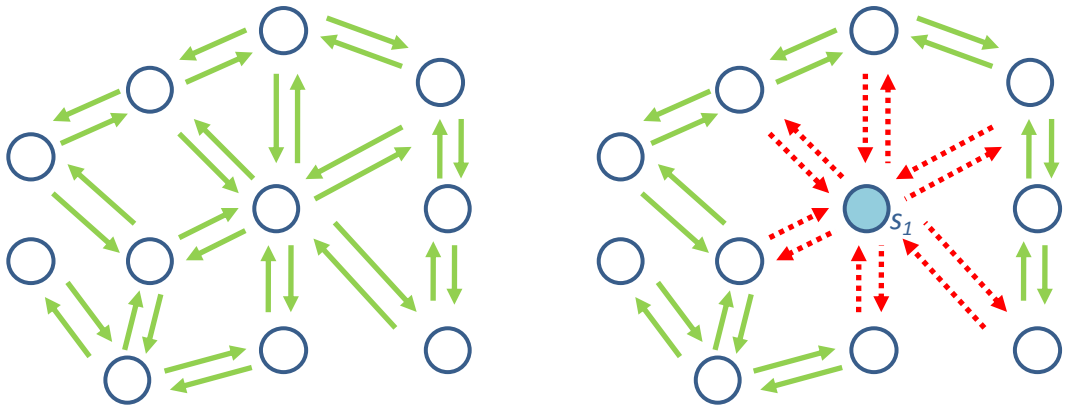
$$Q_s = (1-d)cR_s + dP_s^- \tag{6.1}$$

Here, the transition matrix Q_s is identical to Q defined in the basic NegativeRank. If we reorganize the initial modified transition probability matrix Q such that ranked sentences come before unranked sentences, Q can be written as:

$$Q = \begin{bmatrix} Q_s \\ U \end{bmatrix} \tag{6.2}$$

where Q_s corresponds to ranked sentences which have been transformed into a negative-transition state and U consists of unranked ones. Then, we can find the next central item s_2 by computing the stationary distribution of Q and take the one with the largest stationary probability to be s_2. Next, we turn s_2 into a negative-

transition state and keep repeating the process until all or the specified number of sentences are ranked. In contrast to GRASSHOPPER, we can repeatedly calculate the stationary distributions after each selected item has been transformed since the modified transition matrix is still a stochastic matrix where ergodicity properties still hold. This greatly simplifies and improves the efficiency of the ranking model.



a) Stage 1: Random walk over a regular sentence graph

b) Stage 2: Random walk over the sentence graph with negative states, after the first item s_1 is found.

Figure 6.1. An illustration of Multi-stage NegativeRank model.

6.4 Experimental Evaluation

6.4.1 Data Sets

We conduct a query-focused sentence extraction evaluation on three publicly available data sets. These are two test sets from the Document Understanding Conferences 2006 and 2007 (DUC06 and DUC07) and complex interactive question answering test set from TREC 2006 question answering track (ciQA06). All of them contain a collection of newswire articles drawn from the AQUAINT corpus. Each test set consists of a number of query-focused tasks. Each task essentially describes a

specific information need in a form of question. In addition, a ree-form narrative provides elaborate description of the information need. Next a set of answers or information nuggets, created by human experts, are available as a benchmark set.

Table 6.1. Summary of the data sets

Summary	DUC06	DUC07	ciQA06
Number of tasks	50	45	30
Number of reference nuggets per task	10	10	16
Number of candidate sentences per task	680	528	2,321

DUC06 and DUC 2007 are two test sets prepared by domain experts at NIST (National Institute of Standards and Technology) to be used in Document Understanding Conferences for evaluating focused summarization tasks. DUC06 contains 50 query-focused tasks while DUC07 contains 45 query-focused tasks. Each task is given a set of 25 relevant news articles and a set of gold standard sentences. The task description comprises a short title and a free-form narrative describing the information need in detail. On average, there are 10 reference nuggets per task.

ciQA06 contains 30 query-focused sentence tasks which require generating a set of answer passages for complex relationship questions. Similar to the DUC data sets, task description consists of a short query topic and a free-form narrative describing the specific aspects of the information need. Documents containing candidate answers to the test topics are drawn from the AQUAINT corpus. On average, there are 2,320.87 answer sentences per test question. In addition, NIST

assessors also create the answer key which consists of a list of vital/okay information nuggets for each question. There are, on average, 16 nuggets per task.

6.4.2 Evaluation Metrics

To evaluate the performance of diversity ranking methods, we employ two sets of performance metrics. The first set of metrics measure diversity based on n-gram co-occurrences between a set of reference nuggets and an automatically extracted set.

Formally, given a task q, we define $R = \{r_1, r_2, \ldots, r_m\}$ as a set of m reference nuggets and $A = \{a_1, a_2, \ldots, a_n\}$ as a set of n automatically extracted nuggets. To measure diversity between set R and A, an n-gram coverage between R and A is defined as coverage(R,A) = $|R \cap A|$. If coverage(R,A_1) is greater than coverage(R,A_2), then A_1 is more diverse than A_2. Consider the following example:

Task: What effect does steroid use have on an athlete's performance?

Reference Set	• Steroids enhance athletic performance. • Steroids act like testosterone, the male sex hormone, in building muscle mass. • Steroids have the same adverse health and social effects as narcotics.
Set *A*	• Steroids help boost athletic performance by improving muscle mass. • Steroids can cause many adverse effects.
Set *B*	• Steroids enhance athletic performance. • Athletes use steroids to improve their performance.

According to the above example, set A contains three matching nuggets while set B only contains one matching nuggets, compared to the reference set. Therefore,

set A is more diverse than set B according to our definition of diversity. In this study, we employ ROUGE (Lin and Hovy 2003) and nugget pyramid (Lin and Demner-Fushman 2005) metrics to evaluate the coverage-based diversity of automatically extracted sets. Specifically, we compute ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4), and pyramid F_1 score. The formulas for these metrics have been described in the previous chapter (equation 5.7-5.11).

The second set of metrics is based on an evaluation framework proposed by Clark et al. (2008). In their work, Clark et al. introduced α-NDCG, a performance metric that rewards diversity based on the Normalized Discounted Cumulative Gain (NDCG) measures used in evaluating a ranked list of search results (Järvelin and Kekäläinen 2002). Their framework makes a clear distinction between novelty and diversity as the need to avoid redundancy vs. the need to resolve ambiguity. According to this framework, documents which contain more information nuggets should be ranked higher than those with fewer nuggets.

The first step to compute NDCG is to create a gain vector G. Formally, the k-th element of G is defined as the following.

$$G[k] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}} \tag{6.3}$$

where $J(d_k,i) = 1$ if the document d contains information nugget n_i. Otherwise, $J(d_k,i) = 0$. α is a constant that reflects the possibility of judgment error, where $0 < \alpha \leq 1$. If $\alpha = 0$, then G represents standard binary relevance. Furthermore, $r_{i,k-1}$ is the number of documents that contain nugget n_i ranked up to position k-1. It is formally defined as follow.

$$r_{i,k-1} = \sum_{j=1}^{k-1} J(d_j, i) \tag{6.4}$$

The next step is to compute the cumulative gain vector *CG*.

$$CG[k] = \sum_{j=1}^{k} G[j] \tag{6.5}$$

In addition, to model the user effort required to reach documents at the lower rank, a discount maybe applied at each rank to penalize such documents. Normally, $log_2(1+k)$ is used as a discount function. The k-th element of a discount cumulative gain vector *DCG* is defined as follow.

$$DCG[k] = \frac{\sum_{j=1}^{k} G[j]}{log_2(1+j)} \tag{6.6}$$

Finally, we normalize *DCG* by the ideal discounted cumulative gain vector *DCG'* to compute *NDCG*. The ideal cumulative gain represents the ideal ordering that maximizes cumulative gain at every level. We formally describe NDCG as the following.

$$NDCG[k] = \frac{DCG[k]}{DCG'[k]} \tag{6.7}$$

In this work, we calculate both *DCG* and *NDCG* according to the previous study by Al-Maskari et al. (2007), which suggested that *DCG* tends to correlate better with user satisfaction than *NDCG*. Note that NDCG framework requires a manual assessment of information nuggets contained in each document. In this work, we automatically obtain the nugget judgments for each sentence using Pourpre script.

It has been shown by Lin and Demner-Fushman (2005) that Pourpre's judgments correlate very well with human judgments (over 80% Kendall's Tau and over 90% R^2).

Table 6.2. Summary of performance metrics

Metrics	Diversity Measure
ROUGE-1, ROUGE-2, ROUGE-SU4, Pyramid F_1	N-gram coverage
DCG, NDCG	Nugget-based discounted cumulative gain

6.4.3 Evaluation Settings

In the previous chapter, we found significant differences in the performance of ranking methods across the different sizes of extracted sets, $p<0.05$. Due to the fact that some methods may perform well when extracting a smaller set of sentences while some may perform better on a larger extracted set, we investigate the performance of diversity ranking methods on three different sizes of extracted set: *small*, *medium* and *large*. The sizes are defined in either the number of words or characters unit, depending on which evaluation metric is being computed. In order to compute ROUGE scores, a small size is defined as 100-word set. A medium size contains 250 words. Lastly, a large extracted set is 500 words in length. Next, to compute nugget pyramid scores, we set the sizes of the extracted set to be 500, 1,500, and 3,000 characters for small, medium, and large set, respectively. The experiment has a similar setup compared to a top-*k* sentence retrieval experiment.

In this case, the predefined value is the size of the extracted set, not the total number of sentences retrieved.

Table 6.3. Summary of the data sets

Extracted Set	# of words	# of characters
Small	100	500
Medium	250	1,500
Large	500	3,000

Next, we compare the performance of the proposed method with many state-of-the-art ranking models. We briefly describe them as follows.

Topic-Sensitive LexRank. Erkan and Radev (2004) initially proposed LexRank algorithm to extract sentences in generic summarization task. Later, it has been extended to handle query-focused sentence extraction applications, e.g. focused summarization and question answering by Otterbacher et al. (2005). Although it does not reward diversity upfront, its ranking model accounts for information subsumption among sentences (Radev 2000). That is, important sentences are selected from the centroids of sentence clusters. Thus, each selected sentences is inherently novel from one another. Consequently, the extracted set of sentences is expected to be diverse. For a formal description of LexRank , please refer to equation 5.14.

NegativeRank. In the previous chapter, we adapted a notion of negative endorsement to the eigenvector centrality ranking. The key idea is to represent redundancy between vertices as negative-signed edges. Then, a random walk is defined over negative-edge graph to find the stationary distribution of each vertex.

As a result of negative endorsements, the stationary distributions of redundant sentences will be lowered than those which are relatively more novel. And subsequently, a diversity in the extracted set of sentences is promoted.

GRASSHOPPER. In contrast to the topic-sensitive LexRank, Zhu et al. (2007) addressed the diversity issue in eigenvector-centrality ranking by introducing an absorbing Markov chain random walks approach called GRASSHOPPER (Graph Random-walk with Absorbing States that HOPs among Peaks for Ranking). They argued that, since eigenvector centrality does not consider diversity upfront, it is likely that top ranked sentences are going to be dominated by those from the central cluster. To avoid such issue, each ranked sentence will be transformed into absorbing state as the ranking process continues. Effectively, the importance of similar unranked nodes will be lowered by the absorbing nodes. Therefore, each ranked sentences are novel and thus encouraging diversity of the extracted set.

The overall GRASSHOPPER algorithm can be described as follows. Starting from the undirected sentence graph $G(V,E)$, the modified transition probability matrix Q, as described earlier, is constructed from the transition matrix P, the prior distribution matrix R, and the coefficient d. Next, it computes the stationary distribution and select the vertex with the largest stationary probability to be the first ranked sentence s_1: $s_1 = argmax_{i=1}^{n} \vec{\pi}_i$. Then, s_1 is turned into an absorbing state by setting $P_{ss} = 1$ and $P_{si} = 0, \forall i \neq s$. Next, to find the subsequent ranked sentence, the expected number of visits to each vertex after the ranked sentence become an absorbing node is computed. Intuitively, those sentences which are highly similar to the absorbing nodes will have fewer visits by the random walk as the walk will get absorbed eventually. On the other hand, sentences which are less similar to the

absorbing ones will have relatively more visits. A sentence with the largest expected number of visits will be selected as the next ranked sentence s_2. After which, it is transformed into an absorbing node. The process continues until all sentences are ranked.

Super-centrality. Chen et al. (2009) recently introduced a greedy approximation ranking model which measures the importance of a subset of vertices as a whole. Their method considers both centrality and diversity together in the ranking process similar to GRASSHOPPER. The main goal of the super-centrality method is extracting a set of super vertices. The super vertex *s* is defined as an important vertex whose content subsumes the content of all vertices. All super vertices have an outgoing edge weight of 1 while an incoming edge weight is equal to the maximal weight of edges between any other vertices. That is, suppose $w(i \rightarrow j)$ denotes an edge weight from vertex *i* to vertex *j*, the super vertex *s* is formally defined as $s = subset(V)$ where $\forall i \in V, w(s \rightarrow i) = 1$ and $w(i \rightarrow s) = argmax_{j \in V} w(i \rightarrow j)$. The importance of super vertices is quantitatively measured as super centrality.

The process to find a set of super vertices is described as follows. First, given an undirected sentence graph *G*(*V*,*E*), graph *G* is represented as a transition probability matrix *P*. A random walk over *G* is defined to induce the stationary distribution $\vec{\pi}P = \vec{\pi}$. Then, the sentence with the largest stationary probability is selected as the first super vertex s_1. After the super vertex is found, the original sentence graph *G* is modified into an extended graph *G'(V',E')* where $V' = V \cup s_1$. Edge weights of the super vertex s_1 are set to $\forall i \in V, w(s_1 \rightarrow i) = 1$ and $w(i \rightarrow s_1) = argmax_{j \in V} w(i \rightarrow j)$. To find subsequent super vertices, multiple iterations are required in the similar manner as GRASSHOPPER's ranking process. In this work, we implement a heuristic

algorithm proposed by Chen et al. to find all super vertices. To encourage diversity, the selected super vertex s_k needs to maximize the objective function $\triangle\pi_s(s_k)$. Simply put, if s_k is redundant to any ranked sentences in $S = \{s_1, s_2,\ldots, s_{k-1}\}$, then $w(i \rightarrow s_k)$ and $w(i \rightarrow S)$ will be virtually identical. Thus, adding s_k to S does not contribute to the objective function $\triangle\pi_s(s_k)$.

Structure Learning. Recent progress in machine learning has addressed the complementary issue of designing the classification algorithms that can deal with complex inputs and outputs, such as sets (Tsochantaridis et al. 2005). In those methods, structural SVMs have shown high potential for building highly complex and accurate models in areas like language processing, protein structure prediction, and information retrieval.

Yue and Joachims (2008) have employed such method to predict diverse subsets using loss functions to penalize low diversity. We adapt their structural SVM approach and apply it to sentence extraction task. Our method utilizes reference sentences in a gold standard set as low-dimensional semantic representation which proves to be a novel solution for sentence extraction tasks.

Given a set of documents, each of which include a set of candidate sentences $X = \{X_1 \ldots X_n\}$ and a set of reference sentences $T = \{T_1 \ldots T_m\}$, our goal is to select a subset Y of X for each document which maximizes factual diversity and coverage. To achieve that, we need to learn a hypothesis function h: $X \rightarrow Y$ to predict a subset Y when given a set of sentences X from training data sets. That is, given a set of training examples, $S = (X^{(n)}, T^{(m)})$ where n denotes the number of candidate sentences while m denotes the number of reference sentences in gold standard set, we want to

predict m sentences from set X. Specifically, we want to find a function h which minimizes the empirical risk.

$$R = \frac{1}{M} \times \sum_{i=1}^{m} \triangle(T_{(i)}, h(X^{(n)}) \tag{6.8}$$

Here, diversity is promoted by the loss function $\triangle(T, y)$. We use structural SVM classifier following the optimization problem 1 (Yue and Joachim 2008) to learn a weight of vector w.

$$min_{w,\xi \geq 0} \frac{1}{2}||w||^2 + \frac{C}{N} \sum_{i=1}^{N} \xi_i \tag{6.9}$$

$$s.t. \forall i, \forall y \in Y \backslash y^{(i)} : \tag{6.10}$$

$$w^Y \psi(x^{(i)}, y^{(i)}) \geq w^T \psi(x^{(i)}, y) + \triangle(T^{(i)}, y) - \xi_i \tag{6.11}$$

where C is a parameter that controls the tradeoff between model complexity and a hinge loss relaxation of the training loss for each training example. To solve the optimization problem 1 efficiently, we can use cutting plane algorithm (Yue and Joachim 2008) to iteratively add constraints until the original problem is solved within a desired tolerance.

We set the threshold to 0.2 for the similarity of sentences in document and reference sentences in the gold standard. Each sentence is represented as a TFIDF-weighted unigram feature vector. Next, we use loss function $\triangle(T, y)$ to be the weighted percentage of reference sentences in the gold standard set which are not covered. For a given candidate set, each sentence's weight is proportional to the number of sentences in the documents whose similarity to this sentence is above the threshold. Therefore, it penalizes the redundancy. Each dataset is split into training

data sets and test data sets. 10-fold cross-validation is employed in the experiment. Table 6.4 displays the summary of methods we compare in the experiment.

Table 6.4. Summary of methods compared in the experiment

Abbreviation	Method	Diversity Ranking Principle
LR	Topic-sensitive LexRank	Cross-sentence information subsumption
NR	NegativeRank	Negative endorsements
GH	GRASSHOPPER	Absorbing states random walk
SC	Super-centrality	Maximizing the stationary distribution gain
MNR	Multi-stage NegativeRank	Negative states random walk
SVM	SVM Diversity	Subtopic diversity

6.5 Results and Discussion

6.5.1 N-gram coverage

Table 6.5. F_1 scores of each method on DUC06 data set

Size	Method	R-1	R-2	R-SU4	Pyramid
Small	LR	0.2555	0.0560	0.0893	0.1809
	NR	0.2544	0.0538	0.0872	0.1698
	GH	**0.2712**	0.0572	**0.0958**	**0.2025**
	SC	0.2565	0.0482	0.0871	0.1969
	MNR	0.2649	**0.0577**	0.0942	0.2018
	SVM	0.2202	0.0344	0.0661	0.1282
Medium	LR	0.3728	0.0782	0.1313	0.2670
	NR	0.3688	0.0744	0.1276	0.2683
	GH	**0.3833**	0.0783	0.1341	0.2766
	SC	0.3741	0.0741	0.1282	0.2791
	MNR	0.3805	**0.0824**	**0.1367**	**0.2797**
	SVM	0.3524	0.0600	0.1125	0.2218
Large	LR	0.3514	0.0783	0.1303	0.3072
	NR	0.3463	0.0743	0.1263	0.3012
	GH	0.3580	0.0803	**0.1329**	0.3039
	SC	**0.3589**	0.0808	0.1324	**0.3110**
	MNR	0.3534	**0.0809**	0.1316	0.3071
	SVM	0.3588	0.0674	0.1201	0.2541

Table 6.6. F_1 scores of each method on DUC07 data set

Size	Method	R-1	R-2	R-SU4	Pyramid
Small	LR	0.2567	0.0554	0.0910	0.1826
	NR	0.2543	0.0555	0.0900	0.1766
	GH	0.2775	0.0651	0.1007	0.2082
	SC	0.2611	0.0605	0.0938	0.2048
	MNR	**0.2825**	**0.0720**	**0.1071**	**0.2212**
	SVM	0.2365	0.0537	0.0842	0.1526
Medium	LR	0.3865	0.0879	0.1437	0.2837
	NR	0.3820	0.0834	0.1380	0.2739
	GH	0.3960	0.0901	0.1458	0.2979
	SC	0.3995	0.0893	0.1453	0.2944
	MNR	**0.4016**	**0.0966**	**0.1518**	**0.3087**
	SVM	0.3686	0.0887	0.1365	0.2555
Large	LR	0.3656	0.0912	0.1423	0.3209
	NR	0.3675	0.0921	0.1429	0.3156
	GH	0.3746	**0.0994**	0.1487	0.3347
	SC	0.3739	0.0934	0.1457	0.3252
	MNR	0.3703	0.0985	**0.1487**	**0.3374**
	SVM	**0.3764**	0.0958	0.1438	0.2630

Table 6.7. F_1 scores of each method on ciQA06 data set

Size	Method	R-1	R-2	R-SU4	Pyramid
Small	LR	0.2083	0.0532	0.0793	0.2222
	NR	0.2297	0.0617	0.0908	0.2292
	GH	0.2360	0.0690	0.0964	**0.2822**
	SC	0.2188	0.0534	0.0811	0.2437
	MNR	**0.2437**	**0.0706**	**0.1000**	0.2622
	SVM	0.2406	0.0624	0.0913	0.2467
Medium	LR	0.3169	0.0822	0.1224	0.3174
	NR	0.3192	0.0867	0.1275	0.3282
	GH	0.3269	0.0950	0.1333	0.3671
	SC	0.3225	0.0822	0.1254	0.3636
	MNR	0.3312	0.0926	0.1336	0.3439
	SVM	**0.3738**	**0.1090**	**0.1526**	**0.3696**
Large	LR	0.3157	0.0900	0.1298	0.3531
	NR	0.3141	0.0920	0.1326	0.3655
	GH	0.3241	0.0975	0.1366	0.3668
	SC	0.3203	0.0908	0.1321	0.3636
	MNR	0.3261	0.0991	0.1386	0.3832
	SVM	**0.4048**	**0.1375**	**0.1795**	**0.4185**

The performance comparison of different ranking methods as measured by R-1, R-2, R-SU4 and nugget pyramid metrics on DUC06, DUC07, and ciQA06 data sets is shown in table 6.5, 6.6, and 6.7, respectively. Overall, the unified eigenvector centrality and diversity ranking methods, i.e. GH, SC, and MNR, consistently produced the best results across all three data sets. For instance, consider the task

of extracting a small set of sentences from DUC06 test data. The best methods performed 6.13%, 3.05%, 7.34%, and 11.94% better than LR, which only focuses on eigenvector centrality ranking, according to R-1, R-2, R-SU4, and nugget pyramid scores, respectively. In a similar case, the best methods also performed 6.62%, 7.29%, 9.95%, and 19.26% better than NR, which focuses on diversity ranking. In addition, the unified eigenvector centrality and diversity ranking methods also produced more diverse sets of sentences, compared to the supervised SVM diversity method. For example, The best methods performed 23.19%, 67.64%, 44.95%, and 57.92% better than SVM in extracting small sets of sentences for DUC06 tasks.

Within the centrality and diversity ranking methods, MNR is the best overall method considering the results across data sets and evaluation metrics. Table 6.8, 6.9, and 6.10 summarize the top three performers at the different extraction sizes and test data. On DUC06, MNR produced the best results according to R-2@small, R-2@medium, R-SU4@medium, nugget pyramid@medium, and R-2@large. Next, on DUC07, MNR is the best method in almost all combinations, except R-1@large and R-2@large. Lastly, on ciQA06, MNR performed the best at R-1@small, R-2@small, and R-SU4@small. We found that GH performed quite as effective as MNR in many DUC06, DUC07, and ciQA06 tasks. Within those tasks, they performed approximately within 1% to 10% of each other. This is not surprising considering that they employ the similar ranking principle. Furthermore, the performance gaps between MNR and SC were relatively larger in many cases where MNR's results were about 10% - 30% better than SC's results.

The results at the different extraction sizes also favored the centrality and diversity ranking methods. The differences were larger when the extracted sets were small. As the size grows, the performance gaps between GH, SC, and MNR became

smaller. Interestingly, SVM did not produce the extracted set as diverse as those methods under these performance metrics. It did, however, significantly outperform all of them when extracting the medium and large sentence sets for ciQA06 tasks at $p<0.05$. We expect that the addition of complex feature sets should improve SVM's performance in many cases.

In conclusion, when diversity is measured by an n-gram coverage, the graph-based ranking methods that incorporate both centrality and diversity produced the best results, particularly when the extracted sets were 250 words in size or smaller. The performance of SVM diversity which focuses on optimizing subtopic diversity was much poorer than most methods being evaluated in the study. Nonetheless, it was able to extract increasingly diverse set of sentences as the extraction size was 250 words or larger.

Table 6.8. The top three n-gram coverage performers on DUC06 data at the different extraction sizes.

Small				Medium				Large			
R-1	R-2	R-SU4	Py	R-1	R-2	R-SU4	Py	R-1	R-2	R-SU4	Py
GH MNR SC	MNR GH LR	GH MNR LR	GH MNR SC	GH MNR SC	MNR GH LR	MNR GH LR	MNR SC GH	SC SVM GH	MNR GH SC	GH SC MNR	SC LR MNR

Table 6.9. The top three n-gram coverage performers on DUC07 data at the different extraction sizes.

Small				Medium				Large			
R-1	R-2	R-SU4	Py	R-1	R-2	R-SU4	Py	R-1	R-2	R-SU4	Py
MNR GH SC	MNR GH SVM	MNR GH SC	MNR GH SC	MNR SC GH	MNR SVM GH	MNR GH SC	MNR GH SC	SVM GH SC	GH MNR SVM	MNR GH SC	MNR GH SC

Table 6.10. The top three n-gram coverage performers on ciQA06 data at the different extraction sizes.

Small				Medium				Large			
R-1	R-2	R-SU4	Py	R-1	R-2	R-SU4	Py	R-1	R-2	R-SU4	Py
MNR SVM GH	MNR GH SVM	MNR GH SVM	GH MNR SC	SVM MNR GH	SVM GH MNR	SVM MNR GH	SVM GH SC	SVM MNR GH	SVM MNR GH	SVM MNR GH	SVM MNR GH

6.5.2 Discounted Cumulative Gain

Table 6.11. Discounted cumulative gain scores of each method on DUC06 data set

Size	Method	DCG	NDCG
Small	LR	7.8480	0.86025
	NR	7.8778	0.8859
	GH	8.2387	0.84804
	SC	**8.7085**	**0.90691**
	MNR	8.2898	0.8704
	SVM	7.3774	0.8587
Medium	LR	6.8354	0.77238
	NR	7.2143	**0.7898**
	GH	7.0444	0.75123
	SC	**7.3901**	0.78262
	MNR	6.7694	0.7677
	SVM	6.5519	0.71872
Large	LR	6.0250	0.69256
	NR	6.5975	**0.7326**
	GH	6.4896	0.71774
	SC	**6.6478**	0.7323
	MNR	6.5928	0.7274
	SVM	5.8905	0.65188

Table 6.12. Discounted cumulative gain scores of each method on DUC07 data set

Size	Method	DCG	NDCG
Small	LR	7.2303	0.83497
	NR	7.4631	0.8818
	GH	7.7582	0.8567
	SC	**8.4475**	**0.9431**
	MNR	8.1090	0.8974
	SVM	7.5225	0.85204
Medium	LR	5.8363	0.66941
	NR	6.0668	0.7216
	GH	6.1688	0.6966
	SC	**6.8152**	**0.7911**
	MNR	6.6828	0.7784
	SVM	6.1709	0.7108
Large	LR	5.0145	0.59404
	NR	5.4080	0.6676
	GH	5.2621	0.6279
	SC	6.0318	**0.753**
	MNR	**6.2440**	0.7390
	SVM	6.0176	0.70534

Table 6.13. Discounted cumulative gain scores of each method on ciQA06 data set

Size	Method	DCG	NDCG
Small	LR	4.3029	0.1000
	NR	10.7016	**0.9039**
	GH	4.9637	0.82038
	SC	5.7556	0.1000
	MNR	10.9105	0.8839
	SVM	**11.3253**	0.89234
Medium	LR	3.3873	0.6119
	NR	8.5794	0.7310
	GH	4.1564	0.65262
	SC	4.4346	0.71511
	MNR	**8.6931**	**0.7517**
	SVM	8.3981	0.74514
Large	LR	3.0442	0.5084
	NR	**7.6289**	0.6531
	GH	3.8195	0.61809
	SC	3.9733	0.6311
	MNR	7.4757	0.6568
	SVM	7.2605	**0.70746**

The performance comparison of different ranking methods as measured by DCG and NDCG metrics on DUC06, DUC07, and ciQA06 data sets is shown in table 6.11, 6.12, and 6.13, respectively. In this case, the results are different from the previous evaluation metrics. Unlike the coverage-based diversity evaluation, not all unified eigenvector centrality and diversity ranking methods produced the best

results. While SC and MNR were still among the top three methods in most cases according to DCG and NDCG scores, we found that GH performed relatively poor compared to its counterparts. In addition, the results of NR and SVM were highly competitive on many tasks. Since NR employs the negative-edge graph to promote diversity, we were not surprised that it was able to effectively rank the novel sentences at the relatively high positions on the ranked list.

Table 6.14, 6.15, and 6.16 show the top three performers at the different extraction sizes and test data. The best method that performed consistently well on DUC06 and DUC07 tasks is SC. Particularly, it performed quite effectively at the small and medium sizes. On DUC06, SC produced the best results at DCG@small, NDCG@small, DCG@medium, and DCG@large. Next, SC also produced the optimum results on most DUC07 evaluations, except at DCG@large. Finally, there was no single method that completely dominated the others on ciQA06 evaluations. In this case, NR, MNR, and SVM were among the best methods. NR produced the optimum results at NDCG@small and DCG@large, MNR performed the best at DCG@medium and NDCG@medium, and SVM was the best method at DCG@small, and NDCG@large. Interestingly, SC performed quite poorly on ciQA06 tasks. Moreover, we observed that the performance gaps between methods were relatively larger on ciQA06 tasks, compared to those on DUC06 and DUC07 tasks. For example, when extracting the small sets of sentences, the top method (SVM) produced 163.20% higher DCG score than the worst method (LR) on ciQA06 tasks. On the other hand, on DUC06, the top method (SC) produced 18.04% higher DCG score than the worst method (SVM) while, on DUC07, the DCG score of the top method (SC) was 16.83% higher than the DCG score of the worst method (LR). Lastly, the results of SVM displayed the trend similar to the previous evaluation. That is, SVM diversity

produced significantly better results on ciQA06 tasks than DUC06 and DUC07 tasks, $p<0.05$.

Overall, the results from the discounted cumulative gain metrics offer a contrast view from the n-gram coverage. We found that not all methods consistently produced the best results across all evaluation metrics. A method which performed well on n-gram coverage metric, such as GH, might not necessarily be the best method on cumulative gain metrics, and vice versa. Moreover, we observed that the best method on DUC06 and DUC07 tasks (SC) consistently produced the optimum results across all extraction sizes. However, its performance dropped significantly when extracting the sets of sentences for ciQA06 tasks. One reason why SC performed relatively well in many cases is that it used a greedy approximation algorithm to rank sentences. As DCG and NDCG computed the gain values by discounting the rank, the metrics are generally optimized by greedy strategies.

Table 6.14. The top three discounted cumulative gain performers on DUC06 data at the different extraction sizes.

Small		Medium		Large	
DCG	NDCG	DCG	NDCG	DCG	NDCG
SC MNR GH	SC NR MNR	SC NR GH	NR SC LR	SC NR MNR	NR SC MNR

Table 6.15. The top three discounted cumulative gain performers on DUC07 data at the different extraction sizes.

Small		Medium		Large	
DCG	NDCG	DCG	NDCG	DCG	NDCG
SC MNR GH	SC MNR NR	SC MNR SVM	SC MNR NR	MNR SC SVM	SC MNR SVM

Table 6.16. The top three discounted cumulative gain performers on ciQA06 data at the different extraction sizes.

Small		Medium		Large	
DCG	NDCG	DCG	NDCG	DCG	NDCG
SVM MNR NR	NR SVM MNR	MNR NR SVM	MNR SVM NR	NR MNR SVM	SVM MNR NR

6.5.3 Agreements between the performance metrics

We further examined the Pearson correlation coefficient (R^2) between the performance metrics. First, table 6.17 displays R^2 between n-gram coverage metrics. According to the average values, R-SU4 has the strongest positive correlation with the other metrics (R^2=0.901) while R-1 has the weakest positive correlation (R^2=0.7714). This is not surprising given that R-SU4 accounts for a reasonable degree of text variation, i.e. up to 4-word skip bigrams, when comparing between two sets of sentences. In contrast, R-1 only computes simple unigram co-occurrences between sets. Next, the Pearson correlation coefficients between discounted cumulative gain metrics are shown in table 6.18. In this case, DCG moderately correlates with NDCG (R^2=0.6159). Finally, table 6.19 displays R^2 between the n-gram coverage and the discount cumulative gain metrics. As can be seen, there is a

weak inverse correlation between the two types of performance metrics. This observation is illustrated by the performance inconsistency of some methods, e.g. NR, GH, and SC, in which they produced considerably high scores on one type of metrics, but performed relatively worse on the other type of metrics. The fact that the two types of metrics measure diversity very differently makes it difficult to draw a definite conclusion on the absolute performance of diversity-ranking sentence extraction methods. We believe it is helpful to include both types of metrics in the diversity evaluations as they provide different perspectives of the performances.

Table 6.17. The Pearson correlation coefficients between n-gram coverage metrics

	R-1	R-2	R-SU4	Pyramid	Average
R-1	-	0.7632	0.9302	0.6207	0.7714
R-2	0.7632	-	0.9439	0.8969	0.8680
R-SU4	0.9302	0.9439	-	0.8289	**0.9010**
Pyramid	0.6207	0.8969	0.8289	-	0.7821

Table 6.18. The Pearson correlation coefficients between discounted cumulative gain metrics

	DCG	NDCG
DCG	-	0.6159
NDCG	0.6159	-

Table 6.19. The Pearson correlation coefficients between n-gram coverage and discounted cumulative gain metrics

	DCG	NDCG
R-1	-0.3191	-0.0592
R-2	-0.3157	-0.1870
R-SU4	-0.3470	-0.1454
Pyramid	-0.4034	-0.3422
Average	-0.3460	-0.1835

6.5.4 The running-time efficiency among graph-based ranking models

Table 6.20. The average running time (in seconds) of different graph-based ranking models across all tasks

Method	DUC06		DUC07		CIQA06	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
LR	0.060	0.144	0.054	0.169	0.250	0.136
NR	0.662	0.671	0.656	1.163	3.416	1.946
GH	4.429	5.755	3.234	16.476	41.44	34.431
SC	0.414	0.611	0.334	0.807	1.887	1.044
MNR	0.207	0.703	0.180	0.787	0.901	0.506

Table 6.20 displays the running time of various graph-based ranking models averaged across all tasks on three data sets. Overall, it took considerably longer to extract the sets of sentences for CIQA06 tasks than DUC06 and DUC07 tasks as the number of candidate sentences are much larger in the former data set. Next, it is not

surprising that LR was the most efficient method as it did not explicitly employ any diversity promotion process. The regular NR required sharply more running time to produce the outputs than both LR and MNR. This can be explained by the fact that NR transformed every edge weight into a negative edge thus it took longer for random walk to converge. Among the unified centrality and diversity ranking methods, MNR produced the most efficient performance. On the other hand, GH was the least efficient method due to the fact that it requires inverting the transition matrix to compute the expected number of visits in each iteration. This suggests that MNR is a good alternative to GH as both methods performed relatively effective in many cases while it took considerably less time for MNR to produce the outputs.

6.6 Conclusion

In this chapter, we presented a comprehensive evaluation of several state-of-the-art diversity ranking models. Different diversity ranking principles are considered, for example, a cross-sentence information subsumption principle which implicitly promotes diversity, the unified eigenvector centrality and diversity ranking principles, and a subtopic diversity ranking principle. Next, two types of diversity measurements are defined. First, different ROUGE variants and nugget pyramid were employed as the n-gram coverage based diversity measures. Next, the discounted cumulative gain metrics, which include both normalized and un-normalized variants, were considered as the second set of diversity measures. The experimental results suggested that the unified centrality and diversity ranking models significantly outperformed the other methods in most cases, $p<0.05$. Within the unified ranking principle, Multi-stage NegativeRank outperformed both GRASSHOPPER and Super-centrality according to the overall ROUGE and pyramid

scores. It performed equally well with Super-centrality on the discounted cumulative gain measures. SVM diversity performed quite poorly at a small extraction size. However, its performance started to increase as the extraction size increased. Furthermore, we found that there were reasonable disagreements between different types of diversity measurements while there were strong agreements of metrics of the same category according to Pearson correlation coefficients. ROUGE-SU4 produced the scores that correlate strongly with other ROUGE variants and nugget pyramid (R^2=0.9) while DCG and NDCG scores correlate moderately with each other (R^2=0.62). Due to the fact that each type of metrics treats diversity computation differently, many ranking models tended to perform at different levels, depending on the type of evaluation metric.

Finally, this chapter answers research question 4 by proposing Multi-stage NegativeRank model. The proposed method incorporated diversity ranking into the traditional eigenvector centrality ranking by iteratively transforming ranked item into a negative state. It employed the ranking principle similar to the absorbing Markov chain random walks. The major differences are in the mechanism to penalize redundancy and the ranking computation. Based on the results obtained in this chapter, the proposed method was very effective in promoting diversity of the extracted sets, compared to most state-of-the-art methods. As expected, the unified centrality and diversity ranking models significantly outperformed the baseline topic-sensitive LexRank, $p<0.05$, which implicitly promote diversity via the cross-sentence information subsumption principle. The improvements vary proportionately depending on the extraction sizes, the test set, and the performance metrics. In most cases, the unified eigenvector centrality and diversity ranking principle produced the most diverse results, compared to subtopic diversity and

cross-sentence information subsumption principles. Moreover, we found that different types of diversity measurements tend to disagree strongly. This can be explained by the ways diversity are promoted in the performance computation. For example, the n-gram coverage metrics calculate the performance scores for the unranked set while the discounted cumulative gain metrics compute the scores for the ranked set.

CHAPTER 7: CONCLUSIONS AND FUTURE WORK

"Every honest researcher I know admits he's just a professional amateur. He's doing whatever he's doing for the first time. That makes him an amateur. He has sense enough to know that he's going to have a lot of trouble, so that makes him a professional."

-- Ex-vice president of General Motors Research
& inventor of the electronic automobile starter.

This thesis has investigated several methods to improve the results of query-focused sentence extraction. Specifically, two major areas are explored: the similarity measure and the ranking principle. The main improvement is measured in terms of diversity of facts being selected into the extracted sets. In the area of sentence similarity measure, we started exploring the issue of measuring the semantic similarity between interrogative sentences. Chapter 3 demonstrated that a hybrid framework that include semantic, syntactic, and question category information considerably helped identify similar factoid question pairs. Chapter 4 focused on the variability of natural language expression problem. As the same sentences can be reformulated in various linguistic forms, most document-level similarity measures are not effective in finding semantically similar sentences. The results presented in this chapter indicated that sentence semantic structure is helpful in addressing this problem. Chapter 5 further investigated the effects of sentence similarity measures in sentence extraction contexts. It also explored a sentence ranking principle based on the idea of negative endorsements between redundant items. The combination of

sentence semantic structure and negative endorsements has proved to effectively promote diversity of the extracted sets. Lastly, chapter 6 focused on incorporating negative-endorsement based diversity ranking into centrality ranking. The experimental results demonstrated that diversity can be further improved when the unified centrality and diversity ranking model was utilized.

We reiterate the major contributions of the thesis in this chapter. Additionally, the implications for future directions of this work are discussed.

7.1 Contributions

As previously discussed in the first chapter, five major contributions of the thesis are summarized as follows. This thesis:

1. Empirically demonstrates that a hybrid framework, which consists of a linear combination of semantic, syntactic, and question category components, is very effective in identifying similar interrogative sentences. Compared with word occurrence based approaches, the proposed method is able to predict the similar factoid question pairs much more accurately.
2. Introduces a novel sentence similarity measure which utilizes sentence semantic structure to deal with the variability of natural language expression. It shows that the similarity judgment between sentences can be significantly improved when the similarity computation is done at the sentence structure level. Particularly, the proposed measure is relatively better at judging textual entailment pairs than most word-overlap or co-co-occurrence based measures.
3. Presents a novel sentence extraction method which incorporates the semantic structure of sentence in the sentence similarity judgment. The experimental

evaluation on focused summarization and question answering shows that the sets of sentences, extracted by the proposed method, contain fewer number of factually redundant sentences. It suggests that the quality of focused summaries and complex answers can be significantly improved when the more effective similarity measure is employed as part of sentence extraction method.

4. Demonstrates that redundancy can be modeled as a negative relations in sentence graph. By assigning a negative sign to edges between the sentence vertices, the random walks over the negative-edge graph will lower a redundant sentence's rank if it contains a significant degree of negative endorsements from the similar sentences. The results of focused summarization and question answering indicates that the proposed method significantly increases the diversity of the extracted sets, compared to other sentence extraction methods.
5. Presents a unified centrality and diversity ranking model based on negative state random walk. The proposed method extends the negative-endorsement ranking principle by treating centrality ranking and diversity ranking together in one integrated process The thesis provides an empirical evidence that the unified centrality and diversity approaches are very effective in promoting diversity of the set of sentences.

7.2 Future Work

We present a number of potential directions for future work in this section. It describes several ways to further extend the methods proposed in this thesis and the relevant application domains to which they can be applied.

7.2.1 Identifying the similarity or relation between sentences

The results presented in chapter 3, 4, and 5 provide a starting point toward a more robust similarity computation of sentences. Still, there are many interesting areas of improvement to be pursued in future work.

First, this thesis demonstrates that, by focusing on dealing with the variability of natural language expression, the accuracy of similarity judgment can be improved. Still, we have not invested much effort on the variability at the word level. The lexical mismatch or semantic gap problem (Berger et al. 2000) has been investigated in factoid question answering area for a decade. And it's still one of the ongoing research topics for many text mining applications. One common approach is to enrich a short-text representation with world knowledge using Wikipedia (Banerjee et al. 2007; MacKinnon and Vechtomova 2008; Phan et al. 2008; Hu et al. 2009a; Hu et al. 2009b). We believe the lexical enrichment is an acceptable tradeoff to efficiency since sentences are relatively short to begin with. The question is what is the best way to expand the representations to achieve the best results without introducing noisy or redundant information. Moreover, apart from Wikipedia, there are other external knowledge sources which can be utilized. These include extended WordNet (Harabagiu et al. 1999), VerbNet (Kipper et al. 2008) VerbOcean (Chklovski and Pantel 2004), and search result snippet (Sahami and Heilman 2006). Next, apart from semantic equivalence and entailment relations, other types of judgment can be explored. For instance, a similarity function which is able to identify a contradiction between two statements at a sentence similarity level can be helpful in a task of mining contradictory opinions (Kim and Zhai 2009).

In the past few years, there has been a renewed interest in question retrieval methods (Jeon et al. 2005a; Jeon et al. 2005b; Jijkoun and de Rijke 2005) due to the

popularity of question answering communities (Bian et al. 2008; Liu et al. 2008). The work in chapter 3 can be extended to community question answering domain. The key ideas in our hybrid question similarity framework can be implemented in other retrieval models. For example, a recent work by Cao et al. (2009) demonstrated that the inclusion of question category helps improve the performance of language-model based question retrieval system. Next, question retrieval models may consider the quality of answers (Suryanto et al. 2009) as an additional feature.

7.2.2 Negative Edges and Diversity in Ranking

Chapter 5 and 6 emphasizes the importance of diversity in ranking. The thesis explores the idea of modeling redundancy as negative edges in a sentence graph. Several improvements can be incorporated in the graph representation in a context of sentence extraction. For example, inter-sentence relations can be modeled as directed graph. This can be achieved by employing different edge weighting function, e.g. affinity weight (Zhang et al. 2005; Wan and Yang 2008). Next, different graphical ranking models, such as bipartite graph (Jeh and Widom 2002; Singh et al. 2007), is another interesting direction to explore. In addition, the results in chapter 5 and 6 indicate that the initial ranking distributions significantly affect the performance of the ranking models. We can examine the effects of the more sophisticated models, e.g. Latent Dirichlet Allocation (Blei et al. 2003), statistical translation model (Berger and Lafferty 1999), on the performance of sentence extraction tasks.

Diversity ranking may be applied in other application domains as well. For instance, social network analysis is one apparent domain to which graph-based ranking models can be directly applied. The earlier work by Zhu et al. (2007)

explored diversity ranking of actors based on country coverage and movie coverage as two diversity measurements. To our knowledge, there have not been many works investigating the issue of centrality and diversity ranking in other types of social networks. Citation ranking of academic papers could benefit from diversity ranking as more representative subfields can be ranked higher. Other possible application domains include review summarization where the customers may benefit from a diverse set of opinion about products and services, collaborative filtering (Singh et al. 2007), and automatic term recognition (Zhang et al. 2009).

Furthermore, Several trust and reputation ranking models have explored in various domains. For instance, EigenTrust (Kamvar et al. 2003), TrustRank (Gyöngyi et al. 2004), and propagation of trust and distrust (Guha et al. 2004) have been proposed in web search domain. We believe trust and credibility of content are other interesting directions for future work, especially in focused summarization and question answering domains. For example, an n-partite graph representation may include trust/distrust as another set of relations apart from similarity or redundancy of content. Ultimately, future research in sentence extraction, diversity, and trust may naturally converge to create an automated fact checker system.

BIBLIOGRAPHY

Achananuparp, P., Yang, C.C., and Chen, X. (2009) Using Negative Voting to Diversify answers in Non-Factoid Question Answering. In Proc. of CIKM 2009, Hong Kong.

Achananuparp, P., Hu, X., and Yang, C.C. (2009) Addressing the Variability of Natural Language Expression in Sentence Similarity with Semantic Structure of the Sentences. In Proc. of PAKDD 2009, Bangkok, 548-555.

Achananuparp, P., Hu, X., and Xiajiong, X. (2008) The evaluation of sentence similarity measures. In the Proceedings of 10th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2008), Turin, Italy.

Achananuparp, P., Hu, X., Zhou, X., and Zhang, X. (2008) Utilizing Sentence Similarity and Question Type Similarity to Response to Similar Questions in Knowledge-Sharing Community. In Proceedings of QAWeb 2008 Workshop, Beijing, China.

Achananuparp, P., Han, H., Nasraoui, O., and Johnson, R. (2007) Semantically enhanced user modeling. In Proceedings of the 2007 ACM Symposium on Applied Computing, ACM Press, New York, NY, 1335-1339.

Agrawal, R., Gollapaudi, S., Halverson, A., and Ieong, S. (2009) Diversifying Search Results. In Proc. of WSDM'09, 5-14.

Allan, J., Wade, C., and Bolivar, A. (2003) Retrieval and novelty detection at the sentence level. In Proc. of SIGIR '03, ACM, New York, NY, 314-321.

Al-Maskari, A., Sanderson, M., and Clough, P. (2007) The relationship between IR effectiveness measures and user satisfaction. In Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Amsterdam, The Netherlands, July 23 - 27, 2007). SIGIR '07. ACM, New York, NY, 773-774.

Andersen, R., Borgs, C., Chayes, J., Feige, U., Flaxman, A., Kalai, A., Mirrokni, V., and Tennenholtz, M. 2008. Trust-based recommendation systems: an axiomatic approach. In Proceeding of the 17th international Conference on World Wide Web (Beijing, China, April 21 - 25, 2008). WWW '08. ACM, New York, NY, 199-208.

Baker, C., Fillmore, F., Charles, J., and Lowe, J.B. (1998) The Berkeley FrameNet project. In Proceedings of the COLING-ACL, Montreal, Canada.

Balasubramanian, N., Allan, J., and Croft, W. B. (2007) A comparison of sentence retrieval techniques. In Proceedings of SIGIR '07, Amsterdam, the Netherlands, 813-814.

Banerjee, S., Ramanathan, K., and Gupta, A. (2007) Clustering short texts using wikipedia. In Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Amsterdam, The Netherlands, July 23 - 27, 2007

Banerjee, S., and Pedersen, T. (2003). Extended gloss overlap as a measure of semantic relatedness. In Proc. of IJCAI'03, Acapulco, 805-810.

Barzilay, R. and Elhadad, N. (2003) Sentence Alignment for Monolingual Comparable Corpora. In Proceedings of EMNLP 2003, Sapporo, Japan, 25-33.

Berger, A., Caruana, D., Cohn, D., Freitag, D., and Mittal, V. (2000) Bridging the lexical chasm: Statistical approaches to answer-finding. In Proceedings of SIGIR 2000, 222-229.

Berger, A. and Lafferty, J. (1999) Information retrieval as statistical translation. In Proceedings of SIGIR '99, 222-229.

Berkhin, P. (2005) A Survey of PageRank Computing. Internet mathematics, 2(1), 73-120.

Bilotti, M. W., Ogilvie, P., Callan, J., and Nyberg, E. (2007) Structured retrieval for question answering. In Proceedings SIGIR '07. ACM, New York, NY, 351-358.

Bian, J., Liu, Y., Agichtein, E., and Zha, H. (2008) Finding the right facts in the crowd: factoid question answering over social media. In WWW'08,467-476.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003) Latent dirichlet allocation. Journal of Machine Learning Research, 993-1022.

Boyce, B. (1982) Beyond topicality: A two stage view of relevance and the retrieval process. Info. Processing and Management, 18(3), 105-109.

Brandow, R., Mitze, K., and Rau, L.F. (1995) Automatic condensation of electronic publications by sentence selection. Inf. Process. Manage., 31(5), 675-685.

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7).

Budanitsky, A. and Hirst, G. (2006) Evaluating WordNet-based measures of semantic distance. Computational Linguistics, 32(1).

Burke, R. D., Hammond, K. J., Kulyukin, V. A., Lytinen, S. L., Tomuro, N., and Schoenberg, S. (1997) Question answering from frequently asked question files: Experiences with the FAQ finder system. Technical report.

Carbonell, J. and Goldstein, J. (1998) The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Proc. of SIGIR'98, 335–336.

Carterette, B., and Chandar, P. (2009) Probabilistic Models of Novel Document Rankings for Faceted Topic Retrieval. In Proc. of CIKM'09, November 2-6, Hong Kong, China.

Cao, X., Cong, G., Cui, B., Jensen, C. S., and Zhang, C. (2009) The use of categorization information in language models for question retrieval. In Proceeding of the 18th ACM Conference on information and Knowledge Management (Hong Kong, China, November 02 - 06, 2009). CIKM '09. ACM, New York, NY, 265-274.

Chen, S.Y., Huang, M.L., and Lu, Z.Y. (2009) Summarizing Documents by Measuring the Importance of a Subset of Vertices within a Graph. In Proc. of WI 2009.

Chen, H. and Karger, D. R. (2006) Less is more: Probabilistic models for retrieving fewer relevant documents. In 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 429–436.

Chklovski, T. and Pantel, P. (2004)VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04). Barcelona, Spain.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008) Novelty and diversity in information retrieval evaluation. In Proc. of SIGIR'08, 659-666.

Collobert, R. and Weston, J. (2007) Fast Semantic Extraction Using a Novel Neural Network Architecture. In Proceedings of ACL 2007, Prague, Czech Republic, June 23–30.

Corley, C. and Mihalcea, R. (2005) Measuring the semantic similarity of texts. In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, 13-18, Ann Arbor, Michigan, June.

Dagan, I., Glickman, O., and Magnini, B. (2005) The PASCAL recognising textual entailment challenge. In Proceedings of the PASCAL Workshop.

de Kerchove, C., and Dooren, P.V. (2008) The PageTrust algorithm: how to rank web pages when negative links are allowed? In Proc. SDM 2008 2008, 346-352.

Dolan, W., Quirk, C., and Brockett, C. (2004) Unsupervised construction of large paraphrase corpora: Exploiting massively parallel new sources. In Proceedings of the 20th International Conference on Computational Linguistics.

Erkan, G. and Radev, D. (2004) LexPageRank: Prestige in multi-document text summarization. In Proc. of EMNLP 2004.

Fellbaum, C. (1998) WordNet: An Electronic Lexical Database. MIT Press.

Francis,W. N. and Kucera, H. (1982) Frequency Analysis of English Usage: Lexicon and Grammar. Houghton Mifflin, Boston.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007) The third PASCAL recognizing textual entailment challenge. In Proc. of the Workshop on Textual Entailment and Paraphrasing, 1-9.

Gildea, D., Jurafsky, D. (2002) Automatic labeling of semantic roles. Computational Linguistics, 28(3), 245–288.

Goffman, W. (1964) A searching procedure for information retrieval. Info. Storage and Retrieval, 2, 73-78

Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. (2004) Propagation of trust and distrust. In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM, New York, NY, 403-412.

Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. (2004) Combating web spam with trustrank. In Proceedings of the Thirtieth international Conference on Very Large Data Bases - Volume 30 (Toronto, Canada, August 31 - September 03, 2004). M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, Eds. Very Large Data Bases. VLDB Endowment, 576-587.

Harabagiu, S., Miller, G. A., and Moldovan, D. I. (1999) WordNet 2 - A Morphologically and Semantically Enhanced Resource. Proceedings of ACL-SIGLEX99: Standardizing Lexical Resources, Maryland, June 1999, 1-8.

Hatzivassiloglou, V., Klavans, J., and Eskin, E. (1999) Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In Proceedings of EMNLP.

Hickl, A. and Bensley, J. (2007) A Discourse Commitment-Based Framework for Reconizing Textual Entailment. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, 171-176.

Hirst, G. & D. St-Onge (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, 305-332. Cambridge, Mass.: MIT Press.

Hoad, T. and Zobel, J. (2003) Methods for identifying versioned and plagiarized documents. Journal of the American Society of Information Science and Technology, 54(3), 203-215.

Hovy, E.H., Lin, C.Y., Zhou, L., and Fukumoto, J. (2006) Automated Summarization Evaluation with Basic Elements. In Proceedings of LREC. Genoa, Italy.

Hu, X., Sun, N., Zhang, C., and Chua, T. (2009a) Exploiting internal and external semantics for the clustering of short texts using world knowledge. In Proceeding of the 18th ACM Conference on information and Knowledge Management (Hong Kong, China, November 02 - 06, 2009). CIKM '09. ACM, New York, NY, 919-928

Hu, X., Zhang, X., Lu, C., Park, E. K., and Zhou, X. (2009b) Exploiting Wikipedia as external knowledge for document clustering. In Proceedings of the 15th ACM

SIGKDD international Conference on Knowledge Discovery and Data Mining (Paris, France, June 28 - July 01, 2009). KDD '09. ACM, New York, NY, 389-396.

Huang, M.L. and Chen, S.Y. (2009) Finding Representative and Diverse Vertices within Graphs. Technical Report, Tsinghua University, July.

Järvelin, K. and Kekäläinen, J. (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20, 4 (Oct. 2002), 422-446.

Jeh, G. and Widom, J. (2002) SimRank: a measure of structural-context similarity. In Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Edmonton, Alberta, Canada, July 23 - 26, 2002). KDD '02. ACM, New York, NY, 538-543.

Jeon, J. Croft, W. B., and Lee, J. H. (2005a) Finding semantically similar questions based on their answers. In SIGIR'05, 617-618.

Jeon, J. Croft, W. B., and Lee, J. H. (2005b) Finding similar questions in large question and answer archives. In CIKM'05, 84-90.

Jiang, J. J. and Conrath, D. W. (1998) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In Proceedings of the International Conference on Research in Compuation Linguistic, Taiwan.

Jijkoun, V. and de Rijke, M. (2005) Retrieving answers from frequently asked questions pages on the web. In CIKM'05, 76-83.

Jing, H. and McKeown, K. (2000) Cut and paste based summarization. In Proceedings of NAACL.

Joachims, T. (1998) Text categorization with support vector machines: Leaning with many relevant features. In Proceedings of European Conference on Machine Learning, 137-142.

Jurczyk, P. and Agichtein, E. (2007) Discovering authorities in question answer communities by using link analysis, In Proc. of CIKM 2007, November 06-10, 2007, Lisbon, Portugal.

Kamvar, S. D., Schlosser, M. T., and Garcia-Molina, H. 2003. The Eigentrust algorithm for reputation management in P2P networks. In Proceedings of the 12th international Conference on World Wide Web (Budapest, Hungary, May 20 - 24, 2003). WWW '03. ACM, New York, NY, 640-651.

Kelly, D. and Lin, J. (2007) Overview of the TREC 2006 ciQA Task. SIGIR Forum 41, 1 (June 2007), 107-116.

Kim, H. and Zhai, C. (2009) Generating comparative summaries of contradictory opinions in text. In Proceeding of the 18th ACM Conference on information and Knowledge Management (Hong Kong, China, November 02 - 06, 2009). CIKM '09. ACM, New York, NY, 385-394.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008) A Large-scale Classification of English Verbs, Language Resources and Evaluation Journal, 42(1), 21-40, Springer, Netherland.

Kleinberg, J.M. (1999) Authoritative sources in a hyperlinked environment, Journal of the ACM (JACM), 46(5), 604-632.

Kunegis, J., Lommatzsch, A., and Bauckhage, C. (2009) The Slashdot zoo: Mining a social network with negative edges. In Proc. of WWW 2009, 741-750.

Landauer, T. K. & Dumais, S. T. (1997) A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T.K., Laham, D., Rehder, B., and Schreiner, M.E. (1997) How Well Can Passage Meaning Be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans, in Proc. 19th Ann. Meeting of the Cognitive Science Soc., 412-417.

Leacock, C. and Chodorow, M. (1998) Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), WordNet: An electronic lexical database, MIT Press, Cambridge, MA., 265-283.

Lempel, R. and Moran, S. (2000) The Stochastic Approach for Link-Structure Analysis and the TKC Effect, Proc. of the 11th Int. Conf. on WWW 2000, 387–401.

Lesk, M. (1986) Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, in Proceedings of the 5th annual international conference on Systems documentation, 24 – 26.

Li, L., Xue, G.R., Zha, H., and Yu, Y. (2009) Enhancing Diversity, Coverage and Balance for Summarization through Structure Learning. In Proc. of WWW 2009, 71-80.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., and Crockett, K. (2006) Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering 18, 8 (Aug. 2006), 1138-1150.

Li, X., and Roth, D. (2002) Learning Question Classifiers. COLING'02, Aug.

Li, M. and Vitanyi, P. (1997) An Introduction to Kolmogorov Complexity and its Applications. Springer-Verlag.

Lin, C.Y. and Hovy, E.H. (2003) Automatic Evaluation of Summaries using n-Gram Co-occurrence Statistics. In Proceedings of the HLT2003 conference.

Lin, D. (1998) An Information-Theoretic Definition of Similarity. In Proceedings of the Fifteenth international Conference on Machine Learning, San Francisco, CA, 296-304.

Lin, J. (2007) Is Question Answering Better Than Information Retrieval? Toward a Task-Based Evaluation Framework for Question Series. In Proc. of NAACL HLT 2007, Rochester, NY, 212-219.

Lin, J., and D., Demner-Fushman (2005) Automatically Evaluating Answers to Definition Questions. In Proc. of HLT/EMNLP, Vancouver, 931-938.

Liu, Y., Bian, J., and Agichtein, E. (2008) Predicting Information Seeker Satisfaction in Community Question Answering. In Proc. of SIGIR'08, Singapore, July 20-24.

Long, C., Huang, M., Zhu, X., and Li, M. (2009) Multi-document Summarization by Information Distance. In Proceedings of the 2009 Ninth IEEE international Conference on Data Mining (December 06 - 09, 2009). IEEE Computer Society, Washington, DC, 866-871.

Lytinen, S. and Tomuro, N. (2002) The Use of Question Types to Match Questions in FAQFinder. In Papers from the 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, 46-53.

MacKinnon, I. and Vechtomova, O. (2008) Improving Complex Interactive Question Answering with Wikipedia Anchor Text. In Proc. of ECIR 2008, 438-445.

Malik, R., Subramaniam, V., and Kaushik, S. (2007) Automatically Selecting Answer Templates to Respond to Customer Emails. In Proceedings of IJCAI'07, Hyderabad, India, 1659-1664.

McHale, M. (1998) A Comparison of WordNet and Roget's Taxonomy for Measuring Semantic Similarity. Proc COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems. Montreal, Canada, August, 115-120.

Metzler, D., Bernstein, Y., Croft, W., Moffat, A., and Zobel, J. (2005) Similarity measures for tracking information flow. Proceedings of CIKM, 517–524.

Metzler, D. and Croft, W.B. (2005) Analysis of Statistical Question Classification for Fact-based Questions. In Information Retrieval, 8(3), 481-504.

Metzler, D., Dumais, S. T., and Meek, C. (2007) Similarity Measures for Short Segments of Text. In Proceedings of ECIR 2007, 16-27.

Mihalcea, R. and Tarau, P. (2004). TextRank: bringing order into texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain.

Mihalcea, R., Corley, C., and Strapparava, C. (2006) Corpus-based and Knowledge-based Measures of Text Semantic Similarity, in Proceedings of AAAI 2006, Boston, July.

Miller, G. A. & W. G. Charles (1991). Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1):1-28.

Murdock, V. and Croft, W.B. (2005) A Translation Model for Sentence Retrieval. In Proceedings of HLT/EMNLP '05, 684-691.

Murdock, V. (2006) Aspects of sentence retrieval. Ph.D. Thesis, University of Massachusetts.

Nenkova, A. and Vanderwende, L. (2005) The impact of frequency on summarization. MSR-TR-2005-101.

Otterbacher, J. and Radev, D. (2004) Comparing semantically related sentences: The case of paraphrase versus subsumption. In proc. of COLING 2004, August 23-27.

Otterbacher, J., Erkan, G., and Radev, D.R. (2005) Using Random Walks for Question-focused Sentence Retrieval. In Proc. of the HLT/EMNLP 2006, Vancouver, 915-922.

Otterbacher, J., Erkan, G., and Radev, D. R. (2009) Biased LexRank: Passage retrieval using random walks with question-based priors. Inf. Process. Manage. 45, 1 (Jan. 2009), 42-54.

Palmer, M., Gildea, D., and Kingsbury, P. (2005) The Proposition Bank: An Annotated Corpus of Semantic Roles. Comput. Linguist. 31, 1 (Mar. 2005), 71-106.

Park, Y., Byrd, R. J., and Boguraev, B. K. (2002) Automatic glossary extraction: beyond terminology identification. In Proceedings of the 19th international Conference on Computational Linguistics, Taipei, Taiwan, August 24 - September 01, 1-7.

Phan, X., Nguyen, L., and Horiguchi, S. (2008) Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceeding of the 17th international Conference on World Wide Web (Beijing, China, April 21 - 25, 2008). WWW '08. ACM, New York, NY, 91-100.

Ponzetto, S. P. and Strube, M. (2007) Knowledge Derived From Wikipedia for Computing Semantic Relatedness, Journal of Artificial Intelligence Research, 30, 181-212.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J.H., and Jurafsky, D. (2004) Shallow Semantic Parsing using Support Vector Machines, in Proceedings of HLT/NAACL-2004, Boston, MA, May 2-7.

Radev, D. (2000) A common theory of information fusion from multiple text sources, step one: Cross-document structure. In Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue.

Resnik, P. (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. International Joint Conference for Artificial Intelligence (IJCAI-95), 448-453.

Rubenstein, H. & J. Goodenough (1965). Contextual correlates of synonymy. Communications of the ACM, 8(10):627-633.

Sahami, M. and Heilman, T. (2006) A web-based kernel function for measuring the similarity of short text snippets. In Proceedings of WWW 2006, 377-386.

Seco, N., Veale, T., and Hayes, J. (2004) An intrinsic information content metric for semantic similarity in WordNet. In Proceedings of the 16th European Conference on Artificial Intelligence, Valencia, Spain, 23-27 August 2004, 1089-1090.

Shehata, S., Karray, F., and Kamel, M. (2007) A concept-based model for enhancing text categorization. In Proceedings of KDD '07. ACM, New York, NY, 629-637.

Singh, A. P., Gunawardana, A., Meek, C., and Surendran, A. C. (2007) Recommendations using Absorbing Random Walks. In Proceedings of NESCAI 2007.

Suryanto, M. A., Lim, E. P., Sun, A., and Chiang, R. H. (2009) Quality-aware collaborative question answering: methods and evaluation. In Proc. of WSDM '09. Barcelona, Spain, 142-151.

Tang, J., Yao, L., and Chen, D. (2009) Multi-topic based query-oriented summarization. In Proc. of The SIAM International Conference on Data Mining (SDM 2009).

Tatu, M. and Moldovan, D. (2007) COGEX at RTE3. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, 22-27.

Tomuro, N. (2003) Interrogative Reformulation Patterns and Acquisition of Question Paraphrases. In Proceedings of the Second international Workshop on Paraphrasing, 33-40.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005) Large Margin Methods for Structured and Interdependent Output Variables, Journal of Machine Learning Research (JMLR).

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. (2007) Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. Inf. Process. Manage. 43, 6 (Nov. 2007), 1606-1618.

Voorhees, E. (2001) Overview of TREC-9 Question Answering Track. In The Ninth Text Retrieval Conference (TREC-9), 71–80. NIST SP 500-249.

Wan, X. and 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 (Singapore, Singapore, July 20 - 24, 2008). SIGIR '08. ACM, New York, NY, 299-306.

Wan, X., Yang, J., and Xiao, J. (2006) Using Cross-Document Random Walks for Topic-Focused Multi-Document. In Proc. of 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), 1012-1018.

Wang, D., Li, T., Zhu, S., and Ding, C. (2008) Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization. In Proc. of SIGIR'08, July 20-24, Singapore, 307-314.

Wu, Z. and Palmer, M. (1994) Verb Semantics and Lexical Selection. In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics (ACL'94), 133-138, Las Cruces, New Mexico.

Xu,, Y. and Yin, H. (2008) Novelty and topicality in interactive information retrieval. Journal of the American Society for Information Science and Technology, 59(2):201–215.

Yang, B., Cheung, W., and Liu, J. (2007) Community Mining from Signed Social Networks. IEEE Trans. on Knowl. and Data Eng. 19, 10 (Oct. 2007), 1333-1348.

Yih, W. T., Goodman, J., Vanderwende, L., and Suzuki, H. (2007) Multi-document summarization by maximizing informative content-words. In Proceedings of IJCAI 2007.

Yue, Y. and Joachims, T. (2008) Predicting Diverse Subsets Using Structural SVMs, In Proceedings of ICML 2008.

Zhai, C. and Lafferty, J. (2006) A risk minimization framework for information retrieval. Info. Processing and Management, 42(1), 31-55.

Zhai, C., Cohen, W.W., and Lafferty, J. (2003) Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In Proc. of SIGIR'03, 10-17.

Zhang, Z., Xia, L., Greenwood, M. A. and Iria, J. (2009) Too Many Mammals: Improving the Diversity of Automatically Recognized Terms. In Proceedings of the International Conference on Recent Advances in Natural Language Processing 2009 (RANLP'09), Borovets, Bulgaria, September.

Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., and Ma, W. (2005) Improving web search results using affinity graph. In Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM, New York, NY, 504-511.

Zhang, D. and Lee, W. S. (2003) Question classification using support vector machines. In Proceedings of SIGIR 2003. ACM Press, New York, NY, 26-32.

Zhu, X., Goldberg, A., Van Gael, J., and Andrzejewski, D. (2007) Improving Diversity in Ranking using Absorbing Random Walks. In Proc. of NAACL-HLT 2007.

APPENDIX A: TREC9 QUESTION VARIANTS

Id	Question	Type
408	What kind of animal was Winnie the Pooh?	Definition
701	Winnie the Pooh is what kind of animal?	Definition
702	What species was Winnie the Pooh?	Definition
703	Winnie the Pooh is an imitation of which animal?	Definition
704	What was the species of Winnie the Pooh?	Definition
409	What's another name for aspartame?	Entity
705	Aspartame if also known as what?	Entity
706	What is a synonym for aspartame?	Entity
707	Aspartame is known by what other name?	Entity
708	Aspartame is also called what?	Entity
410	What does hazmat stand for?	Definition
709	Hazmat stands for what?	Definition
710	What is the definition of hazmat?	Definition
411	What tourist attractions are there in Reims?	Reference
711	What are the names of the tourist attractions in Reims?	Reference
712	What do most tourists visit in Reims?	Reference
713	What attracts tourists to Reims?	Reference
714	What are tourist attractions in Reims?	Reference
715	What could I see in Reims?	Reference
716	What is worth seeing in Reims?	Reference
717	What can one see in Reims?	Reference
412	Name a film in which Jude Law acted	Reference
718	Jude Law was in what movie?	Reference
719	Jude Law acted in which film?	Reference
720	What is a film starring Jude Law?	Reference
721	What film was Jude Law in?	Reference
722	What film or films has Jude Law appeared in?	Reference
413	Where are the U.S. headquarters for Procter & Gamble?	Location
723	What city houses the U.S. headquarters of Procter and Gamble	Location
724	Where is Procter & Gamble headquartered in the U.S.?	Location
725	What is the U.S. location of Procter & Gamble corporate offices?	Location
726	Procter & Gamble is headquartered in which U.S. city?	Location
727	Where is Procter & Gamble based in the U.S.?	Location
203	How much folic acid should an expectant mother get daily?	Degree
728	What is the recommended daily requirement for folic acid for pregnant women?	Degree
729	How much folic acid should a pregnant woman get each day?	Degree
730	What is the daily requirement of folic acid for an expectant mother?	Degree
731	What amount of folic acid should an expectant mother take daily?	Degree
201	What was the name of the first Russian astronaut to do a spacewalk?	Entity
732	Name the first Russian astronaut to do a spacewalk.	Entity
733	Who was the first Russian astronaut to walk in space?	Entity

734	Who was the first Russian to do a spacewalk?	Entity
415	What does CNN stand for?	Definition
735	CNN is the abbreviation for what?	Definition
736	CNN is an acronym for what?	Definition
416	When was CNN's first broadcast?	Time
737	What was the date of CNN's first broadcast?	Time
738	CNN began broadcasting in what year?	Time
739	CNN's first broadcast occurred on what date?	Time
740	When did CNN begin broadcasting?	Time
741	When did CNN go on the air?	Time
417	Who owns CNN?	Entity
742	Who is the owner of CNN?	Entity
743	CNN is owned by whom?	Entity
418	What is the name of a Salt Lake City newspaper?	Entity
744	What newspaper serves Salt Lake City?	Entity
745	Name a Salt Lake City newspaper.	Entity
419	Who was Jane Goodall?	Entity
746	What is Jane Goodall famous for?	Entity
747	What is Jane Goodall known for?	Entity
748	Why is Jane Goodall famous?	Entity
749	What made Jane Goodall famous?	Entity
421	What is thalassemia?	Definition
750	Define thalassemia.	Definition
751	What is the meaning of thalassemia?	Definition
752	How is thalassemia defined?	Definition
423	What soft drink contains the largest amount of caffeine?	Reference
753	What soft drink is most heavily caffeinated?	Reference
754	What is the most heavily caffeinated soft drink?	Reference
755	To get the most caffeine, what soda should I drink?	Reference
756	Which type of soda has the greatest amount of caffeine?	Reference
757	What soft drink would provide me with the biggest intake of caffeine?	Reference
424	What do you call a group of geese?	Reference
758	What is the collective term for geese?	Reference
759	What is the collective noun for geese?	Reference
760	What is the term for a group of geese?	Reference
761	What is the name given to a group of geese?	Reference
425	How many months does a normal human pregnancy last?	Time
762	What is the gestation period for human pregnancies?	Time
763	How long is human gestation?	Time
764	What is the gestation period for humans?	Time
765	A normal human pregnancy lasts how many months?	Time
426	What format was VHS's main competition?	Entity
766	What was the alternate to VHS?	Entity
767	What video format was an alternative to VHS?	Entity
768	What format was the major competition of VHS?	Entity

427	What culture developed the idea of potlatch?	Entity
769	What ethnic group introduced the idea of potlatch?	Entity
770	What is the cultural origin of the ceremony of potlatch?	Entity
771	Who developed potlatch?	Entity
428	Where is Logan International located?	Location
772	Where is Logan Airport?	Location
773	What city is Logan Airport in?	Location
774	Logan International serves what city?	Location
775	Logan International is located in what city?	Location
776	What city's airport is named Logan International?	Location
777	What city is served by Logan International Airport?	Location
429	What university was Woodrow Wilson president of?	Entity
778	Woodrow Wilson was president of which university?	Entity
779	Name the university of which Woodrow Wilson was president.	Entity
780	Woodrow Wilson served as president of what university?	Entity
431	What does CPR stand for?	Definition
781	What does the acronym CPR mean?	Definition
782	What do the initials CPR stand for?	Definition
783	CPR is the abbreviation for what?	Definition
784	What is the meaning of "CPR"?	Definition
433	Who was Darth Vader's son?	Definition
785	What was the name of Darth Vader's son?	Definition
786	What was Darth Vader's son named?	Definition
435	How did Bob Marley die?	Procedure
787	What caused the death of Bob Marley?	Procedure
788	What killed Bob Marley?	Procedure
789	What was the cause of Bob Marley's death?	Procedure
436	What instrument is Ray Charles best known for playing?	Entity
790	What instrument does Ray Charles play?	Entity
791	Musician Ray Charles plays what instrument?	Entity
792	Ray Charles plays which instrument?	Entity
793	Ray Charles is best known for playing what instrument?	Entity
437	What is Dick Clark's birthday?	Time
794	When was Dick Clark born?	Time
795	When is Dick Clark's birthday?	Time
796	What is Dick Clark's date of birth?	Time
440	Where was Poe born?	Time
797	What was Poe's birthplace?	Time
798	What was the birthplace of Edgar Allen Poe?	Time
799	Where is Poe's birthplace?	Time
441	What king was forced to agree to the Magna Carta?	Entity
800	What monarch signed the Magna Carta?	Entity
801	Which king signed the Magna Carta?	Entity
802	Who was the king who was forced to agree to the Magna Carta?	Entity
803	What king signed the Magna Carta?	Entity

804	Who was the king who signed the Magna Carta?	Entity
393	Where is your corpus callosum?	Location
805	Where is one's corpus callosum found?	Location
806	What part of your body contains the corpus callosum?	Location
807	The corpus callosum is in what part of the body?	Location
394	What is the longest word in the English language?	Reference
808	What English word has the most letters?	Reference
809	What English word contains the most letters?	Reference
810	What is the longest English word?	Reference
396	Who invented silly putty?	Entity
811	What is the name of the inventor of silly putty?	Entity
812	Silly putty was invented by whom?	Entity
813	Who was the inventor of silly putty?	Entity
397	When was the Brandenburg Gate in Berlin built?	Time
814	When was Berlin's Brandenburg gate erected?	Time
398	When is Boxing Day?	Time
815	What is the date of Boxing Day?	Time
816	What date is Boxing Day?	Time
817	Boxing Day is celebrated on what date?	Time
451	Where is McCarren Airport?	Location
818	What city does McCarren Airport serve?	Location
819	What city is served by McCarren Airport?	Location
820	McCarren Airport is located in what city?	Location
821	What is the location of McCarren Airport?	Location
822	Where is McCarren Airport located?	Location
452	Who created "The Muppets"?	Entity
823	Who invented "The Muppets"?	Entity
824	What was the name of "The Muppets" creator?	Entity
825	"The Muppets" was created by whom?	Entity
826	Name the creator of "The Muppets"	Entity
827	Who is the creator of "The Muppets"?	Entity
453	When is Bastille Day?	Time
828	What is the date of Bastille Day?	Time
829	Bastille Day occurs on which date?	Time
454	What is the Islamic counterpart to the Red Cross?	Entity
830	What is the equivalent of the Red Cross in the Middle East?	Entity
831	What is the name of the Islamic counterpart to the Red Cross?	Entity
832	Name the Islamic counterpart to the Red Cross.	Entity
833	What is the Islamic equivalent of the Red Cross?	Entity
834	What is the name given to the Islamic counterpart of the Red Cross?	Entity
455	What is Colin Powell best known for?	Reference
835	Colin Powell is most famous for what?	Reference
836	Colin Powell is best known for what achievement?	Reference
837	Who is Colin Powell?	Reference
838	Colin Powell is famous for what?	Reference

456	What is the busiest air travel season?	Time
839	What time of year do most people fly?	Time
840	What time of year has the most air travel?	Time
841	What time of year is air travel the heaviest?	Time
842	At what time of year is air travel at a peak?	Time
458	What's the name of a golf course in Myrtle Beach?	Entity
843	Name a golf course in Myrtle Beach.	Entity
442	What's the name of Pittsburgh's baseball team?	Entity
844	What is Pittsburg's baseball team called?	Entity
845	The major league baseball team in Pittsburgh is called what?	Entity
846	Name Pittsburgh's baseball team.	Entity
444	Where is the location of the Orange Bowl?	Location
847	What city is the Orange Bowl in?	Location
848	The Orange Bowl is in what city?	Location
849	The Orange Bowl is located in what city?	Location
850	Where is the Orange Bowl?	Location
445	When was the last major eruption of Mount St. Helens?	Time
851	When did Mount St. Helens last erupt?	Time
852	When did Mount St. Helen last have a major eruption?	Time
853	When did Mount St. Helen last have a significant eruption?	Time
446	What is the abbreviation for Original Equipment Manufacturer?	Reference
854	How do you abbreviate "Original Equipment Manufacturer"?	Reference
855	How is "Original Equipment Manufacturer" abbreviated?	Reference
448	Where is Rider College located?	Reference
856	Where can one find Rider College?	Location
857	What is the location of Rider College?	Location
858	Rider College is located in what city?	Location
859	Where is Rider College?	Location
449	What does Nicholas Cage do for a living?	Reference
860	What is the occupation of Nicholas Cage?	Reference
861	What is Nicholas Cage's profession?	Reference
862	What is Nicholas Cage's occupation?	Reference
450	What does caliente mean (in English)?	Definition
863	What does caliente translate to in English?	Definition
864	What is the English meaning of caliente?	Definition
865	What is the meaning of caliente (in English)?	Definition
866	What is the English translation for the word "caliente"?	Definition
400	What is the name of the Jewish alphabet?	Reference
867	What is the Jewish alphabet called?	Reference
868	The Jewish alphabet is called what?	Reference
869	The Jewish alphabet is known as what?	Reference
402	What nationality was Jackson Pollock?	Reference
870	Jackson Pollock was a native of what country?	Reference
871	Jackson Pollock is of what nationality?	Reference
872	What was the nationality of Jackson Pollock?	Reference

403	Tell me what city the Kentucky Horse Park is near?	Entity
873	The Kentucky Horse Park is close to which American city?	Entity
874	Where is the Kentucky Horse Park located?	Entity
875	Where is the Kentucky Horse Park?	Entity
876	What city is the Kentucky Horse Park near?	Entity
877	The Kentucky Horse Park is located near what city?	Entity
404	What is the state nickname of Mississippi?	Reference
878	What is a nickname for Mississippi?	Reference
879	Mississippi is nicknamed what?	Reference
880	Mississippi has what name for a state nickname?	Reference
881	What is the nickname for the state of Mississippi?	Reference
882	What is the nickname of the state of Mississippi?	Reference
405	Who used to make cars with rotary engines?	Entity
883	Rotary engines were manufactured by which company?	Entity
884	Who made the rotary engine automobile?	Entity
885	Rotary engine cars were made by what company?	Entity
886	Rotary engines used to be made by whom?	Entity
887	What company produced rotary engine vehicles?	Entity
406	What is the tallest mountain?	Entity
888	What is the world's highest peak?	Entity
889	What is the highest mountain in the world?	Entity
890	Name the highest mountain.	Entity
891	What is the name of the tallest mountain in the world?	Entity
407	What is Black Hills, South Dakota most famous for?	Reference
892	What makes Black Hills, South Dakota a tourist attraction?	Reference
893	What are the Black Hills known for?	Reference

APPENDIX B: QUESTION TAXONOMY USED IN QUESTION CLASSIFICATION

Source: http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/definition.html

Classes	Definition
ABBREVIATION	abbreviation
abb	abbreviation
exp	expression abbreviated
ENTITY	entities
animal	animals
body	organs of body
color	colors
creative	inventions, books and other creative pieces
currency	currency names
dis.med.	diseases and medicine
event	events
food	food
instrument	musical instrument
lang	languages
letter	letters like a-z
other	other entities
plant	plants
product	products
religion	religions
sport	sports
substance	elements and substances
symbol	symbols and signs
technique	techniques and methods
term	equivalent terms
vehicle	vehicles
word	words with a special property
DESCRIPTION	description and abstract concepts
definition	definition of something
description	description of sth.

manner	manner of an action
reason	reasons
HUMAN	human beings
group	a group or organization of persons
ind	an individual
title	title of a person
description	description of a person
LOCATION	locations
city	cities
country	countries
mountain	mountains
other	other locations
state	states
NUMERIC	numeric values
code	postcodes or other codes
count	number of something
date	dates
distance	linear measures
money	prices
order	ranks
other	other numbers
period	the lasting time of something
percent	fractions
speed	speed
temp	temperature
size	size, area and volume
weight	weight

APPENDIX C: DUC06 TASKS

Task Id	Topic Title	Narrative
D0601A	Native American Reservation System - pros and cons	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.
D0602B	steroid use among female athletes	Discuss the prevalence of steroid use among female athletes over the years. Include information regarding trends, side effects and consequences of such use.
D0603C	wetlands value and protection	Why are wetlands important? Where are they threatened? What steps are being taken to preserve them? What frustrations and setbacks have there been?
D0604D	anticipation of and reaction to the premier of Star Wars Episode I -- The Phantom Menace	How did fans, media, the marketplace, and critics prepare for and react to the movie? Include preparations and reactions outside the United States.
D0605E	treatment of osteoarthritis	Describe what procedures for treatment of osteoarthritis have been attempted and the result of research on these treatments.
D0606F	impacts of global climate change	What are the most significant impacts said to result from global climate change?
D0607G	civil unrest in China	Note examples of civil unrest in China and the Chinese government's policy toward and reaction to it. Specify the causes of the unrest.
D0608H	automobile safety	What devices and procedures have been implemented to improve automobile safety?
D0609I	Israeli West Bank settlements	What impact have Israeli settlements in the West Bank had on the Israeli/Palestinian peace process? What are the reactions of both parties and of the international community?
D0610A	home-schooling pros and cons	What are the advantages and disadvantages of home schooling? Is the trend growing or declining?
D0611B	organic methods of pest control	What methods or products are used to control pests for organic gardens or farms? Include information on methods of controlling such pests as insects or fungus which do not involve the use of chemical pesticides and are accepted by organizations which certify organic produce for the marketplace.
D0612C	recent developments and theories regarding autism	What are recent developments in autism diagnosis, treatment, and research? What is thought to be the cause? What services are available to patients and families? What is the frequency of occurrence?
D0613D	perceptions of Generation X	Who were the GenXers? What were their perceived habits, preferences, characteristics, and impact?
D0614E	Quebec independence	Describe developments in the movement for the independence of Quebec from Canada.
D0615F	evolution/creationism debate	What are the various perspectives in the U.S. public debate regarding the teaching of evolution, creation science, or intelligent design in public school science classes? What are the key points and counterpoints expressed by people who hold each of those perspectives?
D0616G	terrorist attacks in Chechnya	How have the Chechen militants elected to fight against the Russian government? What is the Russian response to the militancy and what toll is it taking on both sides?

D0617H	EgyptAir Flight 990	What caused the crash of EgyptAir Flight 990? Include evidence, theories and speculation.
D0618I	malaria prevention and treatment	What efforts are being made to combat the spread of malaria and to treat those currently affected?
D0619A	gays and the GOP	Discuss the relationship between gays (homosexuals) and the Republican party. How are Republicans courting gays? How do they alienate gays? Include discussion of the Log Cabin Republicans.
D0620B	school violence prevention measures	Discuss measures that schools and school districts have taken to prevent violent occurrences and shootings, such as those in Littleton, Colorado and Jonesboro, Arkansas.
D0621C	crime and law enforcement in China	Give examples of criminal activity in China. Name those involved, if possible. What is China doing to fight crime?
D0622D	spread of the West Nile virus	Track the spread of the West Nile virus through the United States and the efforts taken to control it.
D0623E	anti-smoking laws	Describe anti-smoking laws passed or rejected world-wide which prohibit smoking in public places or work places. Include any arguments used for or against such laws.
D0624F	Stephen Lawrence	What is known about the murder of Stephen Lawrence, his killers, the actions of the government, and the reactions of the public?
D0625G	types of diseases in Kenya	What are the most prevalent diseases in Kenya and and how are they affecting the population? What is being done to combat them?
D0626H	bombing of US embassies in Africa	How were the bombings of the US embassies in Kenya and Tanzania conducted? What terrorist groups and individuals were responsible? How and where were the attacks planned?
D0627I	international adoption	What are the laws, problems, and issues surrounding international adoption by American families?
D0628A	ADD/ADHD diagnosis and treatment	Describe ADD/ADHD. How is it diagnosed? What kind of treatments are there? Discuss the controversies surrounding its treatment.
D0629B	computer viruses	Identify computer viruses detected worldwide. Include such details as how they are spread, what operating systems they affect, what damage they inflict, their country of origin, and their creators wherever possible.
D0630C	bookselling	What is the current status of bookselling? What challenges face traditional sellers? How are booksellers associations involved? How successful is online bookselling and how has it affected traditional sellers?
D0631D	crash of the Air France Concorde	Discuss the Concorde jet, its crash in 2000, and aftermaths of this crash.
D0632E	Mongolia's foreign relations	What is the extent and nature of Mongolia's diplomatic and economic relations with other countries?
D0633F	U.S. crime trends	Which crime categories have had increasing or decreasing trends nationally in the U.S? Which geographic areas of the U.S. have had increasing or decreasing trends for particular crime categories?
D0634G	Pacific salmon conservation	What conservation measures are being taken locally and nationally to save the salmon species in the Pacific Northwest? What is the nature of Canadian-U.S. relations on Pacific salmon fishing?
D0635H	capital punishment in	How has the administration of Governor George W. Bush

	Texas during Governor Bush's administration	implemented capital punishment and how are those policies viewed outside of Texas?
D0636I	issues between the UAW and American automobile manufacturers	What are the key issues under discussion between the 3 major American automobile manufacturers and the United Auto Workers (UAW)?
D0637A	solar energy around the world	Provide reasons for using solar energy. How widespread internationally is its use and development? Discuss cooperation between nations. Which nations are the leaders in solar energy development?
D0638B	NASA's Galileo Mission	How successful was NASA's Galileo space probe mission of Jupiter? What discoveries were made about the planet and its moons? Include details about when the probe was launched and any troubles it may have encountered.
D0639C	precursor chemicals for weapons of mass destruction (WMDs)	What precursor chemicals are used in making WMDs? What other uses do the chemicals have, if any? Where have the chemicals been found? What controls are placed on WMDs and/or their precursor chemicals?
D0640D	Kursk disaster	Discuss the sinking of the Russian submarine Kursk, the attempts to save it, and salvage operations.
D0641E	global warming	Describe theories concerning the causes and effects of global warming and arguments against these theories.
D0642F	Hugo Chavez	What have been the key policies and outcomes (good or bad) of the Venezuelan Presidency of Hugo Chavez? What supportive or critical statements or actions have come from Venezuelans or leaders of other countries?
D0643G	El Nino and La Nina weather condition	Describe the causes and effects of the El Nino and La Nina weather condition. What programs and scientific techniques are in effect to better predict and cope with the conditions?
D0644H	federal budget surplus	What factors led to the federal budget surplus? What are the expectations for future surpluses? What have been proposed for use of the surplus?
D0645I	need for low-income housing	What are the problems facing low-income Americans in the housing market? How are the problems being addressed?
D0646A	perjury crime and punishment	What is the definition of perjury? Is it a federal offense? How common is it and what kinds of punishments have been given?
D0647B	Elian Gonzales custody battle	Describe the custody battle between Cuban and US relatives of the boy Elian Gonzales. Include details about how he came into the custody of his US relatives, the legal and international issues, and the resolution of the situation.
D0648C	obsessive-compulsive disorder	What are signs of obsessive-compulsive disorder (OCD)? What treatments have been tried and what has been effective? What other disorders are related to OCD?
D0649D	election of Vladimir Putin in 2000	Who is Vladimir Putin, including his experience and background? What led up to his election as Russian President, and what were his actions between his election and swearing in?
D0650E	former President Carter's international activities	Describe former President Carter's international efforts including activities of the Carter Center.

APPENDIX D: DUC07 TASKS

Task Id	Topic Title	Topic Narrative
D0701A	Southern Poverty Law Center	Describe the activities of Morris Dees and the Southern Poverty Law Center.
D0702A	art and music in public schools	Describe the state of teaching art and music in public schools around the world. Indicate problems, progress and failures.
D0703A	steps toward introduction of the Euro	Describe steps taken and worldwide reaction prior to introduction of the Euro on January 1, 1999. Include predictions and expectations reported in the press.
D0704A	Amnesty International	What is the scope of operations of Amnesty International and what are the international reactions to its activities? Give examples of charges lodged by the organization and complaints against it.
D0705A	Basque separatism	Describe developments in the Basque separatist movement 1996-2000.
D0706B	Burma government change 1988	What are the main events and important personalities in Myanmar (formerly Burma) leading up to and since the government changed in September 1988?
D0707B	Turkey and the European Union	What positive and negative developments have there been in Turkey's efforts to become a formal member of the European Union?
D0708B	world-wide chronic potable water shortages	What countries are having chronic potable water shortages and why?
D0709B	Angelina Jolie	What have been the most recent significant events in the life and career of actress Angelina Jolie?
D0710C	Israel / Mossad "The Cyprus Affair"	Two alleged Israeli Mossad agents were arrested in Cyprus. Determine why they were arrested, who they were, how the situation was resolved and what repercussions there were.
D0711C	Microsoft's antitrust problems	Summarize Microsoft's antitrust problems, including its alleged illegal behavior and antitrust proceedings against the company.
D0712C	Salman Rushdie	Summarize events related to the "death sentence" on Salman Rushdie proclaimed by Iran.
D0713C	Pakistan and the Nuclear Non-Proliferation Treaty	What has Pakistan's demonstrated behavior been toward the Nuclear Non-Proliferation Treaty? Include Pakistan's explanation for the behavior and international reaction to it.
D0714D	Napster	Describe the legal battle between various recording artists and members of the record industry and the Internet music site Napster. What support, or lack thereof, have the litigants received?
D0715D	International Land Mine Ban Treaty	Which countries have signed the Ottawa Treaty for the elimination of anti-personnel land mines, and how many have ratified it? What countries have refused to sign, and why? How effective has the treaty been?
D0716D	Jabiluka Uranium Mine	Describe the development of Australia's uranium mine project in its Kakadu National Park and the protests and obstacles encountered.
D0717D	fen-phen lawsuits	Describe the various lawsuits against American Home Products which resulted from the use of fenfluramine, also

		known as Pondimin, and half of the diet drug combination called "fen-phen".
D0718D	Starbucks Coffee	How has Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions, or subsidiaries?
D0719E	unemployment in France in the 1990s	Describe the unemployment situation in France in the 1990s. Discuss social consequences of this situation, Identify possible causes of the unemployment situation and policies proposed to support jobless people and to reduce unemployment.
D0720E	Oslo Accords	Identify the principles of the Oslo Accord of 1993. Describe what happened in subsequent years in attempts to implement these principles?
D0721E	Matthew Shepard's death	Provide details on the murder of Matthew Shepard in 1998. Discuss the culprits and their trials. Describe public reaction to the murder, including legislation proposed as a result of the murder.
D0722E	US missile defense system	Discuss plans for a national missile defense system. Include information about system costs, treaty issues, and technical criticisms. Provide information about test results of the system.
D0723F	Senator Dianne Feinstein	Describe Dianne Feinstein's election to the US Senate and her accomplishments while serving as a member of the Senate.
D0724F	obesity in the United States	Describe the extent of obesity in the United States and possible causes for US obesity.
D0725F	Iran's nuclear capability	Describe Iran's nuclear capabilities and nuclear testing. Also relate concerns of other countries about Iranian nuclear capabilities and their attitudes regarding development, testing, and deployment of Iranian nuclear capabilities.
D0726F	Al Gore's 2000 Presidential campaign	Give the highlights of Al Gore's 2000 Presidential campaign from the time he decided to run for president until the votes were counted.
D0727G	Newt Gingrich's divorce	Describe the charges, counter charges and legal settlement actions involved in Newt Gingrich's divorce.
D0728G	Interferon	Describe the drug Interferon, its uses, effectiveness, patient tolerance and side effects.
D0729G	Eric Rudolph	What crimes have been attributed to Eric Rudolph? What efforts have been made to capture him and how has he eluded capture?
D0730G	line item veto	What has been the argument in favor of a line item veto? How has it been used? How have US courts, especially the Supreme Court, ruled on its constitutionality?
D0731G	Linda Tripp	What role did Linda Tripp play in the Clinton/Lewinsky affair and the Ken Starr investigation?
D0732H	Kenya education developments	Kenya is attempting to raise its economic status. One approach is to raise the educational level of its population. What developments are there in this approach?
D0733H	public programs at Library of Congress	The Library of Congress is available to the public for research. What additional programs and attractions are available?
D0734H	acupuncture treatment in U.S.	It appears that acupuncture treatment is being increasingly accepted in the U.S. for medical problems. How is

		acupuncture being integrated into the American healthcare system and what are its applications?
D0735H	reintroduction program for wolves in U.S.	Note the current situation with the program to reintroduce endangered wolves. What problems exist and what are the prospects for success?
D0736H	Oprah Winfrey TV show	Note the various subjects and controversial incidents on Oprah's show 1998-2000.
D0737I	deep water exploration	What is being learned from the study of deep water, seabeds, and deep water life? What equipment and techniques are used? What are plans for future related activity?
D0738I	mining in South America	What is the status of mining in central and South America? Include obstacles encountered.
D0739I	after "Seinfeld"	What became of the cast and others related to the "Seinfeld" TV series after it ended? What actions were taken by others in response to the show's closing?
D0740I	round-the-world balloon flight	Report on the planning, attempts and first successful balloon circumnavigation of the earth by Bertrand Piccard and his crew.
D0741I	day trader killing spree	Give the background on Atlanta day trader Mark O. Barton's killing spree and the aftermath.
D0742J	John F. Kennedy, Jr., dies in plane crash	Write an account of the sequence of events involving the Kennedy family during and following the plane crash that killed John F. Kennedy, Jr., his wife, and his sister-in-law.
D0743J	earthquakes in Western Turkey in August 1999	Two massive earthquakes occurred in Turkey in 1999. Describe the rescue efforts and the impact of the earthquakes on the economy, society, etc., in Turkey.
D0744J	organic food	Describe the developments in the growth of the organic food industry, U.S. government efforts to set standards, and the public's attitude toward food (especially organic and genetically altered or biotech foods) in 1999 and 2000.
D0745J	OJ Simpson developments	Give an account of the developments in the life of OJ Simpson in the years 1999 and 2000.

APPENDIX E: CIQA06 TASKS

Task Id	Question
26	The analyst is particularly interested in knowing the volume of smuggled VCDs and also the ruses used by smugglers to hide their efforts.
27	The analyst would like to know of efforts to curtail the transport of drugs from Mexico to the U.S. Specifically, the analyst would like to know of the success of the efforts by local or international authorities.
28	The analyst is interested in evidence of transport of goods from Syria to Iraq under the food-for-oil program.
29	The analyst wants to know if there is evidence of transporting human cargo from China to the U.S. and where the ships arrive in the U.S.
30	What is the extent of illegal immigration from Albania to Italy and what steps are being taken to curb it?
31	The analyst is interested in South African arms support to Pakistan and the effect such support or sales has on relations of both countries with India. Additionally, the analyst would like to know what nuclear arms involvement, if any, exists between South Africa and Pakistan.
32	The analyst is concerned about universities which do research on medical subjects slanting their findings, especially concerning drugs, towards drug companies which have provided money to the universities.
33	The analyst is especially interested in opinions of scientists as to whether there is a family link between dinosaurs and birds, and what evidence they cite concerning their opinions.
34	The analyst would like to know if there exists any financial relationship between the Israeli government and the PNA, in particular, is there any evidence of money transfers between these entities?
35	The analyst would like to know what financial relationships exist between Greece and Cyprus. This is intended to include trade between the two countries as well as direct financial grants.
36	The analyst is interested in anything related to financial support from the American government or public for the IRA, Sinn Fein, and any similar group.
37	The analyst wants any information regarding the relationship between the extremist Philippine rebel group Abu Sayyaf and the MILF, a renegade faction of the mainstream rebel group, the Moro National Liberation Front, which signed a peace agreement with the Philippine government in 1996.
38	The analyst would like to know if obesity, when not genetic, is triggered by deep-seated emotional problems or depression. Specifically, does the problem vanish when the underlying cause has been determined?
39	The analyst would like to know to what extent second-hand smoke affects others in proximity to the smoker. Specifically, does it cause lasting and irreparable harm/damage to the non-smoker?
40	The analyst would like to know what influence Title IX had on college wrestling. Title IX was a component of the Federal 1972 Educational Amendments that barred schools receiving federal funds from discrimination on the basis of gender in athletics and other programs.
41	The analyst is interested in the effects and consequences of the use of steroids on athletes' performance and health.
42	The analyst is interested in the effect of aspirin on coronary heart disease and stroke. Specifically, what does aspirin do and how does it do it?

43	The analyst wants to see evidence of perceived impact on Arab-Israeli relations associated with the discovery of natural gas resources in Gaza.
44	The analyst is interested in knowing whether the chairman merely favors reducing tax rates or if he is for or against other features of the tax code.
45	What is John McCain's position toward Jerry Falwell's Moral Majority and Pat Robertson's Christian Coalition? Does he support the organizations or does he oppose them?
46	The analyst would like to know what positions the Saudi Arabian Government has held relative to Osama bin Laden as a political activist. Specifically, the analyst wants to know what positions with respect to him have been expressed by members of the Saudi Government, as well as government-proclaimed positions.
47	The analyst is interested in any actions taken by the United States in reaction to the Mad Cow crisis in England. This can include changes in policy, testing, USDA regulations, and any interactions with the EU in general.
48	The analyst is interested in Saudi Arabia's intentions regarding foreign workers. Specifically, the analyst wants to know how Saudi Arabia views foreign workers and how they treat them.
49	The analyst would like to know how Richard Seed felt about human cloning. Specifically, the analyst would like to know what his feelings were regarding human cloning and what actions he took as a result.
50	The analyst desires to know the evidence presented by the prosecution and what favors were supposedly given by Espy in return for the alleged gratuities.
51	What evidence is there that baseball's famous home run hitting record holder Mark McGwire used illegal, performance-enhancing substances which gave him an advantage over other players? What were the substances?
52	What evidence is there to support the involvement of Christie's chief executive Christopher M. Davidge in a conspiracy by Christie's and Sotheby's to illegally fix buyer and setter fees at the two auction houses?
53	There has been accusations at the U.S. House of Representative hearing that China removes and sells organs from the executed prisoners and even removes organs from the convicts on the death roll, or time the execution to suit the special needs of organ transplants. What evidence is there for or against these accusations?
54	North Korea is so impoverished that it desperately needs to court foreign governments for hard cash and humanitarian aid. But if it opens its doors to foreigners, the totalitarian regime risks eroding its own authority. Is there evidence that North Korea has resorted to counterfeiting foreign currencies to alleviate this problem?
55	The analyst would like to know if there is evidence that Charles Taylor, President of Liberia, was personally involved in diamond smuggling in Sierra Leone. Specifically, the analyst would like to know what evidence exists regarding Taylor's involvement in diamond smuggling.

VITA

EDUCATION

2004 - 2010	Ph.D. in Information Science, Drexel University, Philadelphia PA GPA: 4.0/4.0
2002 - 2004	Master of Information Systems Management (MISM), Carnegie Mellon University, Pittsburgh PA.
1996 - 2000	Bachelor of Economics (2nd class honor), Chulalongkorn University, Bangkok, Thailand

RESEARCH INTERESTS

Topics	Text and web mining, social network mining, text summarization, question answering, information extraction, information retrieval

TEACHING

Fall 2008	**Information Retrieval Systems** The iSchool at Drexel Instructor
Summer 2008	**Information Retrieval Systems** The iSchool at Drexel Teaching Assistant working with Chris Yang

SELECTED PUBLICATIONS

Answer Diversification for Complex Question Answering on the Web, Palakorn Achananuparp, Xiaohua Hu, Tingting He, Christopher C. Yang, Yuan An, and Lifan Guo. In Proc. of PAKDD 2010 [13.1% Acceptance Rate]

Using Negative Voting to Diversify Answers in Non-Factoid Question Answering, Palakorn Achananuparp, Christopher C. Yang, Xin Chen. In Proc. of CIKM 2009, Hong Kong. [20% Acceptance Rate]

Probabilistic Models for Topic Learning from Images and Captions in Online Biomedical Literatures, Xin Chen, Caimei Lu, Yuan An, Palakorn Achananuparp. In Proc. of CIKM 2009, Hong Kong. [15% Acceptance Rate]

Addressing the Variability of Natural Language Expression in Sentence Similarity with Semantic Structure of the Sentences, Palakorn Achananuparp, Xiaohua Hu, Christopher C. Yang. In Proc. of PAKDD 2009, Bangkok, Thailand, 548-555. [33% Acceptance Rate]

The Evaluation of Sentence Similarity Measures, Palakorn Achananuparp, Xiaohua Hu, Xiajiong Chen. In Proc. of DaWak 2008, Turin, Italy, 305-316. [33% Acceptance Rate]