Sentence Level Information Patterns

for Novelty Detection

A Dissertation Presented
by

XIAOYAN LI

Submitted to the Graduate School of the
University of Massachusetts at Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

July 2006
Computer Science

Sentence Level Information Patterns

for Novelty Detection

A Dissertation Presented
by

XIAOYAN LI

Approved as to style and content by:

_____
W. Bruce Croft, Chair

_____
James Allan, Member

_____
Andrew McCallum, Member

_____
Donald Fisher, Member

_____

W. Bruce Croft, Department Head

Department of Computer Science

DEDICATION

To my patient and loving parents, husband and daughters.

ACKNOWLEDGMENTS

I would like to thank my advisor, Distinguished Professor W. Bruce Croft, for his many years of thoughtful, patient and inspiring guidance and support. Thanks are also due to Professor James Allan and all the members of the Center for Intelligent Information Retrieval (CIIR) at UMass. Together their friendship and selfless contributions to my professional development have been invaluable to the rest of my career. I would also like to extend my gratitude to other members of my committee, Professor Andrew McCallum and Professor Donald Fisher for their helpful comments and suggestions on all stages of my PhD research.

A special thank you to all the sisters and brothers of the Amherst Chinese Christian Church for their love, encouragement and prayers.

This work was supported in part by the Center for Intelligent Information Retrieval, by SPAWARSYSCEN-SD grant numbers N66001-99-1-8912 and N66001-1-8903, and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsors.

ABSTRACT

SENTENCE LEVEL INFORMATION PATTERNS FOR NOVELTY DETECTION

JULY 2006

XIAOYAN LI, B.E. TSINGHUA UNIVERSITY

M.E., TSINGHUA UNIVERSITY

Ph.D. UNIVERSITY OF MASSACHUSETTS AT AMHERST

Directed by: Professor W. Bruce Croft

The detection of new information in a document stream is an important component of many potential applications. In this thesis, a new novelty detection approach based on the identification of sentence level *information patterns* is proposed. Given a user's information need, some information patterns in sentences such as combinations of query words, sentence lengths, named entities and phrases, and other sentence patterns, may contain more important and relevant information than single words. The work of the thesis includes three parts. First, we redefine "*what* is novelty detection" in the lights of the proposed information patterns. Examples of several different types of information patterns are given corresponding to different types of uses' information need. Second, we analyze *why* the proposed information pattern concept has a significant impact in novelty detection. A thorough analysis of sentence level information patterns is elaborated on data from the TREC novelty tracks, including sentence lengths, named entities (NEs), and sentence level opinion patterns. Finally, we present *how* we perform novelty detection based on information patterns, which focuses on the identification of previously unseen query-related patterns in sentences. A unified pattern-based approach is presented to novelty detection for both specific NE topics and more general topics. Experiments on novelty detection were carried out on data from the TREC 2002, 2003 and 2004 novelty tracks. Experimental results show that the proposed approach significantly improves the performance of novelty detection for both specific and general topics, therefore the overall performance for all topics, in terms of precision at top ranks. Future research directions are suggested.

Keywords: Novelty detection, information patterns, named entities

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXAMPLES

LIST OF OBSERVATIONS AND CONCLUSIONS

Chapter 1

PROBLEM STATEMENT

1.1 What is Novelty Detection

"An information retrieval system is an information system, that is, a system used to store items of information that need to be processed, searched, retrieved and disseminated to various user populations" [44]. Information retrieval is often referred as document retrieval. The basic task of document retrieval is to retrieve documents that are relevant to a user's request or information need. The output of a traditional document retrieval system is a ranked list of documents. Documents are ranked by the relevance scores that are calculated by the system. The system assumes that a document with a higher relevance score is more likely to be relevant to the user's request than a document with a lower relevance score. The relevance of a document is assumed to be independent of other documents.

Novelty detection can be viewed as going a step further than traditional document retrieval. Based on the output of a document retrieval system (i.e., a ranked list of documents), a novelty detection system will further extract documents with new information from the ranked list. The purpose of the research on novelty detection is to provide a user with a list of text passages that both are relevant and contain new information with respect to the user's information need. The goal is for the user to quickly get useful information without going through a lot of redundant information, which is usually a tedious and time-consuming task.

The detection of new information is an important component in many potential applications. It is a new research direction, but it has attracted more and more attention in the information retrieval field. The TREC novelty tracks, which are related to novelty detection, have run for three years

[6, 23, 42]. Many research groups have participated in the TREC novelty tracks. Novelty detection is also a very important component in the current DARPA-sponsored Global Autonomous Language Exploitation (GALE) project. The output of a distillation engine for GALE should consist of "snippets" of English text for each query. The snippets, which may consist exact text extractions, translations, summarizations or paraphrases of the source material, should be marked either "new" or "redundant" to indicate whether or not the relevant information in a snippet is already provided by another snippet. Therefore novelty detection, i.e., new information detection, is an important part of the GALE project.

Novelty detection can be performed at two different levels: the event level and the sentence level. At the event level, a novel document is required to not only be relevant to a topic (i.e., a query) but also to discuss a new event. At the sentence level, a novel sentence should be relevant to a topic and provide new information. This means that the novel sentence may either discuss a new event or provide new information about an old event. Novelty detection at the sentence level is also the basis for novelty detection at the event level. The work in this thesis focuses on novelty detection at the sentence level. Given a query and a set of sentences, the set of sentences is first classified into relevant sentences and non-relevant sentences. Among relevant sentences, some sentences are marked as novel sentences because they provide new information and others are redundant sentences without new information (see Figure 1). The relevance of a sentence given a query is independent of other sentences. However, the novelty of a sentence depends on the previous sentences. In short, the task of a novelty detection system at the sentence level is to provide a user with a list of novel sentences given a query.

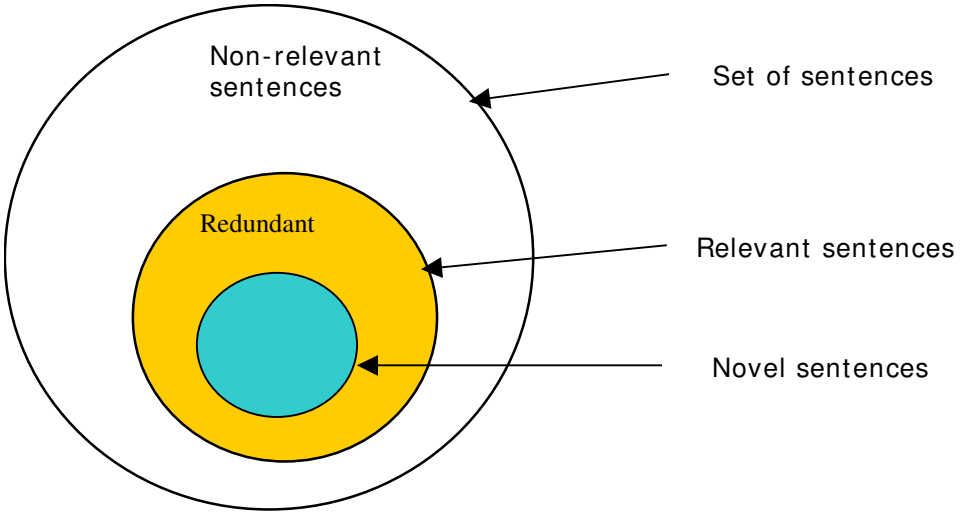

Figure 1. Given a query, the sets of relevant, non-relevant, novel and redundant sentences

1.2. Motivation of the Thesis

A variety of novelty measures have been described in the literature [6, 7, 13, 23, 36]. The various definitions of novelty, however, are quite vague and seem only indirectly related to the intuitive notions of novelty. For example, in [13, 36], novelty detection was informally defined as the opposite of redundancy on the same scale. In [7] it was described as new information based on new words existing in a sentence. Whereas in [6, 23], no clear definition of novelty was provided. Usually, new words appearing in an incoming sentence or document contribute to the novelty scores in various novelty measures in different ways. However, new words are not equivalent to novelty (new information). For example, rephrasing a sentence with a different vocabulary does not mean that this revised sentence contains new information that is not covered by the original sentence. Furthermore, new words appearing in a material that is not relevant to a user's query do not provide any new information with respect to the user's interest, even though it may contain

some new information useful to other users. A more precise definition of novelty or new information is required.

Let us look at a real example to understand the limitations of the related work and why we are proposing a new approach for novelty detection. Topic (query) 306 from the TREC novelty track 2002 is about "African Civilian Deaths". The user is asking for the number of civilian non-combatants that have been killed in the various civil wars in Africa. Therefore a number should appear in sentences that are relevant to the query. Let us consider the following four sentences given in Example 1.1.

Example 1.1. Topic: "African Civilian Deaths"

<Number>: 306

<Title> African Civilian Deaths

<Description>: How many civilian non-combatants have been killed in the various civil wars in Africa?

<Narrative>: A relevant document will contain specific casualty information for a given area, country, or region. It will cite numbers of civilian deaths caused directly or indirectly by armed conflict.

Sentence 1 (Relevant): "It could not verify Somali claims of more than **100** *civilian deaths***"**.

Sentence 2 (Relevant): "Natal's *death* toll includes another massacre of **11** ANC [*African* National Congress] supporters".

Sentence 3 (Non-relevant): "Once the slaughter began, following the *death* of President Juvenal Habyarimana in an air crash on April 6, hand grenades were thrown into schools and churches that had given refuge to Tutsi *civilians*."

Sentence 4 (Non-relevant): "A Ghana News Agency correspondent with the West *African* force said that rebels loyal to Charles Taylor began attacking the *civilians* shortly after

the peace force arrived in Monrovia last Saturday to try to end the eight-month-old civil war."

Each of the four sentences has two terms (in *italic*) that match the key words from the query. However, only the first two sentences are relevant sentences. In addition to the two matching words, each of the first two sentences also has a number, 100 and 11 (in bold and underlined), respectively. Hence, the first two sentences are both topically relevant to the query and have the right type of information with respect to the user's request behind the query. The third sentence and the fourth sentence are not relevant to the query mainly because they do not contain numbers that are required by the user.

For the example query given above, it is very difficult for traditional word-based approaches to separate the two non-relevant sentences (sentence 3 and sentence 4) from the two relevant sentences (sentence 1 and sentence 2). Even worse, the two non-relevant sentences are very likely to be identified as novel sentences simply because they contain many new words that do not appear in previous sentences.

To solve this problem, a deeper query understanding beyond a few key words and a more precise novelty understanding beyond new words are required. This motivated our work on information pattern-based novelty detection. We believe that *information patterns* such as combinations of query words, named entities, phrases and other sentence patterns, which indicate the presence of possible answers, may contain more important and relevant information than single words given a user's request or information need.

As the first attempt, we investigate if identifying query-related named-entities (NEs) patterns in sentences will significantly improve the performance in novelty detection, particularly at top ranks. The NE pattern-based approach is inspired by question answering techniques and is similar to passage retrieval for *factoid* questions. A typical question answering system consists of three

components: query analysis, passage retrieval and answer extraction [19,20]. First, keywords of the topic related to a question are extracted and expected answer type is determined in the process of query analysis. Second, passages with possible answers are retrieved, usually with a modified information retrieval system. Last, possible answers in the retrieved passages are extracted and ranked by their relevance scores, which are usually a combination of passage score and answer score.

The basic concept of our NE-pattern-based approach is that each query could be treated as multiple *specific* NE questions. Each question is represented by a few query words, and it requires a certain type of named entities as answers. Instead of extracting exact answers as in factoid question answering systems [14,19,20], we propose to first extract interesting sentences with certain NE patterns that include both query words and required answer types, indicating the presence of potential answers to the questions, and then to identify novel sentences that are more likely to have new answers to the questions. A *new definition of novelty* and novelty detection based on finding the right answers for the underlying questions using NE-patterns is the first major, original contribution of this work. A deeper understanding of novelty should enable an improvement of performance of novelty-based systems.

However, queries (or topics) that can be transformed into specific NE questions are only a small portion of the possible queries. For example, in TREC 2003, there are only 15 (out of 50) topics that can be formulated into specific NE questions. The rest of the topics will be called general topics throughout the thesis, since they can only be formulated into *general* questions. *New and effective information patterns* are needed in order to significantly improve the performance of novelty detection for those general topics. This will be the second major contribution of this work.

As one of the most interesting sentence level information patterns, we have found that the detection of information patterns related to opinions, i.e., *opinion patterns*, is very effective in

improving the performance of the general topics. As an example, Topic N1, from the TREC novelty track 2003, is about "partial birth abortion ban". This is a query that cannot be easily converted into any specific NE questions. However, we know that the user is trying to find opinions about the proposed ban on partial birth abortions. Therefore, relevant sentences are more likely to be "opinion sentences". Let us consider the following two sentences.

Example *1.2. Topic: "Partial Birth Abortion Ban"*

<Number>: N1

<Title>: partial birth abortion ban

<Toptype>: opinion

<Description>: Find opinions about the proposed ban on partial birth abortions.

<Narrative>: Relevant information includes opinions on partial birth abortion and whether or not it should be legal. Opinions that also cover abortion in general are relevant. Opinions on the implications of proposed bans on partial birth abortions and the positions of the courts are also relevant.

Sentence 1 (Relevant and Novel): "The court's ruling confirms that the entire campaign to *ban 'partial-birth abortion*' -- a campaign that has consumed Congress and the federal courts for over three years -- is nothing but a fraud designed to rob American women of their right to *abortion,***" said** Janet Benshoof, president of Center for Reproductive Law and Policy.

Sentence 2 (Non-relevant): Since the Senate's last *partial birth* vote, there have been 11 court decisions on the legal merits of *partial birth bans* passed by different states.

Both sentence 1 and sentence 2 have five matched terms (in *italic*). But only sentence 1 is relevant to the topic. Note that in addition to the matched terms, sentence 1 also has opinion

patterns, indicated by the word "said" and a pair of quotation marks (both in bold and underlined). The topic is an opinion topic that requires relevant sentences to be opinion sentences. The first sentence is relevant to the query because it is an opinion sentence and therefore topically related to the query. However, for the topic given in the above example, it is very difficult for traditional word-based approaches to separate the non-relevant sentence (sentence 2) from the relevant sentence (sentence 1). This work tries to attack this hard problem.

For a useful novelty detection system, a unified framework of the pattern-based approach is also required to deal with both the specific and the general topics. We propose a unified pattern-based approach that includes the following three steps: query analysis, relevant sentence retrieval and new pattern detection. The *unified pattern-based approach* is the third major contribution of this work.

1.3. Contributions of the Thesis

In summary, this work has made the following three main original contributions:

1. We provide a new and more explicit definition of novelty. Novelty is defined as *new answers to the potential questions* representing a user's request or information need.
2. We propose a new concept in novelty detection - query-related *information patterns*. Very effective information patterns for novelty detection at the sentence level have been identified.
3. We propose a *unified pattern-based approach* that includes the following three steps: query analysis, relevant sentence detection and new pattern detection. The unified approach works for both specific topics and general topics.

1.4. Organization of the Thesis

The rest of the work is organized as follows. Chapter 2 gives a review of related work on novelty detection. Chapter 3 describes data collections from TREC novelty tracks (2002, 2003 and 2004), with examples. Evaluation measures used in TREC and the ones we are using are also discussed in the chapter. Chapter 4 introduces our new definition of "novelty", and the concept of the proposed information patterns for novelty detection of both specific and general topics. Chapter 5 elaborates a thorough analysis of sentence level information patterns including sentence lengths, named entities, and opinion patterns. The analysis is performed on the data from the TREC 2002 and/or 2003 novelty tracks, which provides guidelines in applying those patterns in novelty detection. Chapter 6 describes the proposed unified pattern-based approach to novelty detection for both general and specific topics. The different treatments for specific topics and general topics will be highlighted. Chapter 7 shows experimental results in using the proposed information patterns for significantly improving the performance of novelty detection for specific topics, general topics, and the overall performance of novelty detection for all topics using the unified approach. Chapter 8 summarizes the work, discusses several remaining issues, and indicates some future research directions.

Chapter 2

NOVELTY DETECTION: A REVIEW OF RELATED WORK

Novelty detection has been conducted at two different levels: the event level and the sentence level. In the following we give a review of research related to novelty detection at the two different levels, as well in other applications. Then, we point out the main differences between the pattern-based approach proposed in this thesis and other approaches in the literature.

2.1. Novelty Detection at the Event Level

Work on novelty detection at the event level arises from the Topic Detection and Tracking (TDT) research, which is concerned with online new event detection and/or first story detection [1,2,3,4,5,16,18,31]. Current techniques for new event detection are usually based on clustering algorithms, which include two important issues: event modeling and event clustering. Several *models*, such as vector space models, language models, lexical chains, etc., have been proposed to represent incoming new stories/documents. Each story is then grouped into *clusters*. An incoming story will either be grouped into the closest cluster if the similarity score between them is above the preset similarity threshold or it will be used to start a new cluster. A story which starts a new cluster will be marked as the first story about a new topic, or it will be marked as "old" (about an old event) if there exists a novelty threshold (which is smaller than the similarity threshold) and the similarity score between the story and its closest cluster is greater than the novelty threshold.

Temporal proximity was also considered in [4, 16] based on the observation that news stories discussing the same event tend to be temporally proximate. It is possible that these techniques

may treat two stories both as first stories for a new event if the two stories are actually about the same event with one focusing on the cause of the event and another focusing on the consequences of the event. This is because the similarity score is based on the full text of the two stories, not the seminal events that these two stories discuss. Therefore, document-model based approaches may not work well for the detection of different events from the same topics.

Yang et al [2] proposed a two-level scheme first story detection system. At the first level, the system classifies incoming stories/documents into predefined broad topics. At the second level, the system performs the "first story detection" within each topic. The limitation of this technique is that it needs to predefine the taxonomy of broad topics and to have training data for each broad topic, which makes it inapplicable in the TDT context.

Kumaran and Allan [39] proposed a multi-stage new event detection (NED) system similar to [2]. Stories were classified into categories and NED was performed within categories. They built three vector representations for each story. The main vector that is used for clustering consists of all terms except stopwords in the story. The other two vectors, which consist of named entity terms and non-named entity terms, respectively, were used as additional features to confirm their initial decisions on stories with the main vector.

Strokes et al [3] proposed a composite document representation, which combined a free text vector and a lexical chain created using WordNet, and achieved a marginal increase in effectiveness.

2.2. Novelty Detection at the Sentence Level

Research on novelty detection at the sentence level is related to the TREC novelty tracks [6, 23, 42] described in more detail in the next chapter. The goal is to find relevant and novel sentences,

given a query and an ordered list of relevant documents. Many research groups participated in the TREC 2002, 2003 and 2004 novelty tracks and reported their results [7, 8, 9, 10, 11, 12, 13, 23, 32, 33, 34, 35, 36].

Novelty detection at the sentence level can be conducted mainly in two steps: relevant sentence retrieval and novel sentence extraction. In current techniques developed for novelty detection at the sentence level, new words appearing in sentences usually contribute to the scores that are used to rank sentences. Many similarity functions used in information retrieval are also used in novelty detection. Usually a high similarity score between a sentence and a given query will increase the relevance rank of the sentence, whereas a high similarity score between the sentence and all previously seen sentences will decrease the novelty ranking of the sentence.

The simplest novelty measure, *New Word Count Measure* [7], simply counts the number of new words appearing in a sentence and takes it as the novelty score for the sentence. There are other similar novelty or redundancy measures that consider new words appearing in sentences. One such novelty measure, called New Information Degree (NID), was defined by Jin et al [33]. The authors provided two ways to compute the NID value. One way was to take the sum of the inverse document frequency (IDF) values of new words appearing in a sentence and divided it by the sum of the IDF values of all the terms in the sentence. IDF is the reciprocal of the frequency of a word in a collection of documents. The IDF weight is usually taken as log (N/n) +1, where N is the size of the collection, and n is the number of documents that contain the words. The second way used the fraction of the new bi-gram word sequences over all the bi-gram word sequences in a sentence as the NID value of the sentence.

Zhang et al [36] used an overlap-based redundancy measure in their work on novelty detection. The redundancy score of a sentence against another sentence is the number of matching terms normalized by the length of the sentence. They set the redundancy threshold to 0.55. Sentences

with redundancy scores higher than the threshold were treated as redundant sentences and thus eliminated.

Many similarity functions (such as cosine similarity, language model measures etc.) used in information retrieval have also been tried in novelty detection. Tsai, Hsu and Chen [32] represented sentences as vectors and computed similarity scores for a sentence currently being considered with each previous sentence with the cosine similarity function for detecting novel sentences. The Maximal Marginal Relevance model (MMR), introduced by Carbonell and Goldstein [24], was used by Sun et al [34] in their work on novelty detection.

Instead of counting new words appearing in sentences, Litkowski [35] checked all discourse entities in a sentence. Discourse entities are semantic objects and they can have multiple syntactic realizations within a text. A pronoun, a noun phrase and a person's name could all belong to one single entity and they would be treated as the same discourse entity. The sentence was accepted as novel if it had new discourse entities that were not found in the growing history list of discourse entities. Eichmann et al [38] considered the number of new named entities and noun phrases appearing in a sentence. A sentence was declared novel if the number of new named entities and noun phrases is above a pre-declared number.

2.3. Novelty Detection in Other Applications

Novelty detection can be also seen in other applications, such as temporal summaries of news topics, document filtering and minimal document set retrieval, where new information detection is an important component. Those applications could include novelty detection at the event level, novelty detection at the sentence level, or the combination of both.

In Allan et al's work [57] on temporal summaries of new topics, the task is to extract a single sentence from each event within a news topic, where the stories are presented one at a time and sentences from a story must be ranked before the next story can be considered. Novelty is an important characteristic of sentence selection. They proposed two approaches for measuring novelty. The first approach estimates the probability that two sentences discuss the same event by the probability that the later sentence could arise from the same language model as the earlier sentence. The second approach is the same as the first approach except that the sentence is compared to clusters and there is more information in a cluster to estimate probabilities. The second measure gives better performance than the first approach, which indicates the clustering is useful for modeling the events.

In Zhang et al's work [13] on novelty and redundancy detection in adaptive filtering, the goal is to eliminate redundant documents. A redundant document is defined as the one that is relevant to the user's profile, but only contains information that is covered by previous documents. Five redundancy measures were proposed and evaluated in their experiments. The five measures are set difference measure, cosine similarity metric, and three KL-divergence-based measures with respect to three different language modeling variations: language modeling with Dirichlet smoothing, language modeling with shrinkage smoothing, and a three-component mixture language models. The experimental results showed that both the cosine similarity metric and a redundancy measure based on a mixture of language models were effective for identifying redundant documents. While redundancy and novelty were treated as two ends of a same scale by Zhang et al in their work [13], they are actually different measures. Redundancy measures how much redundant information (information already covered by previous documents) a document may contain and novelty measures how much new information (information relevant to a query and uncovered by the previous documents) the document contains.

Other research that performs novelty detection is Zhai et al's work [17] on subtopic retrieval. The goal of subtopic retrieval is to find documents that cover as many different subtopics as possible. They studied both novelty and redundancy in the language modeling framework and presented two ways to measure the novelty of a document. One was based on the KL-divergence, and another was based on a simple two-component mixture model similar to the three-component mixture model in [13]. The simple mixture model was reported the most effective.

In the recent work by Dai et al [40] on minimal document set retrieval, the authors tried three different approaches to retrieve and rank document sets with maximum coverage but minimal redundancy of subtopics in each set. The first approach was a novelty-based algorithm, which generated documents sets from relevant documents. It starts with an initial document set of top ranked document and adds documents one at a time. The Cosine measure was used to calculate the similarity score between a document and each document already in the current document set. The maximum similarity score of a document was defined as the redundancy score of the document. A document was added to the document set if and only if its redundancy score was less than a pre-learned novelty threshold. The second approach was a cluster-based generation algorithm, which grouped similar documents into clusters and picked a document from each cluster to generate ranked document sets. The third approach was a subtopic extraction based algorithm. It explicitly extracted subtopics from retrieved documents and used those subtopics as dimensions to generate ranked document sets. A subtopic was represented with a cluster of n-gram phrases. The ranked list of documents was re-ranked for each subtopic. The document with the largest relevant value was picked as a document set seed. Then the coverage of the document set for each subtopic was calculated. If the coverage of the set for a subtopic was less than a preset threshold, then the top ranked document from this subtopic's ranking list was picked and added to the document set. The process was repeated until all subtopics were covered. The third approach gave the best performance according to their experiments.

2.4. Comparison of Our Approach and Others

In this thesis, novelty detection at the sentence level is studied. A new definition of novelty is introduced and a unified information patterns based approach was proposed. Both the definition of novelty and the proposed approach can be extended to novelty detection at the event level.

There are three main differences between our pattern based approach and the aforementioned approaches in the literature. First, none of the work described above gives an explicit definition of novelty, which we believe is essential in novelty detection. Usually new words, phrases or named entities appearing in incoming sentences or documents contribute to the novelty score, though in different ways. Our pattern-based approach is based on the new definition of novelty introduced in this thesis, which treats new information as *new answers* to questions that represented users' information requests.

Second, in the aforementioned systems that are related to the TREC novelty tracks, either the title query or all the three sections of a query were used merely as a bag of words, whereas we try to understand a query beyond bag-of-words by transforming it into multiple specific NE questions or a general question. We believe that a user's information request can be better captured with questions than a few keywords, and a deeper understanding of the user's request helps a novelty detection system to serve the user better.

Third, the proposed information pattern based approach is inspired by question answering techniques and is similar to passage retrieval for factoid questions. A typical question answering system consists of three components: query analysis, passage retrieval and answer extraction. The query analysis component first extracts keywords of the topic related to a question and determines the expected answer type of the question, which is usually a certain type of named entities for NE questions. The passage retrieval component, which may be modified from an information

retrieval system, then retrieves passages with possible answers. Last, the answer extraction component extracts possible answers in the retrieved passages and ranks them by their relevance scores, which are usually a combination of passage score and answer score.

In our approach, each query could be treated as multiple questions; each question is represented by a few query words, and it requires a certain type of named entities as answers. Answer sentences are defined as sentences with answers to the multiple questions representing a user's information request. Novel sentences are sentences with new answers to those questions. Instead of explicitly extracting exact answers as in factoid question answering systems [14,19,20], we propose to first retrieve answer sentences with certain information patterns that include both query words and required answer types, indicating the presence of potential answers to the questions, and then identify novel sentences that are more likely to have new answers to the questions.

Chapter 3

TREC NOVELTY TRACK DATA AND EVALUATION

Currently, the three sets of data officially available for experiments on novelty detection at the sentence level are from TREC novelty tracks [6, 23, 42]. The TREC novelty track was first introduced in 2002 and ran through 2004. The direct motivation for these tracks and hence the name "novelty track" came from CMU Professor Jaime Carbonnell's invited talk to the *Automatic Summarization Workshop* of the North American Chapter of the Association for Computational Linguistics (NAACL) in May 2001 [6]. In his talk, he mentioned that one of the ways to optimize search results was to rank sentences/documents on their "novelty of information" measurements, instead of merely using relevance ranking alone. The basic task of the novelty tracks was then defined as the following (rephrased by the author):

> *Given a topic and an ordered set of relevant documents that are segmented into sentences, return sentences that are both relevant to the topic and contain new information that are not covered by previous sentences.*

The novelty tracks are an effort to get beyond the relevant-ranked list output for information retrieval. In this chapter, we will first describe the three data collections from TREC 2002, 2003 and 2004 novelty tracks. Then we will summarize some of their key statistics that are going to be used throughout the thesis, and discuss the evaluation methods that TREC novelty tracks use and the ones we are using.

3.1. Data Collection from TREC 2002 Novelty Track

The data from TREC 2002 novelty track includes 50 topics and an ordered list of up to 25 relevant documents for each topic. All documents were pre-segmented into sentences.

Example 3.1 below is a topic from TREC 2002 novelty track. The topic is about antibiotics ineffectiveness and its ID is 449. Each topic has three main fields to describe the topic: title, description and narrative. They are followed by an ordered list of relevant document IDs. The text of the relevant documents associated with each topic is stored in a separate file. The *title* field gives the key words of the topic from a user. The key words are "antibiotics ineffectiveness" for topic 449. The *description* field actually describes the information request from the user. The user wants to know "what has caused the current ineffectiveness of antibiotics against infections and what is the prognosis for new drugs?" The description of a topic is a very important field because different users may have different requests even though they use the same key words. The last field is the *narrative* field. It indicates that what kinds of information a document or sentence must contain to be relevant to the topic. For topic 449 in Example 3.1, it requires that a relevant document must discuss the reasons or causes for the ineffectiveness of current antibiotics.

Example 3.1. A Topic in TREC 2002 and its relevant document IDs

Topic: Antibiotics Ineffectiveness

<num>Number: 449

<title>: antibiotics ineffectiveness

<desc>Description: What has caused the current ineffectiveness of antibiotics against infections and what is the prognosis for new drugs?

<narr>Narrative: To be relevant, a document must discuss the reasons or causes for the ineffectiveness of current antibiotics. Relevant documents may also include efforts by

pharmaceutical companies and federal government agencies to find new cures, updating current testing phases, new drugs being tested, and the prognosis for the availability of new and effective antibiotics.

<relevant>: (**Author's note: 25 relevant document IDs)**

FBIS4-23026

FT933-7438

FT924-5307

FBIS4-23024

.

.

.

FT921-16061

FT943-9461 (**Author's note: the text of this document is listed below**)

FT922-12251

.

.

Document: "FT943-9461" (**Author's note: the text of all the documents associated with a topic is stored in a file, only one document is listed here.**)

<DOCNO>

<s docid="FT943-9461" num="1"> FT943-9461</s>

</DOCNO>

<PROFILE>

<s docid="FT943-9461" num="2"> _AN-EHKC0ABYFT</s>

</PROFILE>

<DATE>

<s docid="FT943-9461" num="3"> 940811</s>

</DATE>

<HEADLINE>

<s docid="FT943-9461" num="4"> FT 11 AUG 94 / Cancer charity sells drug rights</s>

</HEADLINE>

<BYLINE>

<s docid="FT943-9461" num="5"> By DANIEL GREEN</s>

</BYLINE>

<TEXT>

<s docid="FT943-9461" num="6"> The Cancer Research Campaign, a medical charity, has joined with Xenova, a biotechnology company, to develop a new cancer drug.</s> <s docid="FT943-9461" num="7"> Cancer Research Campaign Technology, the charity's technology transfer arm, has sold the worldwide marketing rights to the drug Daca in return for cash payments and royalties should the drug be marketed.</s> <s docid="FT943-9461" num="8"> The payments are based on the progress of the drug through trials up to a maximum of Pounds 1.7m.</s> <s docid="FT943-9461" num="9"> If clinical trials go to plan, the drug could be approved at the end of the decade.</s> <s docid="FT943-9461" num="10"> Stockbroker Lehman Brothers estimates that it may eventually generate Dollars 250m in sales.</s> <s docid="FT943-9461" num="11"> Daca is still in the early stages of development.</s> <s docid="FT943-9461" num="12"> Less than a quarter of drugs make it to the market from this stage.</s> <s docid="FT943-9461" num="13"> CRC and Xenova, which is based in Slough, Berkshire, are taking the development forward because laboratory tests showed that low doses of Daca overcame two main types of drug resistance - a common cause of failure of chemotherapy - in various tumours, including colon, skin and lung cancers.</s> <s docid="FT943-9461" num="14"> Mr Louis Nisbet, chief executive of Xenova, said he may resell certain marketing rights, such as those in Asia, to large drug companies to secure more funding for the final, most expensive, clinical trial stages.</s> <s docid="FT943-9461" num="15"> The compound was discovered by scientists at the Auckland Cancer Society in New Zealand and handed to the British charity for further investigation.</s> <s docid="FT943-9461" num="16"> Dr Paul Bevan, Xenova's research director, said the drug worked by blocking two enzymes, topoisomerase I and II, involved in the uncontrolled replication of cancer cells.</s> <s docid="FT943-9461" num="17"> Existing anti-cancer drugs of this type hit only one or other of the two.</s> <s docid="FT943-9461" num="18"> The effect is to kill rapidly dividing cells in tumours.</s> <s docid="FT943-9461" num="19"> Such an approach is already used in chemotherapy drugs which kill cells by other means, so side-effects such as hair loss, infection and anaemia are also possible with Daca.</s> <s docid="FT943-9461" num="20"> Daca appears to overcome drug resistance that arises when cancer cells are able to pump out chemotherapy drugs before they can

work.</s> <s docid="FT943-9461" num="21"> The deal is one of a series CRC has done with the private sector.</s> <s docid="FT943-9461" num="22"> But it is more unusual for Xenova, whose collaborations in the past have been with companies such as Genentech and academic centres such as Purdue University of Indiana, rather than charities.</s> <s docid="FT943-9461" num="23"> Xenova, which specialises in rapid screening of large numbers of potential drugs derived from fungi, bacteria and plants, floated its shares on the Nasdaq market in the US last month.</s>

The text of all the documents associated with a topic is stored in a file. One example document is listed above. The document was marked with XML format. The start and end of a sentence were marked with <s docid= “xxx” num=“yyy”> and </s>, respectively. xxx here indicates the document ID of the document a sentence belongs to, and yyy indicates the sequence of the sentence in the document.

The 50 topics for the TREC 2002 novelty track were selected from the full set of 150 topics (topics 300-450) from TRECs 6, 7 and 8 according to two criteria. First, each topic should have between 10 and 70 relevant documents. Second, among its relevant documents, each topic should not have a large number of Federal Register documents, which tend to be very long. The criteria were based on the considerations of having enough relevant documents to work with but not too many for humans to annotate in creating the “ground-truth” data. Judgment data was created by NIST (National Institute of Standards and Technology) assessors who manually annotated the data. They first created a file of relevant sentences and then generated a file of novel sentences. Each topic was independently judged by two different assessors. The judgment data from the assessor who marked the smallest number of relevant sentences, on a per topic basis, was used as the ground truth data. The maximum number of documents was determined to be no more than 25 after the NIST staff tried this important but quite tedious job [6]. Only 49 topics were finally used as the test set in this thesis because there are no relevant sentences marked for topic 310.

3.2. Data Collection from TREC 2003 Novelty Track

The data from the TREC 2003 novelty track [23] also includes 50 topics and an ordered list of up to 25 relevant documents for each topic. All documents were pre-segmented into sentences. The main difference between the TREC 2003 topics and TREC 2002 topics is that the TREC 2003 topics were classified into event topics and opinion topics. Among the 50 topics from TREC 2003 novelty track, there were 22 opinion sentences and 28 event topics. The type of a topic was given in the *topic-type* (“*toptype”)* field. An event topic example is shown in Example 3.2, and an opinion topic is shown in Example 3.3.

Example 3.2. An event topic from TREC 2003

<Num>: Number: N1

<Title>: World Cup soccer

<toptype>: event

<desc>Description: The World Cup Soccer event of 1998.

<narr>Narrative: Any mention of participating nations, preparation for, attendance of, behavior/misbehavior of soccer game fans, planning and preparation for the game, locations, final results and closing ceremonies of the 1998 World Cup Soccer activities were relevant. Mention of any other soccer game was not relevant. Discussion of 'robotic' W.C. Soccer using robotics and artificial languages were irrelevant.

Example 3.3. An opinion topic from TREC 2003

<num>Number: N1

<title>: Partial Birth Abortion Ban

<toptype>: opinion

<desc>Description: Find opinions about the proposed ban on partial birth abortions.

<narr>Narrative: Relevant information includes opinions on partial birth abortion and whether or not it should be legal. Opinions that also cover abortion in general are relevant. Opinions on the implications of proposed bans on partial birth abortions and the positions of the courts are also relevant.

3.3. Data Collection from TREC 2004 Novelty Track

The data from the TREC 2004 novelty track [42] also includes 50 topics, as in the cases for the TREC 2002 and 2003 novelty tracks. The difference is that, in addition to an ordered list of up to 25 relevant documents for each topic, there are zero or more non-relevant documents for each topic. As usual, all documents were pre-segmented into sentences. Similar to TREC 2003 novelty track, the 50 topics were also classified into event topics and opinion topics. There were 25 event topics and 25 opinion topics in the TREC 2004 novelty track.

The major change of TREC 2004 novelty track is the inclusion of irrelevant documents into the document set. This means that a topic from TREC 2004 novelty track may have some non-relevant documents in addition to 25 relevant documents in the ordered list. The document collection for the TREC 2004 novelty track is also unique in that it contains three news sources from overlapping time period: New York Times New Service (Jun 1998 – Sep 2000), AP (also Jun 1998 – Sep 2000), and Xin Hua News Service (Jun 1996 – Sep 2000). As a result, this collection exhibits greater redundancy than other TREC collections, and thus less novel information, increasing the difficulty of novelty detection in this data set.

3.4. Statistics of the Three Novelty Track Collections

Table 3.1 summarizes the statistics of the three data collections from TREC 2002, 2003 and 2004 novelty tracks, respectively, in terms of number of queries (# of topics), number of event topics, number of opinion topics, number of all sentences in the collection, and number of relevant, non-relevant and novel sentences.

Table 3.1. Statistics of the collections from TREC 2002, 2003 and 2004 novelty tracks

Data collections	# of topics	# of event topics	# of opinion topics	# of sentences	# of relevant sentences	# of novel sentences	# of non-relevant sentences
TREC2002	49	N/A	N/A	57227	1365	1241	55762
TREC2003	50	28	25	39820	15557	10226	24263
TREC2004	50	25	25	52447	8343	3454	44104

There are three main differences among the three sets.

(1). The TREC 2003 and 2004 novelty track collections exhibited greater redundancy than the TREC 2002 and thus has less novel sentences [23]. Only 41.4% and 65.7% of the total relevant sentences were marked novel for the TREC 2004 novelty track and the TREC 2003 novelty track, respectively, whereas 90.9% of the total relevant sentences in the 2002 track are novel sentences.

(2). TREC 2003 and TREC 2004 topics were classified into event topics and opinion topics. This allowed these two types of topics to be treated differently for performance improvement.

(3). TREC 2002 and 2004 novelty track collections exhibited greater difficulty in terms of relevant sentence retrieval because there were only a small portion of sentences in the document set marked relevant sentences. TREC 2002 and TREC 2004 novelty tracks had 1365 and 8363 relevant sentences, respectively, which were only 2.4% and 15.9% of the total number of sentences. TREC 2003 novelty track had 15557 relevant sentences, which was 39.1% of the total number of sentences in the TREC 2003 novelty track document collection.

All three data sets are used in the experiments of this thesis. The data from the TREC 2002 novelty track exhibits least redundancy and was not suggested for testing by TREC. But it is still considered in our experiments. The reason is that we believe a good novelty detection system should not hurt the performance even though there is less redundancy in a data collection. Furthermore, it is usually not known in advance that how much redundancy a collection has with respective to a query in real applications.

3.5. Evaluation Methods

The TREC novelty tracks used three evaluation measures for a novelty detection system: set precision, set recall and the F measure. The F measure is the primary evaluation measure for the system. The sentences selected manually by the NIST assessors were considered the ground truth data. Relevant and novelty sentence retrieval results were evaluated separately [6, 23, 42].

Let M be the number of matched sentences, i.e., the number of sentences that are selected by both the assessors and the novelty detection system. Let A be the number of sentences selected by the assessors, and S be the number of sentences retrieved by the system. Then the sentence set recall (R) of the system is defined as M/A, and the sentence set precision (P) of the system is defined as M/S. The F measure (adopted from van Rijsbergen's E measure [50]) is a function of the set precision and recall, which is defined as

$$F_b = \frac{(b^2+1)PR}{b^2P+R} \tag{3.1}$$

where the parameter b determines the relative importance of the recall and the precision of the system. The measurement F_1, i.e., the F measure with a b value of 1, was used for performance evaluation in the TREC novelty tracks. This indicates equal importance of precision and recall.

One problem with set recall and set precision is that they do not average well when the sizes of assessors' sets vary widely across topics. Harman [6] illustrated the problem with an example, where a system did precisely the wrong thing but got an average score of 0.5 for both recall and precision. The average of the F measure is more meaningful than the average of set precision or recall in the case that the sizes of the judgment sets vary largely. In addition, the F measure also combines precision and recall into one number, which is more convenient for users to compare the performance of systems with different values. However, systems with an equal value of the F measures can have a range of precision and recall scores. Hence, the F measure is not very precise in characterizing users' real requirements, such as a quick search on the Web for obtaining useful information within top ranked results.

Usually, the number of relevant (or novel) sentences retrieved by the system is determined by their ranking scores. Those sentences whose ranking scores are larger than a threshold are returned. This indicates that the numbers of sentences for different queries may vary dramatically, and the difference between numbers of sentences for two topics can be more than a hundred. However, in real applications, a user may only care about how much information within a limited number of sentences, say 10 or 20, retrieved by a system. For instance, many users who search information on the Web only look at top 10 results in the first page returned from a search engine. Few users would read the results beyond first two pages.

On the other hand, the pattern-based approach we propose is a *precision-oriented* approach to meet the afore-mentioned requirements from users. In improving the precision of relevance and novelty detection, some relevant and novel sentences are filtered out. For example, in dealing with topics that can be turned into multiple specific questions, only those sentences that "answer" these identified specific questions are selected as relevant; other relevant sentences that are not covered by these specific questions are filtered out. This is also true for novelty detection in that a sentence is treated as novel only if it contains new named entities that answer the specific

questions. Since we apply very strict rules in the selection process, ideally, those sentences that are retrieved by our system are truly relevant sentences or novel sentences therefore the precision of the relevant and novel sentence detection should be high. However, the recall might be low for the same reason.

Therefore, in this thesis, we simply use precision of relevant or novel sentences at top ranks, instead of applying the commonly used set precision, set recall or the F measure. We believe that precision values at top ranks (top 5, 10, 15, 20 or 30 sentences) are more useful in real applications where users want to quickly get some required information without going through a lot of non-relevant as well as redundant information. Note that we will compare the performance of our system with other baselines with the same evaluation measures.

The relevance and novelty judgment files used for evaluation were generated by human assessors, who manually annotated the data. Each topic was independently judged by two different assessors. Since the judgments were somewhat subjective, there were disagreements between the two assessors. The major factors that caused the disagreements include the interpretation of topics and relative strictness. For the TREC 2002 novelty track, the average coverage on relevance was about 0.579. The assessor who picked the smaller number of relevance sentences is designated as the "official assessor". The 0.579 average coverage means that the other assessor, who picked the larger number of relevant sentences, only picked about 57.9% of the relevant sentences that were picked by the official assessor. The difference on relevant sentence selection from two assessors resulted in an even lower coverage on novel sentences. The topics and judgments for the TREC 2003 novelty track were much improved over the TREC 2002 novelty track. The average coverage information was not given for the TREC 2003 data. However, the difference between two different assessors in terms of numbers of selected relevant and novel sentences was much smaller than that of the TREC 2002 data. The topics for the TREC 2004 novelty track were comparable in quality to the TREC 2003 topics, particularly in relevant

sentence selection. There was no significant difference between two different assessors in terms of selected relevant sentences, but an obvious difference for novelty between two assessors was observed. On average, 42% of relevant sentences were marked novel by the first assessor and 52.6% of relevant sentences were marked novel by the second assessor.

The disagreements on judgments might affect the comparisons of different systems, but Voorhees [59] showed that there was no effect on system comparisons as long as enough topics were used for averaging.

Chapter 4

DEFINITIONS OF NOVELTY AND INFORMATION PATTERNS

The definition of novelty or "new" information is crucial for the performance of a novelty detection system. Unfortunately, as we have pointed out, novelty is usually not clearly defined in the literature. Generally, new words in the text of a sentence, story or document are used to calculate novelty scores by various "novelty" measures. However, new words are not equivalent to novelty (new information). Rephrasing a sentence with a different vocabulary does not mean that this revised sentence contains new information. Furthermore, new words appearing in text that is not relevant to a user's query do not provide any new information with respect to the user's interest, even though the text may contain some new information useful to other users.

4.1. Our Novelty Definition

We give our definition of novelty as follows:

> *Novelty or new information means new answers to the potential questions representing a user's request or information need.*

There are two important aspects in this definition. First, a user's query will be transformed into one or more potential questions for *identifying* corresponding query-related *information patterns,* which include both query words and required answer types. Second, new information is obtained by *detecting* those sentences that include previously unseen "answers" corresponding to the query-related patterns. We emphasize that although a user's information need is typically represented as a query consisting of a few key words, our observation is that a user's information

need may be better captured by one or more questions that lead to corresponding information patterns.

Obviously, the new "novelty" definition can be applied to novelty detection at different levels – event level and sentence level. In this work, we will study novelty detection via information pattern identification at the *sentence level*. This novelty definition is also a general one that works for novelty detection with any query that can be turned into questions.

In this thesis, we show that any query (topic) in the TREC novelty tracks can be turned into either one or more specific NE-questions, or a general question. The *NE-questions* (corresponding to specific topics) are those topics whose answers are specific *named entities* (NEs), including persons, locations, dates, time, numbers, and etc. [21]; a detailed description of the definition and extraction of named entities will be given in Section 4.3. The *general questions (*corresponding to general topics), on the other hand, require obtaining additional information patterns for effective novelty detection, which will be detailed in the next section. Other types of questions can be further explored within the framework of this novelty detection.

Before we discuss information patterns, let us first see some real examples on how to turn queries into questions to facilitate relevance and novelty detection. As shown in the first example (Example 1.1) given in Chapter 1, the answer type in that query-related information pattern is a named entity, i.e., a *number*, and the potential answer lies in those numbers in sentences, i.e., 100 and 11. Therefore, the query-related pattern is a combination of query words and a number for the example query. Here are some more examples about queries that require other types of named entities as their answers.

Example 4.1. "Who/Where/When" questions

<Number>: N59 (from TREC 2004 novelty track)

<Title>: Payne Steward Plane Crash

<Topic-type>: Event

<Description>: Identify a document that describes the plane crash that killed the golfer Payne Stewart on Oct. 25, 1999.

<Narrative>: Details about the crash, **who** else was aboard, and information about the **destination** and departure are relevant. The reason for the flight would not be relevant. **Time** and weather conditions are relevant.

This query about Payne Steward plane crash can be transformed into the following three NE questions by identifying the word patterns (marked in bold) in the narrative field:

1. *Who* was on board?
2. *Where* did the crash happen?
3. *When* did the crash happen?

The first question is a "who-question", which requires that at least one *person* name should appear in the answers related to this question. The second is a "where-question", which requires that answer sentences should contain at least one *location* name. The third question is a "when-question", indicating that a *date* should appear in answers to this question.

Example 4.2. "What" questions

<Number>: N67

<Title>: Military Action Kosovo

<Topic-type>: Opinion

<Description>: The necessary or not necessary need for military action in Kosovo.

<Narrative>: The opinion of various **countries**, **leaders**, and diplomats on whether or not military action is required in Kosovo is relevant. Opinions on the news items or short statements on the Kosovo situation without any apparent opinion are not relevant. Actions taken by the U.N. and/or NATO are relevant.

The query about military action Kosovo can be transformed into the following two NE questions:

1. *What* are the opinions of various countries?

2. *What* are the opinions of the leaders and diplomats of various countries?

The first question indicates that a country name should appear in relevant sentences that may answer this question. The second question indicates that a person name should appear in relevant sentences that may answer this question. Even if they are not explicit NE questions, but NEs of country (*location*) names and *persons* are parts of the answers, respectively. Therefore we also treat this topic as a specific NE topic that can turn into two NE questions.

Some queries (topics) cannot be turned into specific NE questions. Here is an example from the TREC 2002 novelty track.

Example 4.3. A general question

<Number>: 420

<Title> Carbon Monoxide Poisoning

<Description>: What are the symptoms, causes, and methods of preventing carbon monoxide poisoning?

<Narrative>: Relevant documents will contain data on what carbon monoxide poisoning is, symptoms, causes, and/or prevention. Advertisements for carbon monoxide protection products or services are not relevant. Discussions of auto emissions and air pollution are not relevant even though they can contain carbon monoxide.

This topic cannot be formulated into NE questions. Instead, a general question "What are the symptoms, causes, and methods of preventing carbon monoxide poisoning?" will be formulated.

4.2. Information Patterns

4.2.1. Information Patterns for Specific Topics

The identification and extraction of *information patterns* is crucial in our approach. The information pattern corresponding to a specific NE-question (generated from a specific topic/query) is represented by both the query words (of the potential question) and an NE answer type (which requires named entities as its potential answer). They are called *NE words patterns*, related to the questions about DATE ("when"), LOCATION ("where"), PERSON ("who"), ORGANIZATION ("what/who") and NUMBER ("how many"). The details on the definitions and extraction of named entities will be discussed in Section 4.3.

For each type of the five NE-questions, a number of word patterns were constructed for question type identification. Typical NE patterns are listed in Table 4.1; some were extracted from the TREC 2002 novelty track queries manually and some were selected from Li & Croft's question answering system [20]. Each NE word pattern is a combination of both query words (of potential questions) and answer types (which requires named entities as potential answers). We have

shown that our pattern-based approach is very effective in improving the performance of novelty detection for those specific topics (queries). This will be detailed in the following chapters.

Table 4.1. Word patterns for the five types of NE questions

Categories	Answer Types	Word Patterns
Name	Person	who, individual, person, people, participant, candidate, customer, victim, leader, member, player, name
	Organization	who, company, companies, organization, agency, agencies, name, participant
	Location	where, location, nation, country, countries, city, cities, town, area, region
Time	Date	when, date, time, which year, which month, which day
Number	Number	how many, how much, length, number, polls, death tolls, injuries, how long,

4.2.2. Information Patterns for General Topics

For a general topic, it is very difficult (if not impossible) to identify a particular type of named entity as its answer. Any types of named entities or even single words or phrases could be part of an answer as long as the answer context is related to the question. Further, in many relevant and novel sentences, no named entities are included. This observation is supported by the data analysis in Chapter 5. Simply using named entities seems not very helpful for improving the performance of novelty detection for these general topics, as has been shown in [22]. Therefore, the challenging questions are *how* to effectively make use of these named entities, and *what* kinds of additional and critical information patterns will be effective for general topics.

After analyzing the TREC data, we have found that the following three kinds of information patterns are very effective for this purpose: *sentence lengths, named-entity combinations* and *opinion patterns*. We have also found that these patterns are effective for both general and

specific topics. Details will be provided in Chapter 5; in the following, we introduce a particularly effective type of information patterns – opinion patterns.

4.2.3. Opinion Patterns and Opinion Sentences

We note that the topics in TREC 2003 and 2004 novelty tracks are either classified as event topics or opinion topics. According to Ian Soboroff [42], "identifying sentences that contain an opinion remained a hard problem. Event topics proved to be easier than opinion topics." As a specific interesting finding, we have found that a large portion of the general questions are about opinions. Opinions can typically be identified by looking at such sentence patterns as "XXX said", "YYY reported", or as marked by a pair of quotation marks. Currently, we have identified about 20 such opinion-related sentence patterns. The full list of opinion patterns is described in Chapter 5. These patterns are extracted manually from a training set that is composed of about 100 documents from the TREC 2003 novelty track. The opinion patterns currently used in our system are individual words or a sequence of words, such as said, say, claim, agree, found that, state that, etc. Note that the terms remain in their original verbal forms without word stemming, in order to more precisely capture the real opinion sentences. For example, a word "state" does not necessarily indicate an opinion pattern, but the word combination "stated that" most probably does. If a sentence includes one or more opinion patterns, it is said to be an *opinion sentence*.

Here are some examples of "opinion sentences" that are indicated by the occurrences of opinion patterns. The opinion patterns appearing in the sentences ("said" and quotation marks "" in Example 4.4(1), "say" in Example 4.4(2), and "believes" in Example 4.4(3)) are marked in bold. These sentences are from the TREC 2003 novelty track, and they are all related to topic N1. The string before the text of each sentence is its document ID and the sentence number.

Example 4.4. Opinion patterns and sentences

Topic: "Partial Birth Abortion Ban"

<Number>: N1

<Title>: partial birth abortion ban

<Toptype>: opinion

<Description>: Find opinions about the proposed ban on partial birth abortions.

<Narrative>: Relevant information includes opinions on partial birth abortion and whether or not it should be legal. Opinions that also cover abortion in general are relevant. Opinions on the implications of proposed bans on partial birth abortions and the positions of the courts are also relevant.

Example 4.4(1) – Sentence NYT19980629.0465-12: **"**It's really trying to get rid of all abortions,**" said** Patricia Baird-Windle, who owns abortion clinics in both Melbourne and West Palm Beach.

Example 4.4(2) – Sentence NYT19980629.0465-14: Proponents **say** the law specifically targets a method known as dilation and extraction - in which a fetus is partially delivered before it is killed by collapsing its skull - that they call tantamount to infanticide.

Example 4.4(3) - Sentence NYT19980629.0465-34: Eventually, he'd like all abortions to be banned because he **believes** they are murder.

We want to point out that there are some opinion patterns that have not been included in our current system due to the limited number of sentences in the training set. Opinion sentences with word patterns that are not in the list of opinion patterns used in our system will be missed. For instance, Example 4.4(4) and 4.4(5) are opinion sentences, which can be recognized by the occurrence of the word "suggested" and the phrase "appears that", respectively. But they were not identified as opinion sentences simply because "suggested" and "appears that" are not in the list.

The next version of our system will have a list with more opinion patterns based on a large training set.

Example 4.4(4) - Sentence APW20000628.0229-40: O'Connor, who supplied the critical fifth vote, **suggested** in a concurring opinion that states can ban some partial-birth abortions.

Example 4.4(5) - Sentence NYT19980722.0224-31: Now it **appears that** we're about to get stuck with that part of the plan without enjoying many of the aspects of the proposal that would have been beneficial to most Americans.

As the first attempt to identify opinion sentences, our current list of opinion patterns only contains individual words or word patterns. We know that some opinion sentences may contain more complicated opinion patterns, such as phrase, or combinations of words and named entities. To identify those opinion sentences, ISI patterns used for question answering [51] maybe considered. Other techniques in the field of sentiment classification [52, 53] and opinion classification [45] may also be considered in our novelty detection system to improve the accuracy of opinion sentences classification. This will be one of the future directions of the thesis work.

4.3. Named Entities: Definitions, Extraction and Examples

Answers and new answers to specific NE-questions are named entities. For many of the general topics (questions), named entities are also major parts of their answers. Therefore, the extraction of named entities is crucial for the success of novelty detection. We use a named entity extraction program that was developed by BBN, Identifier [21], to mark and to extract named entities. Currently, named entities are classified into four categories: Name, Time, Number and Object. In the following examples, these four categories are represented and marked with XML formats as follows:

(1) Name Category: <ENAMEX TYPE = "XXX">YYY</ENAMEX>

where XXX is the name of the type, which includes three types: PERSON, ORGANIZATION and LOCATION, and YYY is the named entity itself.

(2) Time Category: <TIMEX TYPE = "XXX">YYY</TIMEX>

where XXX is the name of the type, which includes three types: DATE, TIME and PERIOD, and YYY is the named entity itself.

(3) Number Category: <NUMEX TYPE = "XXX">YYY</NUMEX>

where XXX is the name of the type, which includes fourteen types: NUMBER, ORDEREDNUM, ENERGY, MONEY, MASS, POWER, TEMPERATURE, DISTANCE, SPEED, LENGTH, HEIGHT, AREA, SPACE and PERCENT, and YYY is the named entity itself.

(4) Object Category: <OBJECT TYPE = "XXX">YYY</OBJECT>

where XXX is the name of the type, which only includes URL type now, and YYY is the named entity itself, ie., a real URL.

Altogether, there are 21 types of named entities in these four categories. With the above definition, it is easy to read the following examples. These example sentences include some of the named entity types in the four NE categories.

Example 4.5. Named entities (NEs)

Ex 4.5(1) : <s docid="XIE19990105.0233" num="19">

<ENAMEX TYPE = "LOCATION">LONDON</ENAMEX> -- <ENAMEX TYPE = "LOCATION">UK</ENAMEX>'s interest rates will be influenced by the strength or weakness of the euro, said <ENAMEX TYPE = "ORGANIZATION">Bank of England</ENAMEX>

governor <ENAMEX TYPE = "PERSON">Eddie George</ENAMEX> on <TIMEX TYPE = "DATE">Monday</TIMEX>.</s>

Ex 4.5(2) : <s docid = "XIE19970217.0023" num = "12">

According to the latest survey by the German public opinion research institute <ENAMEX TYPE = "PERSON">Allensbach</ENAMEX>, which had questioned <NUMEX TYPE = "NUMBER">600</NUMEX> leaders from the German political and business circles, <NUMEX TYPE = "PERCENT">71 percent</NUMEX> of the respondents believe that the single currency will be launched on time.</s>

Ex 4.5(3) : <s docid = "XIE19980808.0014" num = "14">

An estimated <NUMEX TYPE = "MONEY">12.1 billion</NUMEX> euro coins weighing <NUMEX TYPE = "MASS">50,000 tons</NUMEX> will be needed in <ENAMEX TYPE = "LOCATION">Germany</ENAMEX> alone when the euro completely enters into circulation in <ENAMEX TYPE = "ORGANIZATION">EMU</ENAMEX> countries in <TIMEX TYPE = "DATE">2002</TIMEX>.</s>

Ex 4.5(4) : <s docid = "APW20000130.0118" num = "11">

<TIMEX TYPE = "DATE">24, 1999</TIMEX>: A <ENAMEX TYPE = "ORGANIZATION">China Southwest Airlines</ENAMEX> passenger plane crashes in a field <NUMEX TYPE = "DISTANCE">250 miles</NUMEX> south of <ENAMEX TYPE = "LOCATION">Shanghai</ENAMEX> in <ENAMEX TYPE = "LOCATION">China</ENAMEX>'s coastal <ENAMEX TYPE = "LOCATION">Zhejiang</ENAMEX> province.</s>

Chapter 5

INFORMATION PATTERN ANALYSIS

Novelty detection includes two consecutive steps: first retrieving relevant sentences and then detecting novel sentences. In this chapter we perform a statistical analysis of information patterns (or features) in relevant sentences, novel sentences and non-relevant sentences. The three information patterns studied here are: sentence lengths, named entities, and opinion patterns. The goal is to discover effective ways to use these information patterns in distinguishing relevant sentences from non-relevant ones (step 1), and novel sentences from non-novel ones (step 2).

Currently, there are three sets of data officially available for novelty detection at the sentence level. The TREC 2002 novelty track [6] generated 50 queries (in which 49 test queries are used in our experiments). The TREC 2003 novelty track [23] and 2004 novelty track [42] collected 50 queries each. For each query from the 2002 and 2003 novelty tracks, there are up to 25 relevant documents that were broken into sentences. For each query from the 2004 novelty track, in addition to 25 relevant documents, there could be zero or more non-relevant documents. For all the three datasets, a set of sentences has been pre-marked as relevant/non-relevant, and novel/non-novel for each query (topic). Detailed information has been provided in Chapter 3.

5.1. Statistics on Sentence Lengths

The first type of pattern we have studied is the length of a sentence. This is probably the simplest feature of a sentence. The hypothesis is that it may include some information about the importance of the sentence. The statistics of sentence lengths in TREC 2002 and 2003 datasets

are shown in Table 5.1. The length of a sentence is measured in the number of words after stop words are removed from the sentence. Interestingly, we have found that the average lengths of relevant sentences from the 2002 data and the 2003 data are 15.58 and 13.1, respectively, whereas the average lengths of non-relevant sentences from the 2002 data and the 2003 data are only 9.55 and 8.5, respectively.

Table 5.1. Statistics of sentence lengths

Types of Sentences (S.)	TREC 2002: 49 topics		TREC 2003: 50 topics	
	# of S.	Length	# of S.	Length
Relevant	1365	15.58	15557	13.1
Novel	1241	15.64	10226	13.3
Non-relevant	55862	9.55	24263	8.5

Could this piece of information be useful in relevance and novelty detection? To gain a more concrete insight of the statistics, we list the following two paragraphs excerpted from a relevant document to the topic N1 in TREC 2003. The human judgment of relevance and novelty of each sentence is indicated by R (relevant), NR (non-relevant) or R&N (relevant and novel), given in parentheses in front of the sentence. The length of a sentence is calculated in words after stop words are removed from the sentence; it is shown in parentheses after the sentence. The string before the text of each sentence is its document ID and the sentence number.

Example 5.1. Sentence lengths

Topic: "Partial Birth Abortion Ban"

<Number>: N1

<Title>: partial birth abortion ban

<Toptype>: opinion

<Description>: Find opinions about the proposed ban on partial birth abortions.

<Narrative>: Relevant information includes opinions on partial birth abortion and whether or not it should be legal. Opinions that also cover abortion in general are relevant. Opinions on the implications of proposed bans on partial birth abortions and the positions of the courts are also relevant.

Paragraph 1

(R&N) APW20000114.0177-13: Although the current controversy swirls around a specific procedure, abortion-rights activists contend far more may be at stake. (*11 words*)

(R&N) APW20000114.0177-14: They say the court's eventual decision could broadly safeguard -- or dramatically erode -- abortion rights, depending on what state legislatures are allowed to consider when passing laws to regulate abortions. (*19 words*)

(NR) APW20000114.0177-15: The court will hear arguments in the Nebraska case in April. (*7 words*)

(NR) APW20000114.0177-16: Its decision is expected by July. (*3 words*)

Paragraph 2

(R) APW20000629.0004-24: "A ban on partial-birth abortion that only proscribed the D&X method of abortion and that included an exception to preserve the life and health of the mother would be constitutional, in my view," said Justice Sandra Day O'Connor, who supplied the critical fifth vote for striking down Nebraska's law. (*26 words*)

(NR) APW20000629.0004-25: The Senate passed a partial-birth abortion ban last year and the House passed a similar version two months ago. (*12 words*)

(NR) APW20000629.0004-26: But GOP lawmakers decided to wait on the court before adopting a final bill. (*8 words*)

These two paragraphs clearly show that relevant (and novel) sentences are significantly longer than non-relevant sentences. To conclude, here is our first observation on sentence length (SL) patterns:

SL Observation #1. Relevant sentences on average have more words than non-relevant sentences.

Since we are going to use them in our algorithm designs, the list of the observations (including this one) is listed in Observations and Conclusions at the beginning of the thesis for easy referencing. The sentence length feature is quite simple, but very effective since the length differences between non-relevant and relevant sentences are significant. We want to point out here that this feature is ignored in other approaches, mainly because in the past sentence retrieval was performed with information retrieval techniques developed for document retrieval, where document lengths were usually used as a penalty factor. A long document may discuss multiple topics and a short document may focus on one topic. Therefore, in document retrieval, a short document is usually ranked higher than a long document if the two documents have same occurrences of query words. But at the sentence level, it turns out that relevant sentences have more words than non-relevant sentence on average. Therefore, this observation will be incorporated into the retrieval step to improve the performance of relevant sentence retrieval, which will then boost the performance for identifying novel sentences. Further, we have the second observation on sentence length patterns from the statistics:

SL Observation #2: The difference in sentence lengths between novel and non-relevant sentences is slightly larger than the difference in sentence lengths between relevant sentences and non-relevant sentences.

This indicates that the incorporation of the sentence length information in relevance ranking will put the novel sentences at higher ranks in the relevance retrieval step, which increases the chance to be selected as novel sentences.

5.2. Statistics on Opinion Patterns

Topics in TREC 2003 and 2004 novelty track data collections are classified into event and opinion topics. There are 22 opinion topics out of the 50 topics from the 2003 novelty track. The number is 25 out of 50 for the 2004 novelty track. There are no classifications of opinion and event topics in the 2002 novelty track. We declare a sentence as an *opinion sentence* if it has one or more opinion patterns. Opinion patterns are detected in a sentence if it includes quotation marks or one or more of the expressions indicating it states an opinion.

Table 5.2. Examples of opinion patterns

“ ”,
said, say, according to, add,
addressed, agree, agreed, disagreed,
argue,affirmed, reaffirmed,
believe, believes,
claim, concern, consider, expressed,
finds that, found that, fear that,
idea that, insist,
maintains that,
predicted, reported, report,
state that, stated that, states that,
show that, showed that, shows that,
think, wrote

The full list of opinion patterns is shown here in Table 5.2. The list is by no means complete; however this work wants to show that by making use of this piece of information, the performance of novelty detection can be improved. Intuitively, for an opinion topic, opinion sentences are more likely to be relevant sentences than non-opinion sentences. This hypothesis is tested with a data analysis on the 22 opinion topics from the 2003 novelty track.

We have run our statistical analysis of opinion patterns on the 22 opinion topics from the 2003 novelty track in order to obtain guidelines for using opinion patterns for both 2003 and 2004 data. Statistics show that there are relatively more opinion sentences in relevant (and novel) sentences than in non-relevant sentences. According to the results shown in Table 5.3, 48.1% of relevant sentences and 48.6% of the novel sentences are opinion sentences, but only 28.4% of non-relevant sentences are opinion sentences. This could have a significant impact on separating relevant and novel sentences from non-relevant sentences for opinion topics, therefore we summarize this observation as Opinion Pattern (OP) Observations #1 and #2:

OP Observation #1: There are relatively more opinion sentences in relevant (and novel) sentences than in non-relevant sentences.

OP Observation #2: The difference of numbers of opinion sentences in novel and non-relevant sentences is slightly larger than that in relevant and non-relevant sentences.

Note that the number of opinion sentences in the statistics only counts those sentences that have one or more opinion patterns shown in Table 5.2. We have noticed that some related work [45] has been done very recently in classifying words into opinion-bearing words and non-opinion-bearing words, using information from several major sources such as WordNet, World Street Journal, and General Inquirer Dictionary. Using opinion-bearing words may cover more opinion sentences, but the accuracy of classifying opinion words is still an issue. We believe that a more accurate classification of opinion sentences based on the integration of the results of that work

into our framework could further enlarge the difference in numbers of opinion sentences between relevant sentences and non-relevant sentences.

The second observation indicates that novel sentences may benefit more from the use of opinion patterns.

Table 5.3. Statistics on opinion patterns for 22 opinion topics (2003)

Sentences (S.)	Total # of S.	# of Opinion S. (and %)
Relevant	7755	3733 (48.1%)
Novel	5374	2609 (48.6%)
Non-relevant	13360	3788 (28.4%)

To show the intuitive connections between opinion patterns and relevant (novel) and non-relevant sentences, here we give a real example. Example 5.2 below lists a few paragraphs from a relevant document to the opinion topic N1 in TREC 2003. Each paragraph is broken into several sentences. The string before the text of each sentence represents its document ID and the sentence number. The symbols inside the pair of parentheses before each sentence indicate if the sentence is a relevant, non-relevant or novel sentence (*R for relevant, NR for non-relevant, and R&N for relevant and novel*), and whether it has been identified as an opinion sentence (*O for opinion sentence, X for not*).

Example 5.2. Opinion patterns

Topic: "Partial Birth Abortion Ban"

<Number>: N1

<Title>: partial birth abortion ban

<Toptype>: opinion

<Description>: Find opinions about the proposed ban on partial birth abortions.

<Narrative>: Relevant information includes opinions on partial birth abortion and whether or not it should be legal. Opinions that also cover abortion in general are relevant. Opinions on the implications of proposed bans on partial birth abortions and the positions of the courts are also relevant.

Paragraph 1:

(R - O) NYT20000629.0416-9: Doctors who do abortions as part of their regular medical practice **said** Thursday that they were greatly relieved by the decision of the Supreme Court that the government could not prohibit doctors from performing a procedure that opponents call partial-birth abortion.

(R - O) NYT20000629.0416-10: They **said** that laws trying to ban the procedure were so vague that doctors performing almost any abortion could be prosecuted under them.

(R - O) NYT20000629.0416-11: The doctors **said** that medically, there is no such procedure as a partial-birth abortion, calling the term a political label.

(NR - X) NYT20000629.0416-12 : The Nebraska law struck down by the court described a partial-birth abortion as one in which part of the fetus was outside the woman's body before the fetus was killed.

In this paragraph, the three opinion sentences are relevant sentences, and they are identified by the same word “said”. The fourth sentence is not a relevant sentence, and it is not classified as an opinion sentence by our system.

Paragraph 2:

(R - O) APW20000629.0004-24: **''**A ban on partial-birth abortion that only proscribed the D&X method of abortion and that included an exception to preserve the life and health

of the mother would be constitutional, in my view,**'' said** Justice Sandra Day O'Connor, who supplied the critical fifth vote for striking down Nebraska's law.

(NR - X) APW20000629.0004-25: The Senate passed a partial-birth abortion ban last year and the House passed a similar version two months ago.

(NR - X) APW20000629.0004-26: But GOP lawmakers decided to wait on the court before adopting a final bill.

In this paragraph, the first opinion sentence is a relevant sentence, and it is identified by both a pair of quotation marks and the word “said”. The other two sentences are not relevant sentences, and are not classified as opinion sentences by our system either.

Paragraph 3:

(R&N - X) APW20000114.0177-13: Although the current controversy swirls around a specific procedure, abortion-rights activists ***contend*** far more may be at stake.

(R&N - O) APW20000114.0177-14: They **say** the court's eventual decision could broadly safeguard -- or dramatically erode -- abortion rights, depending on what state legislatures are allowed to consider when passing laws to regulate abortions.

(NR - X) APW20000114.0177-15: The court will hear arguments in the Nebraska case in April.

(NR - X) APW20000114.0177-16: Its decision is expected by July.

In this paragraph, the first two sentences are both relevant and novel, and the last two are non-relevant. Both of the first two sentences are actually opinion sentences, but we only identify the second sentence as an opinion sentence by the word “say”. This first opinion sentence is indicated by the opinion word “contend”, but it is not in our opinion pattern table now so the sentence is not declared as an opinion sentence. Currently the opinion patterns are extracted manually from a

training set of 100 documents. More systematic and probably automatic approaches or a large training set could lead to more accurate and efficient handing of opinion patterns.

5.3. Statistics on Named Entities

As we have pointed out, answers and new answers to specific NE-questions are named entities. And for many of the general topics (questions), named entities are also major parts of their answers. Therefore, understanding the distribution of named entity patterns could be very helpful both in finding relevant sentences and in detecting novel sentences.

The statistics of all the 21 named entities that can be identified by our system are listed in Table 5.4 on TREC 2002 and TREC 2003 novelty track data collections. These named entities are also grouped into four categories: Name, Time, Number and Object. In the table, each item lists two measurements: $N_Y(X)$, and $P_Y(X)$. The former is the number of Y-type (Y is either R for relevant or NR for non-relevant) sentences, each of which has at least one of X-type named entity (X is one of the 21 NE types). The latter is the percentage of those sentences among all the relevant (or non-relevant) sentences in each collection. These two measurements are calculated as

$$N_Y(X) = \#\text{ of Y-type sentences with at least one X-type NE} \quad (5.1)$$

$$P_Y(X) = N_Y(X)/M_Y \quad (5.2)$$

In Eq. 5.2, M_Y denotes the total number of Y-type (either R for relevant or NR for non-relevant) sentences. Those numbers are listed in the head of the table. For example, for TREC 2002, M_R =1365 and M_{NR} =55862. When X = PERSON, we have N_R(PERSON) = 381, P_R(PERSON) = 27.91%, N_{NR}(PERSON) = 1310, and P_{NR}(PERSON) = 23.45%.

From this statistics, we have found that the five most frequent types of NEs are PERSON, ORGANIZATION, LOCATION, DATE and NUMBER. For each of them, there are more than

25% relevant sentences that have at least one named entity of the type in consideration. Note that all the three NE types (PERSON, ORGANIZATION, LOCATION) in the Name category are among the five most frequent types. In the Time category, DATE type is much more significant than TIME and PERIOD types, probably because the significance of a piece of news is characterized by the scale of a day, a month or a year, rather than a specific time of a day, or a period of time. In the Number category, the general Number type (NUMBER) of named entities is overwhelmingly more significant than all the other specific Number types (such as MONEY, POWER, LENGTH, etc.).

Among these five types of NEs, three (PERSON, LOCATION and DATE) of them are more important than the other two (NUMBER and ORGANIZATION) for separating relevant sentences from non-relevant sentences. For example, for TREC 2003, the percentage of relevant sentences that include PERSON type name entities, P_R(PERSON), is about 13% more than that of the non-relevant sentences, P_{NR}(PERSON). The discrimination capability of the ORGANIZATION type in relevance detection is not as significant; P_R(ORGANIZATION) is just marginally higher than P_{NR}(ORGANIZATION). The insignificant role of ORGANIZATION in relevant retrieval has also been validated by our experiments on real data. One of the reasons is that an NE of this type often indicates the name of a news agency; and names of news agencies often occur frequently in news articles. The role of the NUMBER type is not consistent among three TREC datasets (2002, 2003, and 2004). In Table 5.4, P_R(NUMBER) is higher than P_{NR}(NUMBER) for TREC 2002, but is lower for TREC 2003. Therefore only the three effective types will be incorporated into the sentence retrieval step to improve the performance of relevance. This leads to our first important observation on named entities (NEs):

NE Observation #1. Named entities of the PLD types - PERSON, LOCATION and DATE are the more effective in separating relevant sentences from non-relevant sentences.

Table 5.4. The statistics of named entities (2002, 2003)

TREC 2002 Novelty Track Total S# = 57227, M_R =1365, M_{NR} =55862			**TREC 2003** Novelty Track Total S# = 39820, M_R =15557, M_{NR} =24263		
NEs	**Rel#(%)**	**Non-Rel#(%)**	**NEs**	**Rel#(%)**	**Non-Rel#(%)**
Name Category					
PERSON	381(27.91%)	13101(23.45%)	PERSON	6633(42.64%)	7211(29.72%)
ORGANIZATION	532(38.97%)	17196(30.78%)	ORGANIZATION	6572(42.24%)	9211(37.96%)
LOCATION	536(39.27%)	11598(20.76%)	LOCATION	5052(32.47%)	5168(21.30%)
Time Category					
DATE	382(27.99%)	6860(12.28%)	DATE	3926(25.24%)	4236(17.46%)
TIME	9(0.66%)	495(0.89%)	TIME	140(0.90%)	1154(4.76%)
PERIOD	113(8.28%)	2518(4.51%)	PERIOD	1017(6.54%)	705(2.91%)
Number Category					
NUMBER	444(32.53%)	14035(25.12%)	NUMBER	4141(26.62%)	6573(27.09%)
ORDEREDNUM	77(5.64%)	1433(2.57%)	ORDEREDNUM	725(4.66%)	688(2.84%)
ENERGY	0(0.00%)	5(0.01%)	ENERGY	0(0.00%)	0(0.00%)
MONEY	66(4.84%)	1775(3.18%)	MONEY	451(2.90%)	1769(7.29%)
MASS	31(2.27%)	1455(2.60%)	MASS	34(0.22%)	19(0.08%)
POWER	16(1.17%)	105(0.19%)	POWER	0(0.00%)	0(0.00%)
TEMPERATURE	3(0.22%)	75(0.13%)	TEMPERATURE	25(0.16%)	9(0.04%)
DISTANCE	8(0.59%)	252(0.45%)	DISTANCE	212(1.36%)	47(0.19%)
SPEED	0(0.00%)	0(0.00%)	SPEED	32(0.21%)	2(0.01%)
LENGTH	46(3.37%)	682(1.22%)	LENGTH	103(0.66%)	29(0.12%)
HEIGHT	2(0.15%)	25(0.04%)	HEIGHT	1(0.01%)	3(0.01%)
AREA	5(0.37%)	72(0.13%)	AREA	17(0.11%)	11(0.05%)
SPACE	2(0.15%)	54(0.10%)	SPACE	11(0.07%)	10(0.04%)
PERCENT	62(4.54%)	1271(2.28%)	PERCENT	371(2.38%)	1907(7.86%)
Object Category					
URL	0(0.00%)	0(0.00%)	URL	0(0.00%)	62(0.26%)
Others Category					
No NEs	246(18.02%)	15899(28.46%)	No NEs	3272(21.03%)	5533(22.80)
No POLD	359(26.3%)	22689(40.62%)	No POLD	4333(27.85%)	8674(35.75%)
No PLD	499(36.56%)	31308(56.05%)	No PLD	6035(38.79%)	12386(51.05%)

However, the ORGANIZATION type will be also used in the new pattern detection step since a different organization may provide new information. Here is an example:

Example 5.3. Organization in new pattern detection

<docid = "NYT19980629.0465" num = "42">

<ENAMEX TYPE = "ORGANIZATION">**Christian Coalition of Florida**</ENAMEX> director <ENAMEX TYPE = "PERSON">John Dowless</ENAMEX> disagrees with <ENAMEX TYPE = "PERSON">Carres</ENAMEX>' spin but agrees that a judge or judges somewhere, state or federal, is liable to strike down the law.

In this example, the organization name "Christian Coalition of Florida" gives an indication that new information may be provided by this sentence. We summarize this into NE Observation #2:

NE Observation #2: Named entities of the POLD types - PERSON, ORGANIZATION , LOCATION, and DATE will be used in new pattern detection; named entities of the ORGANIZATION type may provide different sources of new information.

Table 5.4 also lists the statistics of those sentences with no NEs at all, with no POLD (Person, Organization, Location or Date) NEs, and with no PLD (Person, Location or Date) NEs. These data show the following facts:

(1) There are obvious larger differences between relevant and non-relevant sentences without PLD NEs, than the differences without POLD NEs, or without any NEs. This confirms that PLD NEs are more effective in re-ranking the relevance score (NE Observation #1).

(2) There are quite large percentages of relevant sentences without NEs or without POLD NEs. Therefore, first, the absence of NEs is not going to be used exclusively to remove sentences from the relevant sentence list. Second, the number of the previously unseen POLD NEs only contributes partly to novelty ranking. This point can be summarized into the following two NE Observations:

NE Observation #3: The absence of NEs cannot be used exclusively to remove sentences from the relevant sentence list.

NE Observation #4: The number of the previously unseen POLD NEs only contributes part of the novelty ranking.

We have also done three particular investigations. First, we examine how to count NEs by analyzing two kinds of NE pattern distributions on the four classes of sentences: relevant, non-relevant, novel and non-novel. Second, we want to understand the role of certain named entities in separating relevant sentences from non-relevant sentences, for event topics and opinion topics, respectively. Finally, we want to further analyze the role of named entities in novel sentence extraction. These investigations will be discussed in the following three subsections.

5.3.1. Named Entities: How to Count Them

To start with, we define two kinds of distributions on relevant and non-relevant sentences respectively. Assume that the total number of relevant sentences in a dataset is M_r, and the total number of non-relevant sentences is M_{nr}. Let us denote the number of named entities in a sentence as N, and the number of different types of named entities in a sentence as ND. In the latter case, if more than one named entity of the same type occurs, only one count is added. We would like to see which measurements are more effective in separating relevant and non-relevant sentences, and novel and redundant sentences. If the occurrence of relevant sentences with N named entities is represented as $O_r(N)$, then the "probability" of the relevant sentences with N named entities can be represented as

$$P_r(N) = O_r(N)/M_r \quad (5.3)$$

Similarly the occurrence and probability of the non-relevant sentences with N named entities can be represented as $O_{nr}(N)$ and $P_{nr}(N)$, where

$$P_{nr}(N) = O_{nr}(N)/M_{nr} \quad (5.4)$$

We can also define the occurrence and probability of the relevant sentences with ND types of named entities as $O_r(ND)$ and $P_r(ND)$, where

$$P_r(ND) = O_r(ND)/M_r \quad (5.5)$$

The occurrences and probability of the non-relevant sentences with ND types of named entities are $On_r(ND)$ and $P_{nr}(ND)$, where

$$P_{nr}(ND) = O_{nr}(ND)/M_{nr} \quad (5.6)$$

The occurrences and probabilities of the novel and non-novel sentences with N named entities or ND types of named entities can be defined in the same way. Note that here "novel" means "relevant and containing new information", while "non-novel" means "non-relevant" or "relevant but containing no new information". Let us assume that the total number of novel sentences in the dataset is M_n, and the total number of non-novel sentences is M_{nn}. Then the occurrence and probability of the novel sentences with N named entities can be represented as $O_n(N)$ and $P_n(N)$, and of the non-novel sentences as $O_{nn}(N)$ and $P_{nn}(N)$, respectively , where

$$P_n(N) = O_n(N)/M_n \quad (5.7)$$

$$P_{nn}(N) = O_{nn}(N)/M_{nn} \quad (5.8)$$

The occurrence and probability of the novel sentences with ND different types of named entities can be represented as $O_n(ND)$ and $P_n(ND)$, and of the non-novel sentences as $O_{nn}(ND)$ and $P_{nn}(ND)$, respectively , where

$$P_n(ND) = O_n(ND)/M_n \quad (5.9)$$

$$P_{nn}(ND) = O_{nn}(ND)/M_{nn} \quad (5.10)$$

In the following, we show and explain the results from our novelty data investigation. We use 101 queries where 53 queries are from the TREC 2002 novelty track (49 testing queries and 4 training queries from TREC 2002) and 48 queries are from the dataset collected by a research group in CIIR at UMass-Amherst [7]. This group developed 48 topics in order to better train their novelty detection system when they participated in the TREC 2002 novelty track. They used a method almost identical to that used by NIST to create the dataset. Therefore, the UMass data is very similar to the TREC 2002 novelty track data that has been described in detail in Chapter 3. For each query there is a set of sentences that have been pre-marked as relevant/non-relevant, and novel/non-novel. The total number of sentences for all 101 queries is 146,319, in which the total number of relevant sentences M_r is 4,947, and the total number of non-relevant sentences M_{nr} is 141,372. The total number of novel sentences M_n is 4,170, and the number of non-novel sentences M_{nn} is 142,149. In our experiments, named entities include all the 21 types listed in Table 5.4.

We perform two sets of data analyses. In the first set, we compare the distributions of named entities in relevant and non-relevant sentences to the given queries. In the second set, we further compare the distributions of named entities in *novel* and *non-novel* sentences. We have performed the t-test for significance on the data analysis, and the distributions of named entities in relevant/novel and non-relevant/non-novel sentences are significantly different from each other at the 95% confidence level except those two that are marked with an asterisk.

Table 5.5 and Table 5.6 show the results of the first set of statistical analyses. In Table 5.5, the second and third columns show the distributions of relevant sentences and non-relevant sentences with different types of named entities, indicated by the numbers in the first row (ND), whereas the fourth and fifth columns show the distributions of relevant/non-relevant sentences with certain numbers of named entities, also indicated by the numbers in the first row (N). Table 5.6 gives statistical results on the number of relevant/non-relevant sentences that have some combinations

of named entity types (patterns) that might be more important in novelty detection: person and location, person and date, location and date, and person, location and date.

The particular NE combinations we select (in Table 5.6) have more impact on relevant sentence retrieval. For general combinations of two types of named entities (ND = 2 in Table 5.5), the ratios of named entity occurrence percentiles $P_r(ND)/P_{nr}(ND)$ between relevant and non-relevant sentences is only 22.4%/19.4% =1.16. But the average ratio for three types of combinations of two different named entities (in Table 5.6) is 2.41. The ratios for the combinations of three types of named entities (ND=3) are 1.85 in the general case (Table 5.5) and 3.21 for a particular person-location-date combination (in Table 5.6). Note that the combination of PERSON, LOCATION and DATE has the highest ratio.

Table 5.5. Named Entities (NE) distributions in relevant/non-relevant sentences

	NE Type Distributions		**NE # Distributions**	
ND or N	$O_r(ND)$ ($P_r(ND)$)	$O_{nr}(ND)$ ($P_{nr}((ND))$	$O_r(D)$ ($P_r(D)$)	$O_{nr}(D)$ ($P_{nr}(D)$)
0	1141 (23.1%)	45508 (32.2%)	1141 (23.1%)	45508 (32.2%)
1	1301 (26.3%)	49514 (35.0%)	987 (20.0%)	40294 (28.5%)
2	1110 (22.4%)	27465 (19.4%)	807 (16.3%)*	22877 (16.2%)*
3	816 (16.5%)	12548 (8.9%)	635 (12.8%)	13323 (9.4%)
4	425 (8.6%)	4616 (3.3%)	482 (9.7%)	7832 (5.5%)
5	124 (2.5%)	1351 (1.0%)	351 (7.1%)	4627 (3.3%)
>5	30 (0.6%)	370 (0.3%)	544 (11.0%)	6911 (4.9%)

Table 5.6. NE combinations in relevant / non-relevant sentences

NE Combination	**# of Relevant Sentences (%)**	**# of Non-Relevant Sentences (%)**
PersonLocation	582 (11.8%)	8543 (6.0%)
PersonDate	427 (8.6%)	4705 (3.3%)
LocationDate	604 (12.2%)	5913 (4.2%)
PersonLocationDate	225 (4.5%)	2028 (1.4%)

The results in Table 5.5 and Table 5.6 indicate that overall, relevant sentences contain more named entities than the non-relevant sentences (as a percentage). More importantly, we have the following two observations:

NE Observation #5. The number of different types of named entities is more significant than the number of entities in discriminating relevant form non-relevant sentences.

NE Observations #6. Some particular NE combinations have more impact on relevant sentence retrieval.

This is particularly true when ND and N ≥ 2. These two observations are incorporated into relevant sentence re-ranking where the number of different types of named entities (in particular PERSON, LOCATION and DATE) are used instead of merely the number of named entities.

In the second analysis, we further study the distributions of named entities in *novel* and *non-novel* sentences. Table 5.7 and Table 5.8 show the results. The design of the "novelty distribution" experimental analysis in Tables 3 and 4 is the same as the design of relevance distribution, except that in novelty distribution analysis, we measure the distributions of named entities with respect to novel and non-novel sentences respectively. We found similar results to those in relevant and non-relevant sentences. The most important findings are summarized as the following two NE Observations:

NE Observation #7: There are relatively more novel sentences (as a percentage) than non-novel sentences that contain at least 2 different types of named entities (Table 5.7)

NE Observation #8: There are relatively more novel sentences (in percentiles) than non-novel sentences that contain the four particular NE combinations of interest (Table 5.8)**.**

Table 5.7. Named Entities in novel/ non-novel sentences

	NE Type Distributions		NE # Distributions	
ND or N	O_n(ND) (P_n(ND))	O_{nn}(ND) (P_{nn}(ND))	O_n(D) (P_n(D))	O_{nn}(D) (P_{nn}(D))
0	947 (22.7%)	45702 (32.2%)	947 (22.7%)	45702 (32.2%)
1	1058 (25.4%)	49757 (35.0%)	814 (19.5%)	40467 (28.5%)
2	937 (22.5%)	27638 (19.4%)	660 (15.8%)*	23024 (16.2%)*
3	714 (17.1%)	12650 (8.9%)	541 (13.0%)	13417 (9.4%)
4	375 (9.0%)	4666 (3.3%)	417 (10.0%)	7897 (5.6%)
5	111 (2.7%)	1364 (1.0%)	313 (7.5%)	4665 (3.3%)
>5	28 (0.7%)	372 (0.3%)	478 (11.5%)	6977 (4.9%)

Table 5.8. NE combinations in novel and non-novel sentences

NE Combination	# of Novel Sentences (%)	# of Non-Novel Sentences (%)
PersonLocation	498 (11.9%)	8627 (6.1%)
PersonDate	373 (8.9%)	4759 (3.3%)
LocationDate	519 (12.4%)	5998 (4.2%)
PersonLocationDate	200 (4.8%)	2053 (1.4%)

5.3.2. Named Entities in Event and Opinion Topics

As we have discussed, topics in TREC 2003 and TREC 2004 novelty track data collections are classified into two types: opinion topics and event topics. If a topic can be transformed into multiple NE-questions, no matter it is an opinion or event topic, the relevant and novel sentences for this "specific" topic can be extracted by mostly examining required named entities (NEs) as answers to these questions generated from the topic. Otherwise we can only treat it as a general topic for which no specific NEs can be used to identify sentences as answers. The analysis in Section 5.2 shows that we can use opinion patterns to identify opinion sentences that are more probably relevant to opinion topics (queries) than non-opinion sentences. However, opinion topics only consist of part of the queries (Table 5.9). There are only 22 opinion topics out of the

50 topics from the 2003 novelty track. The number is 25 out of 50 for the 2004 novelty track. Now the question is: how to deal with those event topics?

Table 5.9. Key data about opinion and event topics (TREC 2003)

R: Relevant; NR: Non-Relevant	# Topics	# Sentences	$M_{R\text{-}Z}$	$M_{NR\text{-}Z}$
Opinion Topics (Z = O)	22	21115	7755	13360
Event Topics (Z=E)	28	18705	7802	10903

Since an event topic (query) is usually about an event with persons, locations and dates involved, naturally we turn to named entities about the types of PERSON, LOCATION and DATES for help. Statistics also verifies this hypothesis. Table 5.10 compares the difference in the statistics of NEs between event topics and opinion topics for the TREC 2003 novelty track. In Table 5.10, each item lists two measurements: $N_{Y\text{-}Z}(X)$, and $P_{Y\text{-}Z}(X)$. The former is the number of Y-type (Y is either R for relevant or NR for non-relevant) sentences for Z-category topics (Z is either O for opinion, or E for event), each of which has at least one of X-type named entity (X is PERSON, LOCATION or DATE). The latter is the percentage of those sentences among all the relevant (or non-relevant) sentences in each category (event or opinion). These two measurements are calculated as

$$N_{Y\text{-}Z}(X) = \#\text{ of Y-type sentences for Z-category topics with at least one X-type NE} \quad (5.11)$$

$$P_{Y\text{-}Z}(X) = N_{Y\text{-}Z}(X)/M_{Y\text{-}Z} \quad (5.12)$$

In Eq. 5.12, $M_{Y\text{-}Z}$ denotes the total number of Y-type (Y is either R for relevant or NR for non-relevant) sentences in the Z category (Z is either E for event or O for opinion). Those numbers are listed in Table 5.9. When X = PERSON, we have $P_{R\text{-}E}(PERSON) = 49.13\%$, $P_{NR\text{-}E}(PERSON) = 29.61\%$, $P_{R\text{-}O}(PERSON) = 36.11\%$, and $P_{NR\text{-}O}(PERSON) = 29.81\%$. For PERSON, LOCATION and DATE types, the relevant to non-relevant percentage ratio for event topics, i.e.,

$P_{R\_E}(X)/P_{NR\_E}(X)$, is close to 2:1 on average, whereas the ratio for opinion topics, i.e., $P_{R\_O}(X)/P_{NR\_O}(X)$, is only about 5:4 on average.

Therefore, we obtain the following observation about named entities and opinion/event topics:

NE Observation #9 (OP Observation #3): Named Entities of the PERSON, LOCATION and DATE types play a more important role in event topics than in opinion topics.

Table 5.10. Statistics of named entities in opinion and event topics (2003)

TREC 2003 Novelty Track Event Topics Total = 18705, Total Rel#= 7802, Total Non-Rel#= 10903			**TREC 2003** Novelty Track Opinion Topics Total S# = 21115, Total Rel#= 7755, Total Non-Rel#= 13360		
NEs	**Rel#(%)**	**Non-Rel#(%)**	**NEs**	**Rel#(%)**	**Non-Rel#(%)**
PERSON	3833(49.13%)	3228(29.61%)	PERSON	2800(36.11%)	3983(29.81%)
LOCATION	3100(39.73%)	2567(23.54%)	LOCATION	1952(25.17%)	2601(19.47%))
DATE	2342(30.02%)	1980(18.16%)	DATE	1584(20.43%)	2256(16.89%))

This is further verified in our experiments of relevance retrieval. In the equation for NE-adjustment (Eq. 6.9), the best results are achieved when α takes the value of 0.5 for event topics and 0.4 for opinion topics. Note that while opinion patterns in sentences are used for improving the performance of novelty detection for opinion topics, named entities play a more important role for event topics.

In the following, two examples are provided to somewhere show the point we have made. In Example 5.4, an event topic and a paragraph from its relevant document are shown, whereas in Example 5.5 an opinion topic and a paragraph from its relevant document are shown. Each paragraph is divided into sentences, each marked by a pair (X-Y), where X can be either R (relevant) or NR (non-relevant), and Y can be either NE (having named entities) or X (no named entities) . Named entities are also labelled by XML formats that we have discussed in Section 4.3.

In the “event” paragraph, there are 4 relevant sentences and 4 non-relevant sentences. All the 4 relevant sentences have NEs, whereas only one of the 4 non-relevant sentences has NEs. In comparison, in the “opinion” paragraph, while all the 3 relevant sentences have NEs, 3 of the 4 non-relevant sentences also have NEs. Fortunately, we can use opinion patterns for this opinion topic. All the three relevant sentences to this opinion topic are opinion sentences indicated by words “said”, “fear”, “disagrees” and “agrees”. Only one of the three non-relevant sentences that have NEs is an opinion sentence indicated by the word “said”. Another non-relevant sentence also includes a word “said” (hence will be judged as an opinion sentence) but it does not include NEs.

Example 5.4. An event topic and its related sentences

<num>Number: N2

<title>clone Dolly sheep

<toptype>event

<desc>Description: Cloning of the sheep Dolly

<narr>Narrative: To be relevant information there must be specific reference to 'Dolly' or 'the first cloned sheep' or 'large animal.' References to Dolly's children are relevant if Dolly's name is included. Mention of the company that cloned Dolly is not relevant if nothing more is said about Dolly. References to the consequences of her being a clone are relevant. Mention of Polly and Molly are not relevant.

(R-NE)<s_ne docid="NYT19981216.0443" num="24">Until <ENAMEX TYPE="PERSON">Dolly</ENAMEX> the lamb, the <NUMEX TYPE="ORDEREDNUMBER">first</NUMEX> clone, was born, it was a scientific truism that the cloning of adults was biologically impossible.</s_ne>

(NR-X)<s_ne docid="NYT19981216.0443" num="25">Scientists reasoned that even though every cell in an animal's body has the same genetic material, adult cells are the result of a process of differentiation that begins in the womb.</s_ne>

(NR-X)<s_ne docid="NYT19981216.0443" num="26">After that, the theory went, a brain cell stays a brain cell, a heart cell remains a heart cell.</s_ne>

(NR-X)<s_ne docid="NYT19981216.0443" num="27">To clone, scientists would have to take such a developed cell and reverse its genetic program so that it could direct the development of a new animal, an identical twin of the adult.</s_ne>

(NR-NE)<s_ne docid="NYT19981216.0443" num="28">That is what scientists achieved, <NUMEX TYPE="ORDEREDNUMBER">first</NUMEX> with <ENAMEX TYPE="

PERSON">Dolly</ENAMEX> and now with other animals.</s_ne>

(R-NE)<s_ne docid="NYT19981216.0443" num="29">But with <ENAMEX TYPE="PERSON">Dolly</ENAMEX>, cloning seemed almost impossibly arduous.</s_ne>

(R-NE)<s_ne docid="NYT19981216.0443" num="30">Scientists at the <ENAMEX TYPE="ORGANIZATION">Roslin Institute</ENAMEX> in <ENAMEX TYPE="LOCATI

ON">Scotland</ENAMEX> started with about <NUMEX TYPE="NUMBER">400</NUMEX> unfertilized sheep eggs.</s_ne>

(R-NE)<s_ne docid="NYT19981216.0443" num="31">Out of those eggs, they got <NUMEX TYPE="NUMBER">one</NUMEX> lamb.</s_ne>

Example 5.5. An opinion topic and its related sentences

<num>Number: N1

<title>partial birth abortion ban

<toptype>opinion

<desc>Description: Find opinions about the proposed ban on partial birth abortions.

<narr>Narrative: Relevant information includes opinions on partial birth abortion and whether or not it should be legal. Opinions that also cover abortion in general are relevant.

Opinions on the implications of proposed bans on partial birth abortions and the positions of the courts are also relevant.

(R-NE) <s_ne docid="NYT19980629.0465" num="36">As to whether the partial-birth is the proverbial foot-in-the door: **"**I hope so,**"** <ENAMEX TYPE="PERSON">Ball</ENAMEX> **said**.</s_ne>

(R-NE)<s_ne docid="NYT19980629.0465" num="37">And that is precisely what abortion-rights activists **fear**, and why they plan to argue in court <TIMEX TYPE="DATE">today</TIMEX> that the partial-birth ban, by its broadness, takes away some of the abortion rights guaranteed by the <ENAMEX TYPE="ORGANIZATION">U.S. Supreme Court</ENAMEX> Roe vs.</s_ne>

(NR-NE)<s_ne docid="NYT19980629.0465" num="38">Wade decision <NUMEX TYPE="PERIOD">25 years</NUMEX> ago.</s_ne>

(NR-NE)<s_ne docid="NYT19980629.0465" num="39">The law is the <NUMEX TYPE="ORDEREDNUMBER">first</NUMEX> abortion restriction in <ENAMEX TYPE="

LOCATION">Florida</ENAMEX> since the parental consent law that lasted <NUMEX TYPE="PERIOD">two months</NUMEX> before it was overturned in <TIMEX TYPE="DATE">1989</TIMEX>.</s_ne>

(NR-NE)<s_ne docid="NYT19980629.0465" num="40"><ENAMEX TYPE="PERSON">Charlene Carres</ENAMEX>, a <ENAMEX TYPE="LOCATION">Tallahassee</ENAMEX> abortion-rights lawyer who will ask the federal judge to stop the new law from taking effect, **said** most of the bans in the other <NUMEX

TYPE="NUMBER">17</NUMEX> states that have passed them have also been struck down by judges.</s_ne>

(NR-X)<s_ne docid="NYT19980629.0465" num="41">``Those judges can't all be reading it wrong," she **said**.</s_ne>

(R-NE)<s_ne docid="NYT19980629.0465" num="42"><ENAMEX TYPE="ORGANIZATION">Christian Coalition of Florida</ENAMEX> director <ENAMEX TYPE="PERSON">John Dowless</ENAMEX> **disagrees** with <ENAMEX TYPE="PERSON">Carres</ENAMEX>' spin but **agrees** that a judge or judges somewhere, state or federal, is liable to strike down the law.</s_ne>

5.3.2. New Named Entity Pattern Analysis

The third investigation is to study the relationship of *new* named entities and novelty/redundancy, which is probably more important in novelty detection. For NE questions, relevant sentences should contain named entities as potential answers to given questions, and novel sentences should contain new answers or previously unseen named entities. Thus a relevant sentence with no new answers/named entities is said to be redundant.

Table 5.11. Previously unseen NEs and Novelty/Redundancy

Types of Sentences	Total # of Sentences	# of Sentences /w New NEs (%)	# of Queries
Novel S.	4170	2801 (67.2%)	101
Redundant S.	777	355 (45.7%)	75*

* Among 101 topics, only 75 of them has redundant sentences

We use 101 queries where 53 queries are from the TREC 2002 novelty track and 48 queries are from the dataset collected by the CIIR research group at UMass [7]. Table 5.11 shows that 67.2% of novel sentences do have new named entities while only 45.7% of relevant but redundant sentences have new named entities. This leads two the 10th Named Entity (NE) Observation:

NE Observation #10. There are more new named entities in novel sentences than in relevant but redundant sentences.

There are two further interesting questions based on these statistics. First, there are 32.8% novel sentences that don't have any new named entities. Why are these sentences marked novel if they do not contain previously unseen named entities? Second, there are 45.7% redundant sentences that do contain new named entities. Why are these sentences redundant if they have previously unseen named entities?

To answer these two questions, we did a further investigation on the novel/redundant sentences and its corresponding queries. We have found that most of the novel sentences *without* new named entities are related to particular queries. These queries can only be transformed into general questions, but not NE questions that ask for certain types of named entities/patterns as answers. For example, query 420 from TREC novelty track data (Example 4.3) is concerned about the symptoms, causes and prevention of carbon monoxide poisoning. A relevant sentence to this query doesn't have to have any named entities to be relevant, let alone new named entities. In fact, most of the relevant sentences for this query don't contain any named entities at all. There are about 18 such queries out of the 101 queries investigated. This investigation is related to NE Observation #4, which is about the use of NE for general topics.

For the second question, all types of new named entities that could be identified by our algorithms and appear in a sentence are considered in the statistics. However, for each NE question, only a particular type of named entity appeared in a relevant sentence is of interest. For example, query 306 (Example 1.1) is about "*How many civilian non-combatants have been killed in the various civil wars in Africa*". For this query, a number appearing in a relevant sentence could be an answer, whereas a person name or other named entities may not be of interest. Therefore, a relevant sentence with a previously unseen person name could be redundant. This leads to the last observation about named entities:

NE Observation #11: Only certain types of named entities may contain important information for a specific topic.

This observation is used in extracting both relevant and novel sentences for specific topics.

Chapter 6

PATTERN-BASED APPROACH TO NOVELTY DETECTION

In our definition, novelty means *new answers to the potential questions* representing a user's information need, and answers are characterized by query-related *information patterns*. Given this definition of novelty, it is possible to detect new information patterns for monitoring how the potential answers to a question change. Consequently, we propose a new novelty detection approach based on the identification of query-related information patterns at the sentence level. In the following, we will first give an overview of our unified pattern-based approach for both specific and general topics (queries). Then, with a focus on using information patterns in improving the novelty detection performance, we will detail the three steps:

(1) information pattern identification via query analysis,

(2) information pattern utilization in relevant sentence retrieval, and

(3) information pattern utilization in novel sentence extraction.

Different treatments for the two types of topics, specific topics and general topics, will be highlighted.

6.1. ip-BAND: A Unified Pattern-Based Approach

The unified information-pattern-BAsed Novelty Detection (ip-BAND) approach for novelty detection is illustrated in Figure 2. There are three important steps in the proposed approach: *query analysis, relevant sentence retrieval* and *novel sentence extraction.* In the first step, an

information request from a user will be implicitly transformed into one or more potential questions in order to determine corresponding query-related information patterns. Information patterns are represented by combinations of query words and required answer types to the query. In the second step, sentences with the query-related patterns are retrieved as answer sentences. Then in the third step, sentences that indicate potential new answers to the questions are identified as novel sentences. This section summarizes the unified approach for both specific and general questions. Details will be provided in Sections 6.2 to Section 6.4.

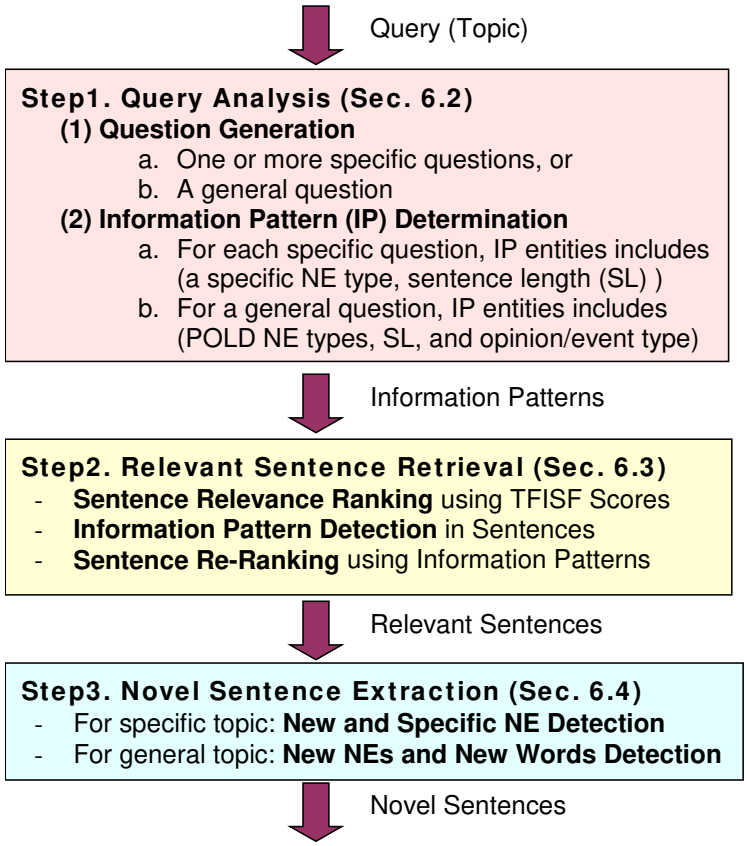

Figure. 2. ip-BAND: a unified information-pattern-based novelty detection approach

6.1.1. Query Analysis

In the first step, a question formulation algorithm first tries to automatically formulate multiple specific NE questions for a query, if possible. Each potential question is represented by a query-related pattern, which is a combination of a few query words and the expected answer type. A specific question would require a *particular* type of named entities (NEs) for answers. Five types of specific questions are considered in the current system: *PERSON, ORGANIZATION, LOCATION, NUMBER* and *DATE*.

If this is not successful, a general question will be generated. A g*eneral question* does not require a particular type of named entity for an answer. Any type of named entities (NEs) as listed in Table 5.4 as well as other single words could be a potential answer or part of an answer, as long as the answer context is related to the question. This means answers for general questions could also be in sentences without any named entities (NEs). However, from our data analysis, the NEs of POLD types (*PERSON, ORGANIZATION, LOCATION, DATE*) are the most effective in detecting novel sentences, and three of them (*PERSON, LOCATION, DATE*) are the most significant in separating relevant sentences from non-relevant sentences. In addition, as we have observed in the statistics in Section 4, sentence lengths and opinion patterns are also important in relevant sentence retrieval and novel sentence extraction. In particular, we can use opinion patterns to identify opinion sentences that are more probably relevant to opinion topics (queries) than non-opinion sentences. PERSON, LOCATION and DATE play a more important role in event topics than in opinion topics. Therefore, for a general question, its information pattern includes topic type (event or opinion), sentence length, POLD NE types, as well as query words.

There are 49 queries in the TREC 2002 novelty track, 50 queries in the TREC 2003 novelty track and 50 queries in the TREC 2004 novelty track. Our question formulation algorithm (Section 6.2) formulates multiple specific questions for 8 queries from the TREC 2002 novelty track, for 15 queries from the TREC 2003 novelty track, and for 11 queries from the TREC 2004 novelty

track, respectively. The remaining queries were transformed into general questions. Details of query analysis will be described in Section 6.2.

6.1.2. Relevant Sentence Retrieval

At this step, the relevant sentence retrieval module first takes key words of the query and searches in its data collection to retrieve sentences that are *topically* relevant to the query. It then re-ranks the sentences retrieved using corresponding information patterns. Sentences that do not satisfy the query-related patterns for specific questions are filtered out because they are unlikely to have potential answers without the required answer patterns.

For a specific topic, multiple specific questions may be generated; only certain types of named entities that the questions expect would be considered for potential answers. Thus a sentence without the expected types of named entities will be removed from the list. For a general topic, all types of named entities, words, and phrases could be potential answers therefore filtering is not conducted.

For both specific and general topics, corresponding information patterns are used in re-ranking the relevance list in order to improve the performance of relevance and therefore novelty. This means that at the retrieval step, the system will revise the relevance sentence retrieval results by adjusting their ranking scores from first round retrieval using sentence lengths, NEs and opinion patterns. Only three types of NEs (Person, Location and Date) are considered in re-ranking. Details will be provided in Section 6.3 and different treatments for specific and general topics will be discussed.

6.1.3. Novel Sentence Extraction

At this step, the new sentence extraction module extracts all query-related named entities and single words for general topics (as possible answers) from each answer sentence and detects previously unseen “answers”. For a specific topic with multiple specific NE questions that are

formulated from the topic (query), an answer sentence may have answers to one or more of these specific NE questions. So named entities related to any one of the specific questions in the answer sentences should be extracted, and are treated as answers. Details will be discussed in Section 6.4.

6.2. Query Analysis

6.2.1. Query Analysis and Question Formulation

The first step in the proposed pattern-based approach is query analysis. It examines a user's query and determines the possible query-related patterns that correspond to one or more potential specific questions, or a general question, transformed from the query. A question formulation algorithm first tries to automatically formulate multiple specific questions for a query if possible. If this is not successful, a general question will be generated. Each potential question is represented by a query-related pattern, which is a combination of a few query words and the expected answer type. In this work, we deal with both specific NE-questions that expect some type of named entities for answers, and general questions that expect more general answer types. Therefore, a specific question would require a *particular* type of named entities for answers. Five types of specific questions are considered in the current system: *PERSON, ORGANIZATION, LOCATION, NUMBER* and *DATE*.

General questions, corresponding to general topics, do not require particular types of named entities for answers. Any type of named entity as well as single words and phrases could be an answer or part of an answer as long as the answer context is related to the question. The types of named entities considered for general topics include the following: *person, location, organization, money, date, time, number, percentage, temperature, ordered number, mass, height,*

length, period, energy, power, area, space, distance and *object*, as listed in Table 5.4. All named entities are identified with an algorithm based on BBN's IdentiFinder [21].

6.2.2. Question Formulation Algorithm

Each query from the TREC novelty tracks has three fields: title, description and narrative. Even though not explicitly provided in the format of questions, a significant number of the queries can be transformed into multiple specific questions. As examples, query 306 from TREC 2002 novelty track and query N37 from TREC 2003 are listed here as Example 6.1 and Example 6.2.

Example 6.1. Question formulation - specific questions

<Number>: 306

<Title> African Civilian Deaths

<Description>: How many civilian non-combatants have been killed in the various civil wars in Africa?

<Narrative>: A relevant document will contain specific casualty information for a given area, country, or region. It will cite numbers of civilian deaths caused directly or indirectly by armed conflict.

Example 6.2. Question formulation - a general question

< Number>: N37

<Title> Olympic scandal Salt Lake City

<Topic Type> event

<Description>: The Salt Lake City Olympic bribery scandal: What were the results?

< Narrative>: Any mention of effects of the bribery scandal was relevant. Mention of things that should not have been affected, such as, souvenir sales or attendance are relevant.

Mention of members being involved in the scandal was relevant. Any mention of kinds of bribery used was relevant. Effects on potential donors or sponsors of the SLC games or future games were relevant.

There are many approaches that can be used for question formulation and pattern determination. In our current implementation, we used a simple algorithm based on word-pattern matching to formulate questions and corresponding information patterns from queries. For each type of the five NE-questions, a number of word patterns, which could be unigram, bi-gram (or further n-gram) word sequences, have been constructed for question type identification. Some word patterns were extracted from the TREC 2002 novelty track queries manually and some patterns were selected from Li & Croft's question answering system [20]. The patterns we have shown Table 4.1 (in Chapter 4) are listed here in Table 6.1 for easy reference. In the table, most of them are unigram, but some of them are bi-gram.

Table 6.1. Word patterns for the five types of NE questions

Categories	No	Answer Types	Word Patterns
Name	0	Person	who, individual, person, people, participant, candidate, customer, victim, leader, member, player, name
	1	Organization	who, company, companies, organization, agency, agencies, name, participant
	2	Location	where, location, nation, country, countries, city, cities, town, area, region
Time	3	Date	when, date, time, which year, which month, which day
Number	4	Number	how many, how much, length, number, polls, death tolls, injuries, how long,

Table 6.2 shows the pseudo code of our question formulation algorithm. For a given query, the algorithm will go through the text in both the description and the narrative fields to identify terms that matches some word-patterns in the list. The query analysis component first tries to formulate at least two specific questions for each query, if possible, because a single specific question probably only covers a small part of a query. This follows NE Observations # 7 and #8 about the

impact of combined NE types – please refer to the List of Observations and Conclusions for a summary of these two and other observations.

Table 6.2. ip-BAND: question formulation algorithm

Step 0. **Load Word Pattern Table** WPT

WPT[at] is a list of word patterns, where answer type at = 0 - 4, representing Person, Organization, Location, Date and Number (see Table 6.1)

Step 1. **Initialize Word-pattern Matching Table** WMT

WMT[at] = 0, at = 0 - 4;

Step 2. **Fill Word pattern Matching Table** WMT

For each word W in both Description and Narrative fields

For each at (Person, Organization, Location, Date and Number)

If W is in WPT[at] then WMT[at]++;

Step 3. **Formulate multiple NE questions or one general question**

Set number of questions Nq = 0;

For (at = 0 to 4) if (WPT[at] != 0) Nq++;

If (Nq >1) then // **multiple specific questions**

{

Question Type QT = Specific;

Answer Types {AT} = {T_{NE}}; // specific Types of NEs from WMT

}

else // **a general question**

{

Question Type QT = General;

Answer Types {AT} = {POLD NEs + Words}; // POLD NEs and general words

}

Step 4. **Extract query words {Wq} and topic type** TT

{Wq} = {all the words in the Title field after stopword removal and stemming};

TT = event or opinion;

Step 5. **Generate information patterns** IP

IP = (QT, {Wq}, {AT}, TT} (6.1)

For a specific topic:

IPs = (Specific, {Wq}, {T_{NE}}, TT} // TT is not used (6.2)

For a general topic:

IP = (General, {Wq}, {Words + POLD}, TT} (6.3)

For the topic in Example 6.1, we will have WMT[2] = 3, and WMT[4] = 2, since in the Description and the Narrative fields, three word patterns "area", "country" and "region" match the word patterns in the answer type Location, and two word patterns "how many" and "numbers" matches the word patterns in the answer type Number. Therefore two specific questions about "Where" and "How many" are formulated, hence the topic is identified as a specific topic. In summary the information pattern IPs (Eq. 6.2) for this topic will have

QT = Specific

{Wq} = { African, Civilian, Deaths }

{AT} = {T_{NE}} = { Location, Number }

If a query only has terms that match with patterns belonging to one type of question, or it does not have any matched terms at all, then a general question is generated for the query. For a general topic, the topic type (event or opinion), as part of the information pattern, is given in the TREC 2003 and 2004 novelty tracks. A number of expected opinion patterns have been identified manually and are listed in Table 4.2 (in Chapter 4).

For the topic in Example 6.2, we will have WMT[0] = 1, since in the Narrative fields, a word pattern "member" matches the one in the answer type Person. Because only one specific question about "Who" is formulated, the topic is classified as a general topic. So in its information pattern IPg (Eq. 6.3), we have

QT = General

{Wq} = {Olympic, scandal, Salt, Lake, City }

{AT} = {POLD NEs, in addition to general words}

TT = event

6.3. Using Patterns in Relevance Re-Ranking

6.3.1. Relevant Sentence Retrieval Algorithm

The task of the relevant sentence retrieval module is to retrieve sentences that are likely to have potential answers to the questions transformed from a given topic. It first takes query words of the query, {Wq}, and searches in its data collection to retrieve sentences that are topically relevant to the query. It then re-ranks the sentences retrieved using the corresponding information patterns determined at the step of query analysis. Sentences that do not satisfy the query-related patterns for *specific questions* are filtered out because they are unlikely to have potential answers without the required answer patterns (this directly follows NE Observation #11).

Our relevant sentence retrieval module is implemented based on an existing search toolkit -- LEMUR [25]. The module first uses a TFISF model adopted from TFIDF models in LEMUR for the first round of relevant sentence retrieval (for details please see Section 6.3.2). Then it removes the sentences that do not contain any "answers" to the potential question. For a specific question, only a specific type of named entity that the question expects would be considered for potential answers. Thus a sentence without an expected type of named entity will be removed from the list. A list of presumed *answer sentences* (which contain expected named entities to the question) is generated. For general questions (topics), all types of named entities as well as single words could be potential answers. Therefore, no filtering is performed after the first round of sentence retrieval using the TFISF model (this follows NE Observation #3).

To improve the performance of finding relevant sentences and to increase the ranks for sentences with the required information patterns, sentence-level information patterns, including sentence lengths, Person-Location-Date NEs, and opinion patterns, are incorporated in the relevance retrieval step. Table 6.3 outlines our relevant sentence retrieval algorithm, which is self-explanatory. The use of information patterns seems to have attracted attention for other

applications. Recent work by Kumaran and Allan [58] tried a pattern-based approach for document retrieval. By asking users simple questions, frequent bigrams within a window of eight terms and phrases in queries were identified and used to improve pseudo relevance feedback results.

Table 6.3. ip-BAND: relevant sentence retrieval algorithm

Step 1	**First Round of Relevant Sentence Retrieval**
	Using {Wq} and IFISF model with pseudo feedback {XWq} (in Eq. 6.7)
	to obtain a relevant sentence set
	{Relevant sentences}, ranking from highest to lowest with the TDISF score S_0 (in Eq. 6.6)
Step 2.	**Answer Sentence Extraction**
	if QT = Specific then
	{Answer sentences} = {Relevant sentences} -{ Sentences without NEs in $\{T_{NE}\}$} (6.4)
	else if QT = General then
	{Answer sentences} = {Relevant sentences} (6.5)
Step 3.	**Answer Sentence Re-Ranking**
	For each sentence in {Answer sentences}
	Adjust relevance score S_0 by incorporating
	Sentence length adjustment (Eq. 6.8),
	PLD NEs adjustment (Eq. 6.9) and
	Opinion patterns adjustment (Eq. 6.10)
	=> {Re-ranked answer sentences}

6.3.2. Ranking with TFISF Models

TFIDF models are one of the typical techniques in document retrieval. TF stands for Term Frequency in a document and IDF stands for Inverse Document Frequency with respect to a document collection. The *term frequency* in the given document gives a measure of the importance of the term within the particular document, which is the number of times the term appears in a document divided by the number of total terms in the document. The *inverse document frequency* is a measure of the general importance of the term, which is the logarithm of

the number of all documents in the collection divided by the number of documents containing the term [44]. In information retrieval systems using TFIDF models, each query term is assigned a weight based on the TF value and IDF value of the term. Documents in the collection are ranked by relevance score, which is calculated based on the TFIDF weights of matched query terms in documents.

We adopt TFIDF models for the relevant *sentence* retrieval step in our novelty detection task simply because it was also used in other systems and was reported to be able to achieve equivalent or better performance compared to other techniques in sentence retrieval [7]. The name of our sentence retrieval model is called TFISF model, to indicate that inverse sentence frequency is used for sentence retrieval instead of inverse document frequency. The initial TFISF relevance ranking score S_0 for a sentence, modified from the LEMUR toolkit [25, 41], is calculated according to the following formula

$$S_0 = \sum_{i=0}^{n} [w(t_i)\, tf_s(t_i)\; tf_q(t_i)\; isf^2(t_i)] \qquad (6.6)$$

where n is the total number of terms, $isf\ (t_i)$ is inverse sentence frequency (instead of inverse document frequency in document retrieval), $tf_s(t_i)$ is the frequency of term t_i in the sentence, and $tf_q(t_i)$ is the frequency of term t_i in the query. The inverse sentence frequency is calculated as

$$isf(t_i) = \log \frac{N}{N_{t_i}},$$

where N is the total number of sentences in the collection, N_{ti} is the total number of sentences that include the term t_i. Note that in the above formulas, t_i could be a term in the original query (with a weight $w(t_i) = 1$) or in an expanded query that has more terms from pseudo feedback (with a weight $w(t_i) = 0.4$). The expanded query with pseudo relevance feedback is represented by

{XWq} = {Wq} + {Top 50 most frequent words from the 1st 100 sentences retrieved} (6.7)

With pseudo relevance feedback, the system assumes that top 100 sentences retrieved are relevant to the query and top 50 most frequent terms within the 100 sentences are added to the original query.

As in other Information Retrieval (IR) systems, stopword removal and word stemming are performed in a preprocessing step for relevant sentence retrieval. Stopword removal is to filter out common words that do not bear much information. Word stemming is to deal with morphology so that variations of a word may be treated equally. There are 419 stop words, such as "about", "almost", etc., in our stopword list. They have been removed from all sentences in the relevant sentence step. The stopword list is a standard list that was used in many systems, such as [17] and [20]. We use the Krovetz Stemmer [43] for word stemming in our system. The Krovetz Stemmer is a light stemmer. It effectively and accurately removes inflectional suffixes in three steps, the conversion of a plural to its single form (e.g. '-ies', '-es', '-s'), the conversion of past to present tense (e.g. '-ed'), and the removal of '-ing'. The conversion process firstly removes the suffix, and then though a process of checking in a dictionary for any recoding (also being aware of exceptions to the normal recoding rules), returns the stem to a word.

The score S_0 will be served as the baseline for comparing the performance increase with the information patterns we have proposed in relevant sentence retrieval and novelty detection. We have tried both the TFISF model with original queries and the TFISF model with expanded queries by pseudo feedback. The TFISF model with pseudo feedback is used in the experiments reported in the rest of the thesis because it provides better performance.

6.3.3. TFISF with Information Patterns

The TFISF score is adjusted using the following three information patterns: sentence lengths, named entities, and opinion patterns. The length-adjustment is calculated as

$$S_1 = S_0 * (L / \bar{L}) \tag{6.8}$$

where L denotes the length of a sentence and $\bar{L}$ denotes the average sentence length. In the current implementation, the average sentence length is set as the same for all topics in each collection of TREC 2002, 2003 and 2004 novelty tracks. This parameter is not critical as long as the comparison is only among sentences for the same topic. A sentence that is longer than the average will increase the TFISF score. This adjustment follows SL Observations #1 and #2.

The NEs-adjustment is computed as

$$S_2 = S_1 * [1 + \alpha(F_{person} + F_{location} + F_{date})] \tag{6.9}$$

where $F_{person} = 1$ if a sentence has at least a person's name, 0 otherwise; $F_{location} = 1$ if a sentence has at least a location name, 0 otherwise; and $F_{date} = 1$ if a sentence has at least a date, 0 otherwise. This adjustment follows NE Observations, particularly #1, #5 and #6. The adjustment is applied to both specific topics and general topics, but the parameter α is set slightly different. For specific topics or general but opinion topics, $\alpha = 0.4$; for general, event topics, $\alpha = 0.5$. This follows the NE Observation #9 (OP #3).

Finally, the opinion-adjustment is computed as

$$S_3 = S_2 * [1 + \beta F_{opinion}] \tag{6.10}$$

where $F_{opinion}$ = 1 if a sentence is identified as an opinion sentence with one or more opinion patterns, 0 otherwise. A number of patterns (i.e. "said", "argue that", see Table 5.2) are used to determine whether a sentence is an opinion sentence. This adjustment follows the OP Observations #1 and #2. This final adjustment step based on opinion patterns is *only* performed for general opinion topics. The opinion pattern information is not used for any queries that convert into specific questions, nor is applied to event topics.

We apply the three adjustments sequentially to tune the parameters on training data based on TREC 2003 novelty track for best performance. We have also tried different ways of adjusting scores and found that above adjustment mechanism (Eqs. 6.8-6.10) achieves the best performance. Details can be found in Appendix A. All three adjustments are applied to the training data to find the best set of parameters, α and β, and the same set of parameters are used for all data sets.

Incorporating information patterns at the retrieval step should improve the performance of relevance and thus help in the following novelty extraction step. After applying the above three steps of adjustments on the original ranking scores, sentences with query-related information patterns are pulled up in the ranked list. For the following two sentences in Example 6.3, which are also shown in Example 1.2 in Chapter 1, the relevant (and novel) sentence (sentence 1) was ranked 14th with the original TFISF ranking scores. It was pulled up to the 9th place in the ranked list after the adjustments with the information patterns. The non-relevant sentence (sentence 2) was initially ranked 2nd, but pushed down to the 81st place after the score adjustments. Complete comparison results of novelty detection performance on TREC 2002, 2003 and 2004 are provided in the experiments in the next chapter.

Example 6.3. Relevant re-ranking with information patterns

Sentence 1 (Relevant and Novel): "The court's ruling confirms that the entire campaign to *ban 'partial-birth abortion*' -- a campaign that has consumed Congress and the federal courts for over three years -- is nothing but a fraud designed to rob American women of their right to *abortion,***" said** Janet Benshoof, president of Center for Reproductive Law and Policy.

Sentence 2 (Non-relevant): Since the Senate's last *partial birth* vote, there have been 11 court decisions on the legal merits of *partial birth bans* passed by different states.

6.4. Novel Sentence Extraction

In general, the new sentence detection module extracts all query-related named entities and single words as possible answers from each answer sentence and detects previously unseen "answers". In this section, we first discuss how novel sentence extraction works for specific and general topics. Then we provide a unified algorithm for novel sentence extraction for both specific and general topics.

6.4.1. Novel Sentence Extraction: How It Works

We start with our treatment for specific topics for easy explanation. For a specific topic with multiple specific NE questions that are formulated from the topic (query), an answer sentence may have answers to one or more of these specific NE questions. So named entities related to any one of the specific questions in the answer sentences should be extracted, and are treated as answers. There is an answer pool associated with each topic, which is initially empty. As

sentences come in, in the order of its re-ranked relevant ranking, new answers (i.e., new and specific NEs) will be added to the answer pool when the novel sentence detection module determines that the incoming answers are previously unseen. A sentence will be marked novel if it contains new answers. Sentences without new answers will be removed from the final list provided to the user.

For a general topic, the novelty score of a sentence is calculated with the following formula:

$$S_n = \omega N_w + \gamma N_{ne} \quad (6.11)$$

where S_n is the overall novelty score of a sentence S, N_w is the number of new words (including NEs) in S that do not appear in its previous sentences, N_{ne} is the number of POLD-type named entities in S that do not appear in its previous sentences, and ω and γ are weights for new words and new named entities, respectively. In fact, Eq. (6.11) is used for both general and specific topics, with different set of parameters (ω, γ and T), where T is the threshold for S_n. The unified novel sentence extraction algorithm is outlined in Table 6.4, and will be discussed in the following sub-section.

Note that stopwords in our stopword list (Table 6.4) have been removed from all sentences in the relevant sentence step, and will not be considered in the process of novelty score calculation. Then the words are stemmed. A sentence is identified as a novel sentence if its novelty score is equal to or greater than a preset threshold. In our experiments, the best performance of novelty detection is achieved when both ω and γ are set to 1, and the threshold T for S_n is set to 4.

We would like to make three notes here.

(1). Named entities considered at the novelty extraction step include all POLD types, i.e., PERSON, ORANIZATION, LOCATION and DATE. Unlike in the relevance sentence retrieval step, the ORGANIZATION type is also considered in this step (NE Observation #2). A previously unseen organization may contain new information.

(2). By the summation of the counts in both new words and new named entities, those relevant sentences that do not include any named entities could also be selected as novel sentences if their novelty score is greater than the preset threshold (NE Observation #4).

(3). The novelty score formula given in Eq. 6.11 is actually a general one that can also be applied to specific topics. In that case, N_{ne} s the number of the specific answer NEs, and we set ω to 0. The threshold for the novelty score S_n is set to 1.

6.4.2. A Unified Novel Sentence Extraction Algorithm

The unified novel sentence extraction algorithm is summarized in Table 6.4. The algorithm is described in pseudo-code and it is self-explanatory. The next paragraph refers to this algorithm.

In Table 6.4, a single answer pool is used for each topic. Ideally, a separate answer pool should be created for each required answer type for a topic. That is to say, for a specific topic with M NE-questions, we should have M answer pools. For a general topic, we should have an answer pool for the general words, and other four separate answer pools for the four possible types of POLD NEs - PERSON, ORGANIZATION, LOCATION and DATE. However, in the current implementation, for coding and space efficiency, we use a single answer pool implemented by a hash table for all types of answers of each topic. Due to the use of the hash table, the search and insertion of a new word in the answer pool (a hash table) is still very efficient in time.

Table 6.4. ip-BAND: novel sentence extraction algorithm

Step 1 **Initialization**

Novel sentence list {Novel sentences} = { };

Answer pool A = {answers} = { };

Topic type TT = Specific or General; // from Table 6.2

New answer threshold T = 1 (if TT = Specific) or 4 (if TT = General) in Eq. 6.11;

Step 2. **Novelty Score Calculation**

For each sentence S in the set of {Answer sentences} // from Table 6.3

Ws = collections of words in S; // after stopword removal and word stemming

if TT = Specific then {

// A_s is defined as Answers in S; {T_{NE}} is defined in Table 6.2

Answers A_s = { NEs of the required answer types {T_{NE}} in Ws}

$N_w = 0$; N_{ne}= NewWord(A, A_s); // see the end of the table for a description

}else{ // TT = General

// A_w is defined as all the answer words, A_{ne} is defined as all the POLD NEs

$A_w = W_s$;

A_{ne} = { NEs of Person, Organization, Location and Date types in Ws};

Answers $A_s = A_w + A_{ne}$;

N_w = NewWord(A, A_w);; N_{ne}= NewWord(A, A_{ne});

}

Novelty score $S_n = \omega N_w + \gamma N_{ne}$; // Eqs. 6.11

Step 3. **Novel Sentence Determination**

if S_n >= T then S is declared novel and is added into {Novel sentences};

NewWord **New Word Counting Function** N = NewWord (A, A_s)

N = 0;

For each word w in the word set A_s

if w is not in A then {

add w into A;

N++;

}

return N;

6.5. Summary

Here we summarize the main features of our unified pattern-based approach for novelty detection (ip-BAND) in using the proposed information patterns at the sentence level.

First, information patterns are defined and determined based on question formulation (in the query analysis step) from queries, and are used to obtain answer sentences (in the relevant sentence retrieval step) and new answer sentences (in the novel sentence detection step).

Second, NE information patterns are used to filter out sentences that do not include the specific NE word patterns in the relevance retrieval step, and information patterns are incorporated in re-ranking the relevant sentences for favoring those sentences with the required information patterns, and therefore with answers and new answers.

Third, new information patterns are checked in determining if a sentence is novel or not in the novel sentence extraction step. Note that after the above two steps, this step becomes relatively simple; however, we want to emphasize that our ip-BAND approach for novelty detection include all the three steps.

Finally, we use a unified approach to novelty detection for both specific NE questions and general questions, in all the three steps (query analysis, relevance detection and novelty detection), and in all the formulas.

Chapter 7

EXPERIMENTAL RESULTS AND ANALYSIS

In this chapter, we present and discuss the main experimental results. The data used in our experiments are from the TREC 2002, 2003 and 2004 novelty tracks. The comparison of our approach and several baseline approaches are described. The parameters used in both the baselines and our approaches were tuned with the TREC 2002 data, except the parameter β in Equation 6.10, which is used for sentence re-ranking for opinion topics in our approach. The best value of β was chosen based on the performance for the opinion topics from the TREC 2003 novelty track. Recall that the topics from the TREC 2003 and 2004 novelty tracks were classified as either "opinion" or "event" topics, but there was no such classification for the TREC 2002 topics. The experiments and analysis include the performance of novelty detection for specific topics using the named entity patterns and sentence lengths, general topics using all the three proposed information patterns (sentence lengths, named entities and opinion patterns), and the overall performance of novelty detection using the unified pattern-based approach.

7.1. Baseline Approaches

We compared our information-pattern-based novelty detection (ip-BAND) approach to four baselines described in the following subsections. They are:

B-NN: baseline with initial retrieval ranking (without novelty detection),

B-NW: baseline with new word detection,

B-NWT: baseline with new word detection with a threshold, and

B-MMR: baseline with maximal marginal relevance (MMR).

For comparison, in our experiments, the same retrieval system based on the TFISF techniques adopted from the LEMUR toolkit [25] is used to obtain the retrieval results of relevant sentences in both the baselines and our ip-BAND approach. The evaluation measure used for performance comparison is precision at rank N (N =5, 10, 15, 20 and 30 in Tables 7.1- 7.12). It shows the fraction of correct novel (or relevant) sentences in the top N sentences delivered to a user, which is defined as

$$precision(N) = \frac{\text{number of novel (relevant) sentences retrieved}}{\text{N (number of sentence retrieved)}} \tag{7.1}$$

Note that precision at top ranks is useful in real applications where users only want to go through a small number of sentences.

7.1.1. B-NN: Initial Retrieval Ranking

The first baseline does not perform any novelty detection but only uses the initial sentence ranking scores generated by the retrieval system directly as the novelty scores (without filtering and re-ranking). One purpose of using this baseline is to see how much novelty detection processes (including NE filtering, relevant sentence re-ranking and new answer detection) may help in removing redundancies.

7.1.2. B-NW: New Word Detection

The second baseline in our comparison is simply applying new word detection. Starting from the initial retrieval ranking, it keeps sentences with at lease one new word that does not appear in previous sentences as novel sentences, and removes those sentences without new words from the

list. As a preprocessing step, all words in the collection were stemmed and stopwords were removed (see Section 6.3.2).

New words appearing in sentences usually contribute to the novelty scores used to rank sentences by various approaches [e.g., 7], but new words do not necessarily contain new information. Our proposed ip-BAND approach considered new information patterns as possible answers to potential questions of queries. Comparing our ip-BAND approach to this baseline helps us to understand which is more important in containing new information: new words (this baseline), or new PLOD named entities (for general questions), or new "answers" of specific NEs (for specific questions).

Note that the novel sentence extraction step in our ip-BAND approach for specific topics determines if a sentence is novel by checking if the sentence includes at least one new NE of the required NE types for the topic. In this sense, the baseline B-NW is designed to mainly compare with ip-BAND for specific topics.

7.1.3. B-NWT: New Word Detection with a Threshold

The third baseline (B-NWT) is similar to B-NW. The difference is that it counts the number of new words that do not appear in previous sentences. A sentence is identified as novel sentence if and only if the number of new words is equal to or greater than a preset *threshold*. The best value of the threshold is 4 in our experiments. This means that a sentence is treated as a novel sentence only if more than 3 new words are included in the sentence. The threshold is selected as the same as for our ip-BAND novel sentence determination for general topics (Eq. 6-11). So, roughly speaking, this baseline is comparable to the ip-BAND applied to general topics.

7.1.4. B-MMR: Maximal Marginal Relevance (MMR)

Many approaches to novelty detection, such as maximal marginal relevance (MMR), new word count measure, set difference measure, cosine distance measure, language model measures, etc.

[6-13,24], were reported in the literature. The MMR approach was introduced by Carbonell and Goldstein [24] in 1998, which was used for reducing redundancy while maintaining query relevance in document re-ranking and text summarization. In our experiments, the MMR baseline approach (B-MMR) starts with the same initial sentence ranking used in other baselines and our ip-BAND approach. In B-MMR, the first sentence is always novel and ranked top in novelty ranking. All other sentences are selected according their MMR scores. One sentence is selected and put into the ranking list of novelty sentences at a time. MMR scores are recalculated for all unselected sentences once a sentence is selected. The process stops until all sentences in the initial ranking list are selected. MMR is calculated by

$$MMR = \arg \max_{S_i \in R/N} \left[\lambda(Sim_1(S_i, Q) - (1-\lambda) \max_{S_j \in N} Sim_2(S_i, S_j) \right] \quad (7.2)$$

where S_i and S_j are the *i*th and *j*th sentences in the initial sentence ranking, respectively, Q represents the query, N is the set of sentences that have been currently selected by MMR, and R/N is the set of sentences have not yet selected. Sim_1 is the similarity metric between sentence and query used in sentence retrieval, and Sim_2 can be the same as Sim_1 or can be a different similarity metric between sentences.

We use MMR as our fourth and main baseline because MMR was reported to work well in non-redundant text summarization [24], novelty detection at document filtering [13] and subtopic retrieval [17]. Also, MMR may incorporate various novelty measures by using different similarity matrix between sentences and/or choosing different value of λ. For instance, if cosine similarity metric is used for Sim_2 and λ is set to 0, then MMR would become the cosine distance measure reported in [7].

We want to note here that our ip-BAND approach also uses the same initial relevance ranking as all the baselines. The novelty detection performance of our ip-BAND is a combined impact of information patterns in sentence filtering for specific topics, relevant sentence re-ranking and new pattern detection. However, in order to see what is the impact of information patterns for the relevant sentence retrieval step and the novel sentence extraction step, we have also conducted experiments comparing relevant sentence retrieval. In the following section, we present the results of the comparisons of the performance of both relevance and novelty detection, and for both specific and general topics. Finally we present the overall performance for all the topics in the three novelty track data collections.

From Table 7.1 to Table 7.12, Chg% denotes the percent change of the precision of our ip-BAND approach (or the 2^{nd}, 3^{rd} or 4^{th} baseline) compared to the first baseline. The Wilcoxon test is used for significance test, with * and ** indicating the 95% and 90% confidence levels, respectively. Here we briefly explain why we select the Wilcoxon test instead of the t-test and the sign test. The t-test is the standard test for matched pairs. However, if the populations are non-normal, particularly for small samples, the t-test may not be valid. The Wilcoxon signed rank test is an alternative that can be applied when distributional assumptions are suspect, even though it is not as powerful as the t-test when the distributional assumptions are in fact valid. The sign test is another alternative to the t test for paired samples when the normality assumption is in doubt. The Wilcoxon signed rank test is generally preferred over the sign test because it takes into account both the sign of the difference and the magnitude of the difference for paired samples, whereas the sign test only takes the difference of the sign into account. Since the sample size in our experiments is very small, varying from 10 to 50, hence the distributions are not guaranteed to be normal, therefore we select the Wicoxon signed rank test.

7.2. Experimental Results and Comparisons

7.2.1. Experimental Results for Specific Topics

First, we tested the information-pattern-based novelty detection (ip-BAND) approach on those specific topics on the data from the TREC 2002, 2003 and 2004 novelty tracks, and compared it to the aforementioned four baselines. Two sets of experimental results are shown here, which are (a) performance of identifying novel sentences for queries that were transformed into multiple specific questions (with query words and specific NE answer types); and (b) performance of finding relevant sentences for these specific topics.

The purpose of the first set of experiments, whose results are shown in Tables 7.1 to 7.3, is to compare the performance of our pattern-based approach to the four baselines for queries with specific question formulations. Our query analysis algorithm formulated multiple specific questions for 8 out of 49 queries from the TREC 2002 novelty track, for 15 out of the 50 queries from the TREC 2003 novelty track, and 11 out of the 50 queries from the TREC 2004 novelty track, respectively. The precision measurements in the tables are the averages of all the specific topics in each data collection at several different ranks. We have the following conclusions (#1 and #2)) on the experimental results for specific topics.

Conclusion #1. The proposed approach outperforms all baselines at top ranks for specific topics.

Conclusions from our experiments are listed in the List of Observations and Conclusions at the beginning of the thesis. Tables 7.1 - 7.3 show comparisons for the top 5, 10, 15, 20 and 30 sentences. For example, the performance of our approach with specific questions beats the first baseline by more than 30% at rank 15 on the data from all the three novelty tracks. The improvements over the first baseline are 35.0%, 35,7% and 31.6%, for TREC 2002, 2003 and 2004, respectively. In comparison, the best results combining all the three baselines (No 2, No 3

and No 4) are only 10.0% (from N-NWT), 32.1% (from N-NWT) and 10.5% (from both N-NW and N-NWT), respectively. In another words, within the top 15 sentences, our approach obtains more novel sentences than all four baselines. The precision values (see Tables 7.1 - 7.3) of our ip-BAND approach are 22.5%, 67.6% and 30.3%, respectively. These indicate that within the top 15 sentences, there are 3.4, 10.1 and 4.6 sentences on average that are correctly identified as novel sentences by the ip-BAND approach (from TREC 2002, 2003 and 2004 respectively). The numbers are 1.5, 4.8 and 1.6 for the combinations of the best results of baselines 2 to 4. For many users who only want to go through a small number of sentences for answers, the number of novel sentences in the top ranks is a more meaningful metric than the F measure, which is a combination of precision and recall.

Conclusion #2. For specific topics, New Word Detection with a Threshold (B-NWT) performs slightly better than New Word Detection (B-NW), but Maximal Marginal Relevance (B-MMR) does not.

On the surface, for specific topics, the new word detection baseline approach and our ip-BAND approach use the similar strategy. That is, once a new word (new specific NE) appears in a sentence, it is declared as a novel sentence. However, a new specific NE in our ip-BAND approach may answer the right question, but a new word in B-NW will often not. Therefore B-NWT is better than B-NW in the sense that it needs more words (4 in our experiments) to declare a sentence to be novel.

The performance of the baseline based on *Maximal Marginal Relevance* in fact is worse than the first baseline that simply uses the results from initial relevant ranking at low recalls. This might indicate that while MMR has good performance in text summarization to obtain summaries of documents, it seems that it does not work well for novelty detection. The reason could be due to the fact that MMR uses a combined score to measure both relevance and redundancy, which are

somewhat in conflict. In reducing redundancy, it might also decrease the performance in relevant ranking, which is very important in novelty detection.

Table 7.1. Performance of novelty detection for 8 specific topics (queries) from TREC 2002 (Note: Data with * pass significance test at 95% confidence level by the Wilcoxon test and ** for significance test at 90% level – same applies to Tables 7.1 – 7.12)

Top# Sentences	B-NN	B-NW		B-NWT		B-MMR		ip-BAND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.1750	0.2000	14.3	0.1750	0.0	0.1750	0.0	0.2000	14.3
10	0.1625	0.1750	7.7	0.2000	23.1	0.1375	-15.4	0.2500*	53.8*
15	0.1667	0.1667	0.0	0.1833	10.0	0.1500	-10.0	0.2250*	35.0*
20	0.1750	0.1750	0.0	0.1875	7.1	0.1562	-10.7	0.1938	10.7
30	0.1625	0.1708	5.1	0.1792	10.3	0.1458	-10.3	0.1750	7.7

Table 7.2. Performance of novelty detection for 15 specific topics (queries) from TREC 2003

Top# Sentences	B-NN	B-NW		B-NWT		B-MMR		ip-BAND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.5067	0.5867**	15.8**	0.5867**	15.8**	0.5200	2.6	0.6933*	36.8*
10	0.5400	0.5867	8.6	0.6467*	19.8*	0.5600	3.7	0.6933*	28.4*
15	0.4978	0.6133*	23.2*	0.6578*	32.1*	0.6000*	20.5*	0.6756*	35.7*
20	0.5200	0.6300*	21.2*	0.6667*	28.2*	0.5867*	12.8*	0.6800*	30.8*
30	0.5133	0.6022*	17.3*	0.6844*	33.3*	0.5978*	16.5*	0.7000*	36.4*

Table 7.3. Performance of novelty detection for 11 specific topics (queries) from TREC 2004

Top# Sentences	B-NN	B-NW		B-NWT		B-MMR		ip-BAND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.2000	0.2364	18.2	0.2364	18.2	0.2182	9.1	0.2727	36.4
10	0.2455	0.2364	-3.7	0.2455	0.0	0.2455	0.0	0.2909	18.5
15	0.2303	0.2545	10.5	0.2545	10.5	0.2364	2.6	0.3030*	31.6*
20	0.2455	0.2545	3.7	0.2773	13.0	0.2500	1.9	0.3182*	29.6*
30	0.2394	0.2515**	5.1**	0.2818	17.7	0.2727	13.9	0.3152*	31.6*

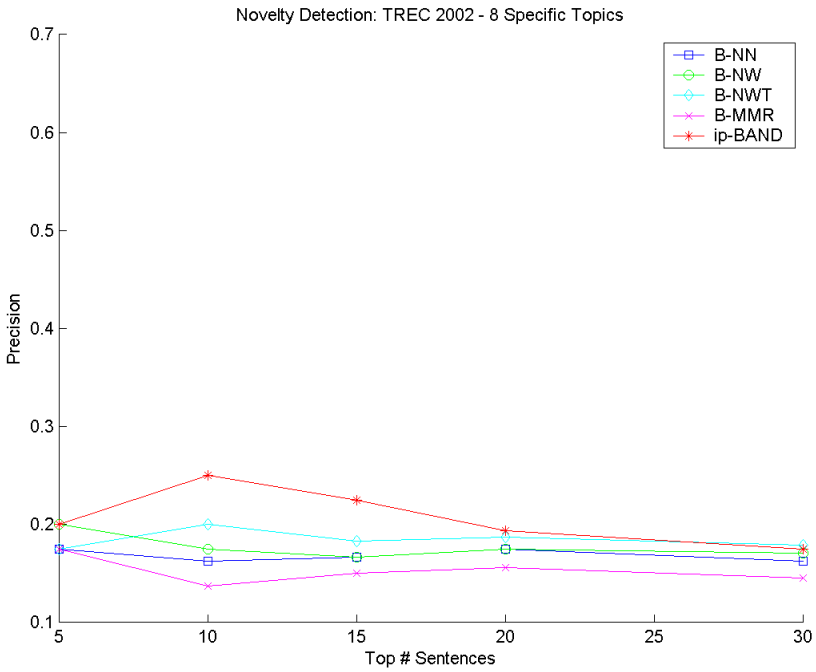

Figure 3a. Performance of novelty detection for specific topics (TREC 2002)

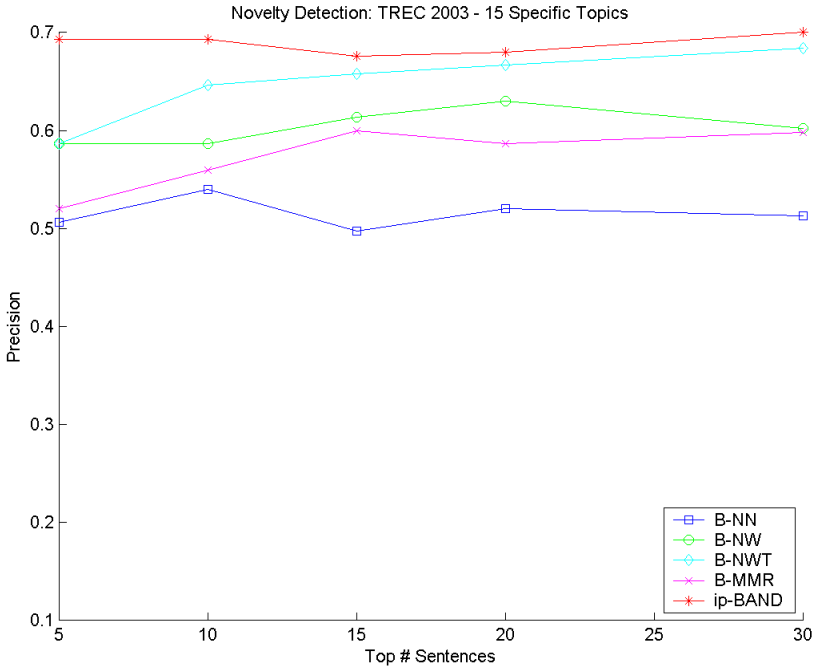

Figure 3b. Performance of novelty detection for specific topics (TREC 2003)

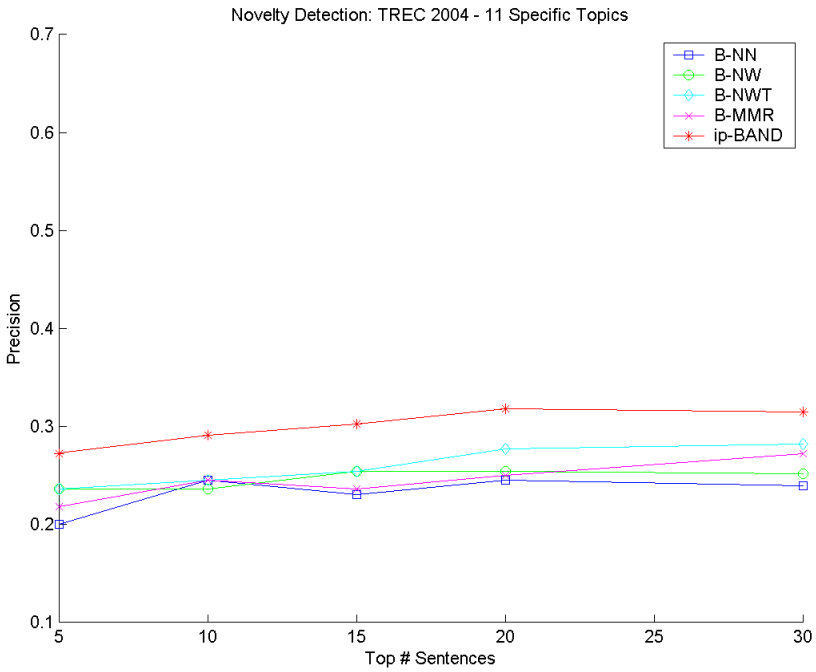

Figure 3c. Performance of novelty detection for specific topics (TREC 2004)

The results of novelty detection for the specific topics are also plotted in Figure 3 for easy comparison. These graphs clearly show that our ip-BAND approach significantly beats all the baselines at top ranks (5, 10, 15, 20, 30). By comparing the three graphs, we can see that the precision values (of our ip-BAND and the baselines) of TREC 2003 novelty track are significantly higher than those of TREC 2002 and 2004, and TREC 2002 has the lowest. The main reason is that in the novelty track data collection of the TREC 2003, the percentages of relevant sentences are much higher, but each of the TREC 2002 and 2004 novelty track collections only has a small portion of sentences in the document set marked as relevant sentences, and the percentages of relevant sentences are the lowest in the TREC 2002 (Chapter 3). These can also be seen in the following experiments on relevance performance comparison.

The second set of experiments is designed to investigate the performance gain of finding relevant sentences with the sentence *re-ranking* step for the specific topics. Remember that, in our ip-BAND approach, the relevant sentence retrieval module re-ranks the sentences by the revised scores that incorporate the information of sentence lengths and PLD-type named entities (i.e., Person-, Location- and Date-type NEs) into adjusting the TFISF scores (note: opinion patterns are not used for specific topics). In addition, sentences without the required named entities are also removed (i.e., filtered) before re-ranking. Our hypothesis is that this filtering plus re-ranking process would improve the performance of finding relevant sentences.

We compare the performance of finding relevant sentences with and without filtering and re-ranking. The comparison results are given in Table 7.4. From this Table, we can see that the retrieval improvements are noticeable for the TREC 2002 novelty tracks, the improvements for TREC 2003 are moderate, and for TREC 2004 improvements are seen for ranks 15, 20 and 30. One reason for the more significant improvements on TREC 2002 data collection could be that the impact of NEs and information patterns are more effective since the percentage of relevant sentences in this data collection is very low.

Nevertheless, the results in Tables 7.1 to 7.3 have shown that the pattern-based approach significantly outperforms all the four baselines at top ranks for identifying novel sentences, for all three novelty data collections. This indicates that our pattern-based approach makes a larger difference at the step of detecting novel sentences than at the step of finding relevant sentences for those specific topics, particularly from the TREC 2003 and 2004 novelty tracks. The reason could be that TREC 2003 and 2004 novelty track collections exhibited greater redundancy than the TREC 2002 and thus has less novel sentences, therefore our information patterns make a big difference here. In combination, information patterns play a balanced role in the two major steps: relevance retrieval (for TREC 2002 in particular) and novelty detection (for TREC 2003 and 2004 in particular).

Table 7.4. Performance of relevance for specific topics (queries) ($\alpha = 0.4$ in Eq. 6.8)

	TREC 2002 (8 Topics)			TREC 2003 (15 Topics)			TREC 2004 (11 Topics)		
Top # Sentences	TFISF	Length + NEs		TFISF	Length + NEs		TFISF	Length + NEs	
	Precision	Precision	Chg%	Precision	Precision	Chg%	Precision	Precision	Chg%
5	0.1750	0.2000	14.3	0.8800	0.9333	6.1	0.6727	0.5636	-16.2
10	0.1750	0.2250	28.6	0.8733	0.9200	5.3	0.6545	0.6182	-5.6
15	0.1750	0.2000	14.3	0.8356	0.9067**	8.5**	0.6061	0.6485	7.0
20	0.1875	0.2062	10.0	0.8567	0.8867	3.5	0.6136	0.6591	7.4
30	0.1708	0.1833	7.3	0.8444	0.8756**	3.7**	0.6030	0.6182	2.5

Table 7.5. Performance of relevance for general topics (queries) (Notes: (1) $\alpha = 0.4$ (2) α =0.5 for event topics, α =0.4, β=0.5 for opinion topics)

	TREC 2002 (41 Topics)			TREC 2003 (35 Topics)			TREC 2004 (39 Topics)		
Top # Sentences	TFISF	Length + NEs [1]		TFISF	Length + NEs + Opinion [2]		TFISF	Length + NEs + Opinion [2]	
	Precision	Precision	Chg%	Precision	Precision	Chg%	Precision	Precision	Chg%
5	0.2049	0.2488**	21.4**	0.6629	0.7086	6.9	0.4615	0.4564	-1.1
10	0.2171	0.2220	2.2	0.6200	0.7000*	12.9*	0.4359	0.4615	5.9
15	0.2114	0.2260	6.9	0.6343	0.6857*	8.1*	0.4308	0.4462	3.6
20	0.2000	0.2159	7.9	0.6386	0.6714**	5.1**	0.4141	0.4410*	6.5*
30	0.1870	0.2033**	8.7**	0.6371	0.6552	2.8	0.4026	0.4342*	7.9*

Table 7.6. Performance of relevance for all topics (queries)

	TREC 2002 (49 Topics)			TREC 2003 (50 Topics)			TREC 2004 (50 Topics)		
Top # Sentences	TFISF	Length + NEs [1]		TFISF	Length + NEs + Opinion [2]		TFISF	Length + NEs + Opinion [2]	
	Precision	Precision	Chg%	Precision	Precision	Chg%	Precision	Precision	Chg%
5	0.2000	0.2408**	20.4**	0.7280	0.7760**	6.6**	0.5080	0.4800	-5.5
10	0.2102	0.2224	5.8	0.6960	0.7660*	10.1*	0.4840	0.4960	2.5
15	0.2054	0.2218	7.9	0.6947	0.7520*	8.3*	0.4693	0.4907**	4.5**
20	0.1980	0.2143	8.2	0.7040	0.7360*	4.5*	0.4580	0.4890*	6.8*
30	0.1844	0.2000*	8.5*	0.6993	0.7213**	3.1**	0.4467	0.4747*	6.3*

7.2.2. Experimental Results for General Topics

Before we show the overall performance of our unified ip-BAND approach, we also want to see how information patterns can improve the performance of novelty detection for those general topics that cannot be easily turned into multiple specific NE questions. As described in Chapter 6, all the three types of information patterns (i.e., sentence lengths, PLD-type NEs and opinion patterns) are incorporated in the relevance retrieval step of novelty detection for general topics. Table 7.5 gives the performance comparison of relevance retrieval with the original TFISF ranking and with the adjustments using these sentence level information patterns for the TREC 2002, 2003 and 2004 data, respectively. Since in TREC 2002, topics are not classified into opinion and event topics, opinion patterns are not applied to this data collection.

The main conclusion here is that incorporating information patterns and sentence level features into TFISF techniques can achieve much better performance than using TFISF alone. Significant improvements are obtained for the 2003 topics and the 2004 topics at top ranks. For example, at rank 15, the incorporation of the information patterns increases the precision of relevance retrieval by 6.9%, 8.1% and 3.6%, for TREC 2002, 2003 and 2004, respectively. This lays a solid ground for the next step - new information detection, and therefore for improving the performance of novelty detection for those general topics.

The overall performance of relevance for all topics is given in Table 7.6. This table shows that re-ranking sentences with information patterns achieves better performance than using TFISF alone, for all topics. Readers can also compare the improvements between the specific and the general topics by comparing the results in Table 7.4 and Table 7.5. Generally speaking, information patterns play a greater role in relevance retrieval for general topics than for specific topics. We believe this is mostly due to the incorporation of opinion patterns into relevant sentence re-ranking for the general and opinion topics (for TREC 2003 and 2004).

Table 7.7. Performance of novelty detection for 41 general topics (queries) from TREC 2002

Top # Sentences	B-NN	B-NW		B-NWT		B-MMR		ip-BAND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.1902	0.1951	2.6	0.2049	7.7	0.2293	20.5	0.2390**	25.6**
10	0.2000	0.1951	-2.4	0.2049	2.4	0.2098	4.9	0.2098	4.9
15	0.1935	0.2000	3.4	0.2016	4.2	0.2033	5.0	0.2114	9.2
20	0.1854	0.1890	2.0	0.1939	4.6	0.1817	-2.0	0.2073	11.8
30	0.1748	0.1772	1.4	0.1707	-2.3	0.1691	-3.3	0.1902**	8.8**

Table 7.8. Performance of novelty detection for 35 general topics (queries) from TREC 2003

Top # Sentences	B-NN	B-NW		B-NWT		B-MMR		ip-BAND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.4229	0.4171	-1.4	0.4457	5.4	0.4343	2.7	0.5257*	24.3*
10	0.4143	0.4371**	5.5**	0.4657**	12.4**	0.4571	10.3	0.5257*	26.9*
15	0.4152	0.4400*	6.0*	0.4552**	9.6**	0.4438	6.9	0.5124*	23.4*
20	0.4057	0.4343*	7.0*	0.4686*	15.5*	0.4200	3.5	0.5029*	23.9*
30	0.3867	0.4238*	9.6*	0.4590*	18.7*	0.4219	9.1	0.4867*	25.9*

Table 7.9. Performance of novelty detection for 39 general topics (queries) from TREC 2004

Top # Sentences	B-NN	B-NW		B-NWT		B-MMR		ip-BAND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.2359	0.2359	0.0	0.2410	2.2	0.2359	0.0	0.2154	-8.7
10	0.2026	0.2026	-0.0	0.2077	2.5	0.2026	0.0	0.2256	11.4
15	0.1949	0.2000	2.6	0.2051	5.3	0.1949	-0.0	0.2239*	14.9*
20	0.1859	0.1962*	5.5*	0.1974	6.2	0.1846	-0.7	0.2128*	14.5*
30	0.1735	0.1821**	4.9**	0.1846**	6.4**	0.1684	-3.0	0.1991*	14.8*

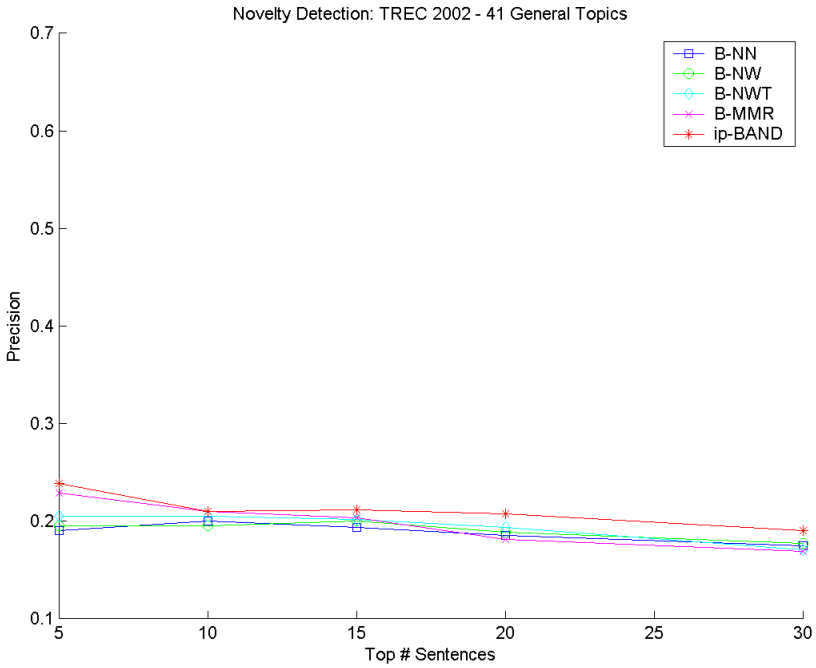

Figure 4a. Performance of novelty detection for general topics (TREC 2002)

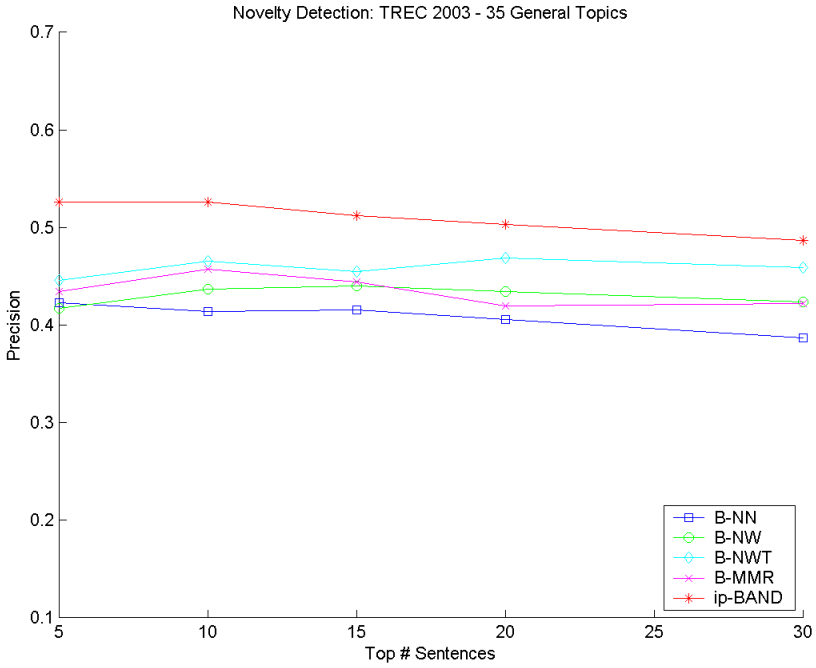

Figure 4b. Performance of novelty detection for general topics (TREC 2003)

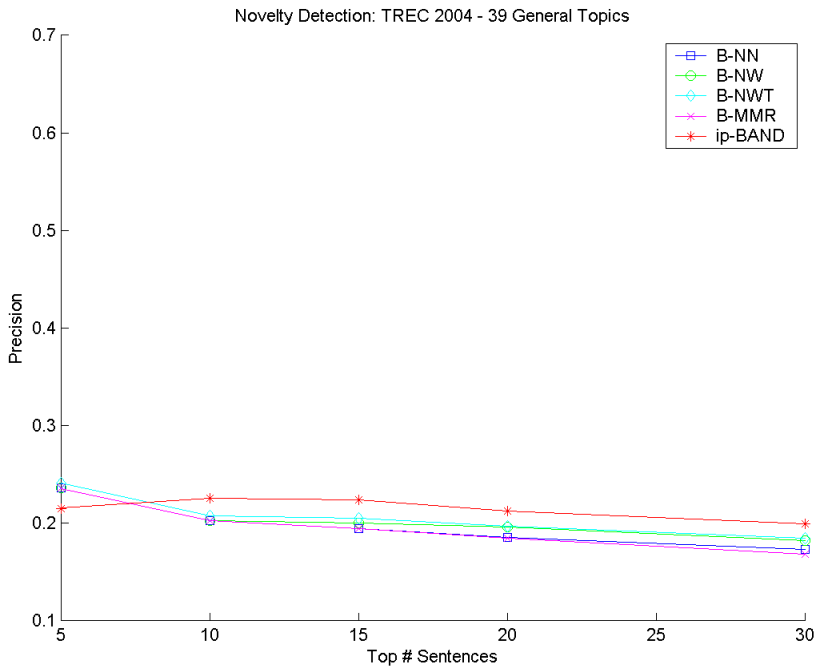

Figure 4c. Performance of novelty detection for general topics (TREC 2004)

We have also conducted experiments to apply opinion patterns to the relevance re-ranking for specific topics, and we found that the performance was worse than the results without opinion patterns as shown in Table 7.4. There could be two reasons. (1) Answers for specific topics are mostly specified by the specific required named entities, no matter the sentences include opinion patterns or not. (2). Answers for general, opinion topics are often indicated by opinion patterns, as have been shown in the statistics of opinion patterns in Chapter 5 (OP Observations #1 and #2).

Tables 7.7 - 7.9 show the performance comparison of our ip-BAND approach with the four baselines on those general topics that cannot be turned into specific NE questions. The results are also plotted in Figure 4 for easy comparison. We can draw the following conclusions from the results.

Conclusion #3. Our ip-BAND approach consistently outperforms all the baseline approaches across the three data sets: the 2002, 2003 and 2004 novelty tracks, for general topics.

The precision values for the top 15 sentences with our ip-BAND approach for general questions of the TREC 2002, 2003 and 2004 data are 21.1%, 50.1% and 22.4%, respectively. This represents that on average, 3.2, 7.5 and 3.4 sentences are correctly identified as novel sentences, among the top 15 recalls. Again, the precision is the highest for the 2003 data since this track has highest ratio of relevant to non-relevant sentences. Compared to the first baseline, the performance increases are 9.2%, 23.4% and 14.7%, respectively and the improvements are significant. In contrast, the best improvements combing all the three baselines (2-4) are only 5.0% (from B-MMR), 9.6% (from B-NWT) and 5.4% (from B-NWT), respectively, which are much lower than the results of our ip-BAND approach.

Conclusion #4. New Word Detection with a Threshold achieves better performance than New Word Detection for general topics.

This is within our expectation because New Word Detection is a special case of New Word Detection with a Threshold when the new word threshold is set to 1. In addition, both B-NWT and ip-BAND use the same threshold of 4 "new words" to declare a relevant sentence to be novel, except that ip-BAND also add additional counts for POLD-type NEs.

Conclusion #5. For general topics, Maximal Marginal Relevance is slightly better than* New Word Detection *and* New Word Detection with a Threshold *on the 2002 data, but it is worse than* New Word Detection with a Threshold *on the 2003 and 2004 data.

Conclusion #6. In comparison, the performance of our ip-BAND approach is slighter better for the specific topics than the general topics.

This may due to the fact that specific, targeted questions and answers are extracted for specific topics. This also indicates that as a future work, more improvements can be expected if more general topics can be turned into NE or other specific questions.

7.2.3. Overall Performance for All Topics

The overall performance comparisons of the unified pattern-based approach with the four baselines on all topics from the TREC 2002, 2003 and 2004 novelty tracks are shown in Tables 7.10, 7.11 and 7.12, respectively. The results for the three datasets are also plotted in Figure 4 for easy comparison. The most important conclusion is the following:

Conclusion #7. The unified pattern-based approach outperforms all baselines at top ranks for all topics.

Significant improvements are seen with the 2003 topics. In the top 15 sentences delivered, our approach retrieves 8.42 (=15*0.5613) novel sentences on average, while the four baseline approaches only retrieve 6.60, 7.38, 7.74 and 7.36 novel sentences, respectively. As anticipated, the overall performance for all topics (including both specific and general ones - Tables 7.10 - 7.12) is slightly better than that for the general topics (Tables 7.7 - 7.9), since the precision for the specific topics (Tables 7.1 – 7.3) are slightly higher than the general ones.

This comparison is also summarized in Table 7.13 at top 15 ranks (sentences). In the table, the following measures are listed for each case of specific, general and all topics, and for TREC 2002, 2003 and 2004: (1) the improvements over the first baseline in percentage (Chg%); (2) the number of correctly identified novel sentences; (3) the number of relevant but redundant sentences; and (4) the number of non-relevant sentences. The last three numbers add up to 15.

According to the results shown in this table, we have the following important observations that could guide further improvements in both relevant and novel sentence detection for different data collections.

(1) There are 3.20 novel sentences, 0.25 redundant sentences and 11.55 non-relevant sentences within the top 15 sentences retrieved for the TREC 2002 topics. It indicates that further improving the performance of relevant sentence retrieval is still a major task for the TREC 2002 topics.

(2) There are 8.42 novel sentences, 2.78 redundant sentences and 3.8 non-relevant sentences within the top 15 sentences retrieved for the TREC 2003 topics. The performance for the TREC 2003 topics is much better than the performance for the TREC 2002 topics; therefore there is less room for performance improvements on either relevant sentence retrieval or novel sentence extraction.

(3) There are 3.62 novel sentences, 3.74 redundant sentences and 7.64 non-relevant sentences within the top 15 sentences retrieved for the TREC 2004 topics. The performance of identifying novel sentences for the TREC 2004 topics is only slightly better than that for the TREC 2002 topics. Further performance gain can be obtained by improving relevant sentence retrieval and/or novel sentence extraction.

Here we want to emphasize that the ip-BAND approach is a unified one for both specific and general topics. As we have seen in the previous chapters, the main differences between the treatments for two classes of topics - specific and general - are the parameters of information patterns, i.e., for relevant sentence filtering (0 or 1), for relevant sentence re-ranking (α, β in Eq. 6.9 and Eq. 6.10), and for new answer detection (γ and T in Eq. 6.11). The procedures and formulas are virtually the same for both specific and general topics.

Table 7.10. Performance of novelty detection for 49 queries (all topics) from TREC 2002

Top# Sentences	B-NN	B-NW		B-NWT		B-MMR		ip-BAND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.1878	0.1959	4.3	0.2000	6.50	0.2204	17.4	0.2327**	23.9**
10	0.1939	0.1918	-1.1	0.2041	5.30	0.1980	2.1	0.2163	11.6
15	0.1891	0.1946	2.9	0.1986	5.00	0.1946	2.9	0.2136	12.9
20	0.1837	0.1867	1.7	0.1929	5.00	0.1776	-3.3	0.2051**	11.7**
30	0.1728	0.1762	2.0	0.1721	-0.40	0.1653	-4.3	0.1878**	8.7**

Table 7.11. Performance of novelty detection for 50 queries (all topics) from TREC 2003

Top# Sentences	B-NN	B-NW		B-NWT		B-MMR		ip-BAND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.4480	0.4680	4.5	0.4880	8.9	0.4600	2.7	0.5760 *	28.6*
10	0.4520	0.4820*	6.6*	0.5200*	15.0*	0.4880	8.0	0.5760*	27.4*
15	0.4400	0.4920*	11.8*	0.5160*	17.3*	0.4907*	11.5*	0.5613*	27.6*
20	0.4400	0.4930*	12.0*	0.5280*	20.0*	0.4700*	6.8*	0.5560*	26.4*
30	0.4247	0.4773*	12.4*	0.5267*	24.0*	0.4747*	11.8*	0.5507*	29.7*

Table 7.12. Performance of novelty detection for 50 queries (all topics) from TREC 2004

Top# Sentences	B-NN	B-NW		B-NWT		B-MMR		ip-BAND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.2280	0.2360	3.5	0.2400	5.3	0.2320	1.8	0.2280	0.0
10	0.2120	0.2100	-0.9	0.2160	1.9	0.2120	0.0	0.2400**	13.2**
15	0.2027	0.2120	4.6	0.2160	6.6	0.2040	0.7	0.2413*	19.1*
20	0.1990	0.2090*	5.0*	0.2150	8.0	0.1990	0.0	0.2360*	18.6*
30	0.1880	0.1973*	5.0*	0.2060*	9.6*	0.1913	1.8	0.2247*	19.5*

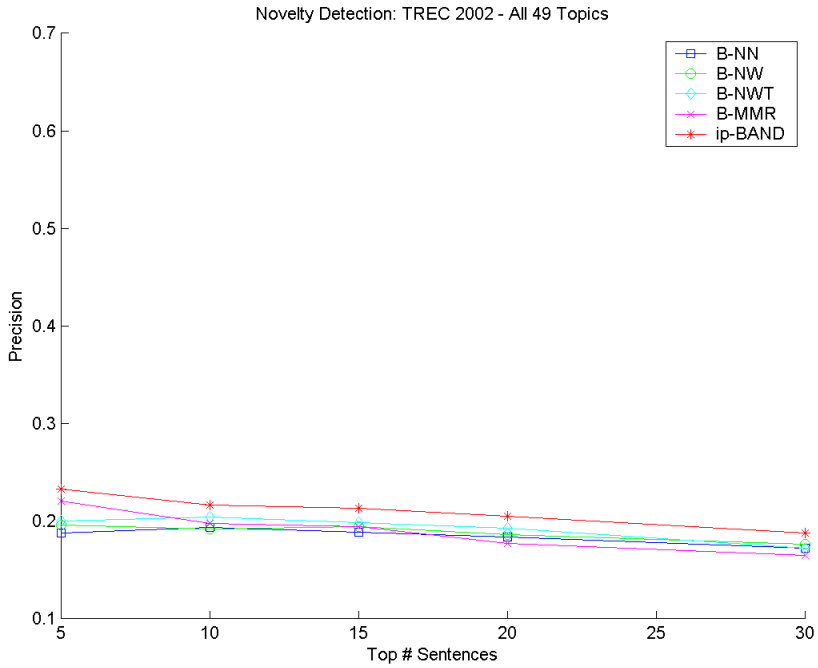

Figure 5a. Performance of novelty detection for all topics (TREC 2002)

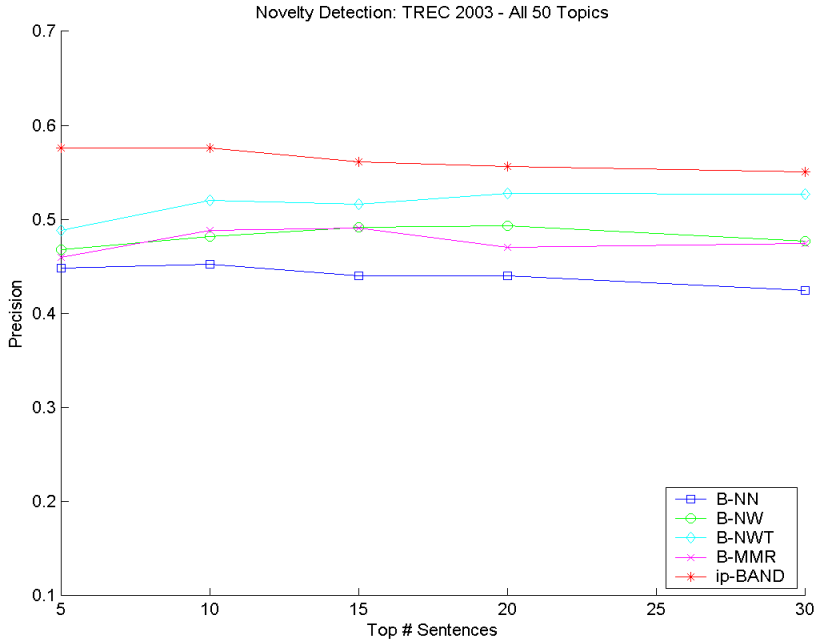

Figure 5b. Performance of novelty detection for all topics (TREC 2003)

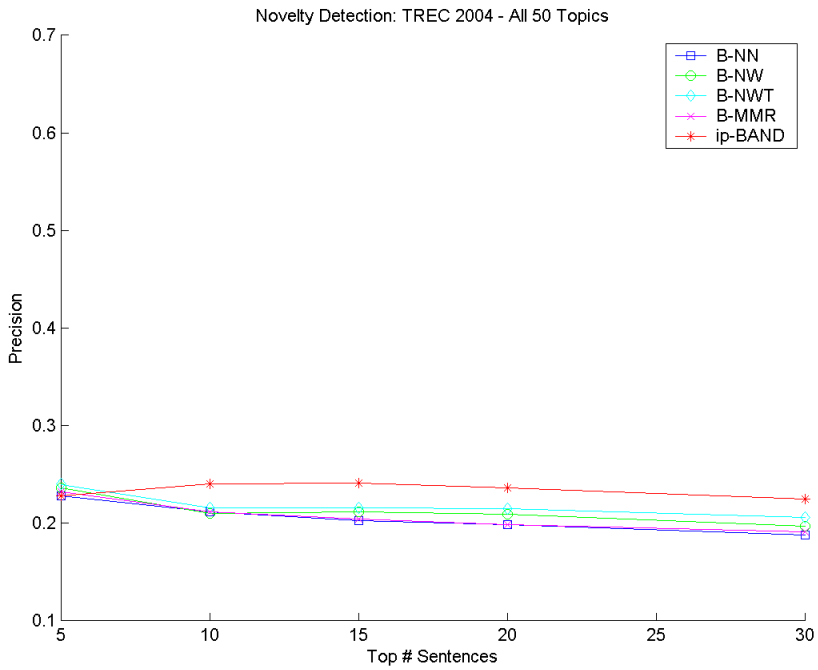

Figure 5c. Performance of novelty detection for all topics (TREC 2004)

Table 7.13. Comparison among specific, general and all topics at top 15 ranks

Data >	**TREC 2002**				**TREC 2003**				**TREC 2004**			
Topic|V	Chg%	Nvl#	Rdd#	NRl#	Chg%	Nvl#	Rdd#	NRl#	Chg%	Nvl#	Rdd#	NRl#
Specific	35.0*	3.38	0.12	11.50	35.7*	10.13	3.07	1.8	31.6*	4.55	5.42	5.03
General	9.2	3.17	0.27	11.56	23.4*	7.68	2.66	4.66	14.9*	3.35	3.34	8.31
All	12.9	3.20	0.25	11.55	27.6*	8.42	2.78	3.8	19.1*	3.62	3.74	7.64

Chg%: Improvement over the first baseline in percentage

Nvl#: Number of true novel sentences;

Rdd#: Number of relevant but redundant sentences;

NRl#: Number of non-relevant sentences

7.3. Discussions: Evaluation, Error and Question Formulation

The experimental results in Chapter 7 show that the proposed pattern-based approach (ip-BAND) gives the best performance among all approaches compared. But there is still much room left for further improvements. There are a few factors that may affect the performance of our system. Discussions and analyses of the factors will lead to some future research issues.

7.3.1. Evaluation Issues

The first factor is the problematic evaluation measure. A relevant sentence, pre-marked redundant by human assessors, could be treated as a novel sentence if it has new information that is not covered by previous sentences in the list provided by a novelty detection system. That is the case where the sentences that make it redundant in the list by human assessors are simply not retrieved by the system.

Relevance or novelty judgment files by human assessors are usually used as the basis for evaluating the performance of different systems. Unlike relevance measures, a novelty or redundancy measure is *asymmetric*. The novelty or redundancy of a sentence S_i depends on the order of sentences (S_1, …, S_{i-1}) that the user has seen before this one. For the TREC novelty track data, only the judgments for a particular set of sentences in a presumed order are available. To collect novelty judgments of each sentence with respect to all possible subsets, a human assessor has to read up to 2^{N-1} subsets. It is impossible to collect complete novelty judgments in reality.

There are two potential problems with this data. First, it is not very accurate to evaluate a system's performance if the ranked sentences of the system have a different order from the particular set. Second, assume sentence A and sentence B are relevant sentences to a topic and sentence A and sentence B contain the same information. If sentence A is before sentence B in the relevant set judged by human assessors, then B will be marked redundant. However, a system

might not retrieve sentence A but only B, or might retrieve both of them but B is in front of A in the ranked list. In this case B will be considered as a novel sentence in the result provided by the system. However it has been treated as redundant in the TREC novelty judgment file.

Some work has been done in this respect. For novelty detection data collected and used in their studies at CMU [13], researchers initially intended to collect judgments for 50 topics, but unfortunately they could only get assessments for 33 topics. They provide the information on which documents before a document makes it redundant. The documents must be listed in chronological order. Thus there are problems when evaluating a novelty detection system in which documents are not output in chronological order. As research interest increases in novelty detection, more accurate and efficient data collection is crucial to the success of developing new techniques in this area. Possible solutions to evaluation issues are further discussed in next chapter.

7.3.2. Error Analysis

The second factor is the misjudgment of relevance and/or novelty by human assessors and disagreement of judgments between the human assessors. Figure 6 and Figure 7 give the distribution of maximum similarity scores (MSSs) in novel and redundant sentences in the TREC 2003 and 2004 novelty tracks, respectively. The MSS score of a sentence is the maximum similarity score among the similarity scores between the sentence and each previous sentence. The similarity measure used here is cosine similarity function. In Figure 6, there are 83 novel sentences that have a maximum similarity score of 1. In Figure 7, there are 31 novel sentences that have a maximum similarity score of 1. That means the 83 novel sentences in Figure 6 and the 31 novel sentences in Figure 7 have exactly the same content as some sentences that occurred before them. These sentences should be marked redundant but were misjudged as novel sentences by human assessors in the judgment files provided by TREC. Therefore, there are at least 83 misjudgments and 31 misjudgments in TREC 2003 and TREC 2004 data, respectively. Note that

there are possible misjudgments on redundant sentences because there are also a number of redundant sentences that have a maximum similarity score of 0 in both Figure 6 and Figure 7.

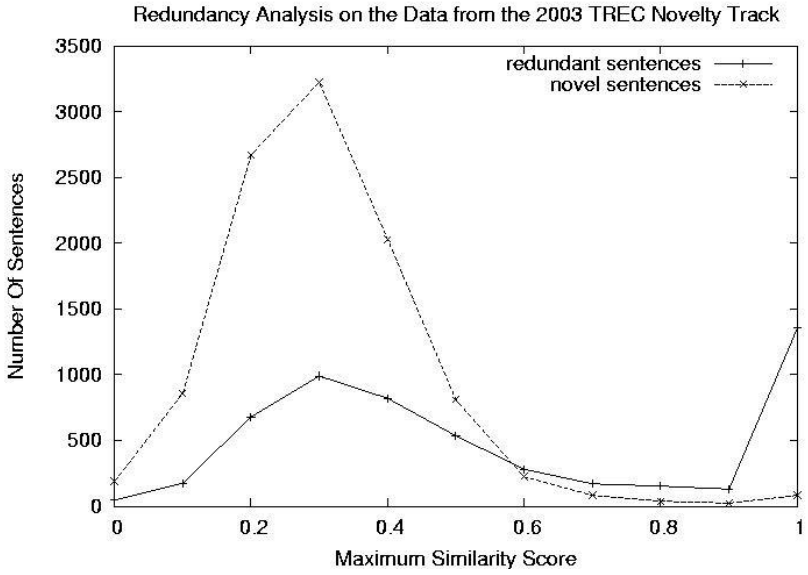

Figure 6. Redundancy analysis on data from the 2003 TREC novelty track

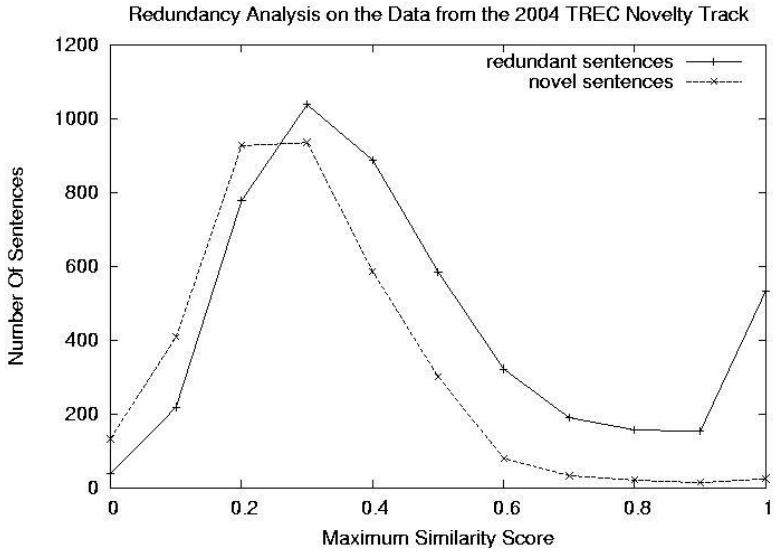

Figure 7. Redundancy analysis on data from the 2004 TREC novelty track

In addition to the misjudgments we discussed above, the human assessors for each topic also disagreed with each other about their judgments. While the assessor who marked a smaller number of relevant sentences (on a per topic basis) was the official assessor, the second assessor who marked the larger number of relevant sentences only picked about 60% of the official assessor's sentences, plus often many more sentences [6]. Possible solutions to the aforementioned judgment problems are discussed in Chapter 8.

7.3.3. Question Formulation Issues

The third factor is related to the limitation and accuracy of the NE question formulation in the proposed pattern-based approach. Currently, for specific topics, only five types of NE questions (PERSON, ORGANIZATION, LOCATION, NUMBER and DATE) are considered for query transformation. A specific topic is transformed into multiple NE questions, which may not completely cover the whole topic. Therefore some relevant or/and novel sentences may be missed because they are only related to the part of the topic that is not covered by the multiple NE questions, and do not contain answers to the multiple NE questions. An example of this case is shown in Example 7.1.

Example 7.1. A specific topic and one of its relevant sentences

<num>Number: N43

<title> Chinese earthquake

<toptype>event

<desc>Description: Events surrounding the Chinese earthquake, January 1998.

<narr>Narrative: Descriptions of the event, its **location** and extent, property damage and human **injuries** and deaths are relevant. Expressions of sympathy and relief donations, both Chinese and foreign, are relevant. Mention of reconstruction plans and results are relevant. Reasons for losses and aftershocks later in the year are not relevant.

(REL) <s docid="XIE19980520.0068" num="10"> Chen said that local farmers mainly rely on themselves in restoring their villages, although financial subsidies from the local government are available. </s>

The topic was transformed into two specific NE questions: LOCATION and NUMBER in the *question formulation* step. An answer sentence accordingly is expected to have a location and/or a number in order to answer at least one of the specific NE questions. Therefore, relevant sentences without any of the two expected types of named entities, such as the sentence shown in Example 7.1, were missed by our system due to the filtering process for specific topics in relevance sentence retrieval step. A topic may be fully covered by multiple specific questions if other types of questions in addition to NE questions are considered thus these missed sentences may be retrieved. For this example, a question about "reconstruction plans and results" will be able to include the missed sentence in the relevant sentence list.

For general topics, the proposed three information patterns only capture part of the characteristics of the required answers. Therefore, better question formulations for both general and specific topics will be one important direction of future work for further performance improvement. Possible solutions will be discussed in the next Chapter.

7.3.4. Other Issues

In addition to the three factors discussed above, there are other factors that may affect the performance of our system. For instance, both the performance of IdentiFinder, which is the named entity extractor used in the system, and the performance of sentence segmentator used by NIST, will affect the performance of our system directly but the degree of their impact needs further study. There are sentences that were not correctly segmented by NIST. One example of this case is shown in Example 7.2. The two sentences in this example, with named entities marked, was actually one sentence that should not be separated. The mistake of the segmentation

also led to an error by the IdentiFinder, which marked “m.” in the beginning of the second sentence as a person name.

Example 7.2. Errors of sentence segmentation

<s_ne docid = "XIE19980110.0208" num = "6"> The tremors, which occurred at <TIMEX TYPE = "TIME">11:50</TIMEX> a.</s_ne>

<s_ne docid = "XIE19980110.0208" num = "7"> <ENAMEX TYPE = "PERSON"> **m.** </ENAMEX>, were obviously felt in the <ENAMEX TYPE = "ORGANIZATION"> Chinese Capital </ENAMEX>, about <NUMEX TYPE = "LENGTH">300 km </NUMEX> Southeast to the epicenter. </s_ne>

Another example of errors made by the IdentiFinder is shown in Example 7.3, where “141 three weeks” were marked as a time period in stead of marking “287-141” as numbers and “three weeks ago” as a date.

Example 7.3. Errors of named entity extraction

<s_ne docid = "APW20000425.0103" num = "37"> The Republican-controlled <ENAMEX TYPE = "ORGANIZATION">House</ENAMEX> voted <NUMEX TYPE ="NUMBER">287</NUMEX>-<NUMEX TYPE = "PERIOD">**141 three weeks**</NUMEX> ago to outlaw partial-birth abortions – the <NUMEX TYPE ="ORDEREDNUMBER">third</NUMEX> time in <NUMEX TYPE="PERIOD">five years</NUMEX> such a ban was backed. </s_ne>

Chapter 8

CONCLUSIONS AND FUTURE WORK

Novelty detection is an important task to reduce the amount of redundant as well as non-relevant information presented to a user. It is an important component of many potential applications, such as new event detection, document filtering, and cross-document summarization. The motivation of this work is to explore new methods for novelty detection.

In this chapter, we will first conclude the thesis work on sentence level information patterns for novelty detection. Then we will discuss some future research issues.

8.1. Conclusions

In this thesis, we introduced a new definition of novelty, which is described as follows:

> *Novelty or new information means new answers to the potential questions representing a user's request or information need.*

Based on this definition, a unified pattern-based approach was proposed for novelty detection. Queries or topics are automatically classified into specific topics and general topics. Multiple effective query-related information patterns, namely sentence lengths, combinations of NEs, and opinion patterns, are identified and considered at both the retrieval step and the novelty detection step. Three sets of experiments were carried out on the data from the TREC novelty tracks 2002-2004, for specific topics, for general topics, and for all the topics. The experimental results show

that the proposed approach achieves better performance at top ranks than the four baseline approaches on topics from the novelty tracks of all the three years.

Here we summarize the main contributions of our unified pattern-based approach for novelty detection in using the proposed information patterns at the sentence level.

First, information patterns are defined and determined based on question formulation (in the query analysis step) from queries, and are used to obtain answer sentences (in the relevant sentence retrieval step) and new answer sentences (in the novel sentence detection step).

Second, NE information patterns are used to filter out sentences that do not include the specific NE word patterns in the relevance retrieval step, and information patterns (sentence lengths, named entities and opinion patterns) are incorporated in re-ranking the relevant sentences in favor of those sentences with the required information patterns, and therefore with answers and new answers.

Third, new information patterns are checked in determining if a sentence is novel or not in the novelty detection step. Note that after the above two steps, this step becomes relatively simple; however, we want to emphasize that our pattern-based approach for novelty detection include all the three steps.

8.2. Future Work

The proposed pattern-based approach opens up some further research issues. By solving these problems, the performance of novelty detection can be further improved. These issues include issues on question formulation and relevant retrieval models. We will discuss each of them in more detail in the following subsections. Solutions to evaluation issues are also discussed.

8.2.1. Improving Question Formulation

Even though we have significantly improved the performance of novelty detection for those "general" topics by using the proposed sentence level information patterns – sentence lengths, named entities and opinion patterns, the novelty detection precision for the specific topics that can be turned into specific questions are somewhat higher. Therefore, there are two ways to further improve the performance for general topics.

First, exploring the possibilities of turning more topics into multiple specific questions will be of great interests. Currently, only NE questions that require named entities for answers are considered. A topic may be not completely covered by the NE questions that were automatically transformed at the step of query analysis. For more topic coverage, other types of question should be considered in addition to NE questions. Therefore, future work should explore other types of question and discover their related patterns that may indicate the existence of potential answers.

One type of question that could be considered is "Why" question that usually asks for the cause of an event or the reason of an opinion. For this type of question, the occurrence of "because", "the reason is that", "due to", or "caused by" in a sentence may indicate possible answers for a "why" question. These words or phrases could be used as useful patterns for identifying relevant sentences.

There are other types of questions that could be considered, such as definition questions ("What" questions) and task questions ("How To" questions). Many other types of questions have been studied in question answering research community. A close monitoring of the development of question answering techniques and integrating them into the proposed pattern-based approach will further improve the performance of novelty detection.

On the other hand, for general topics, the proposed three information patterns only capture part of the characteristics of the required answers. First, sentence lengths and named entities (of Person,

Location and Date in particular) are used for all the topics, including specific and general topics. More of such information patterns, like document creation times (as we have used in the time-based language model [48], see also Section 8.2.2), could be helpful in further improving the performance of novelty detection for all topics. Second, opinion patterns are only used for general, opinion topics. Finding similar patterns for event topics could further improve the performance for those general, event topics. An *event pattern* such as "reported by XXX" where "XXX" could be a person's name or a news agency, might indicate the description of an event is included in a sentence. Some specific named entity combinations such as a specific time, date and location in a sentence could be another type of event pattern.

8.2.2. Incorporating Information Patterns into Retrieval Techniques

Currently, the pattern-based approach is combined with TFISF techniques (Adopted from TFIDF techniques in document retrieval), which is a simple and rather basic technique for sentence retrieval. The pattern-based approach starts with the retrieval results from the TFISF technique and adjusts the relevance ranking scores of sentences according to query-related information patterns, such as sentence lengths, combined named entities and opinion patterns that have been used in this thesis. However there are some limitations. Even with only three types of information patterns, it turns out to be rather difficult to incorporate them into relevance re-ranking. The sequential method that we have used (Appendix A.1) is somewhat ad hoc, and is far from optimal. This becomes more of a problem if more information pattern types are explored. How to systematically incorporating them into relevance retrieval and novelty detection is an interesting research issue. Therefore, in addition to investigating the incorporation of more information patterns with simple techniques such as TFISF/TFIDF models, combining information patterns with other retrieval approaches, such as language modeling approaches [46, 47], might be an

alternative (and better) way to systematically improve the performance of novelty detection using information patterns.

Language modeling frameworks were introduced into information retrieval by Ponte and Croft in 1998 [46]. It was followed by many variations [47, 48]. In the simplest case, the posterior probability of a document given in Equation. (8.1) is used to rank the document in the collection:

$$p(d / q) \propto p(q / d) p(d) \tag{8.1}$$

The prior probability of the document $p(d)$ is usually assumed to be uniform and is ignored for ranking.

The relevance-based language model was proposed by Lavrenko and Croft [47] in 2001. It is a model-based query expansion approach in the language-modeling framework. A relevance model is a distribution of words in the relevant class for a query. Both the query and its relevant documents are treated as random samples from an underlying relevance model R. The main challenge for a relevance-based language model is how to estimate its relevance model with no relevant documents available but only queries. Once the relevance model is estimated, the KL-divergence between the relevance model (of a query and its relevant documents) and the language model of a document can be used to rank the document. Documents with smaller divergence are considered more relevant thus have higher ranks. Equations (1) and (2) are the formulas [47] used for approximating a relevance model for a query:

$$P_o(w \mid R) \approx \frac{P(w, q_1 ... q_k)}{P(q_1 ... q_k)} \tag{8.2}$$

$$P(w, q_1 ... q_k) = \sum_{D \in \mathrm{M}} P(D) P(w \mid D) \prod_{i=1}^{k} P(q_i \mid D) \tag{8.3}$$

where $P_o(w \mid R)$ stands for this original relevance model of the query and its relevant documents, in which $P(w, q_1 \ldots q_k)$ stands for the total probability of observing the word w together with query words $q_1 \ldots q_k$. A number of top ranked documents (say N) returned with a query likelihood

language model are used to estimate the relevance model. In Equation (8.3) *M* is the set of the N top ranked documents used for estimating the relevance model for a query (together with its relevant documents). *P(D)* is the prior probability to select the corresponding document language model *D* for generating the total probability in Equation (8.3). In the original relevance model approach, a uniform distribution was used for the prior.

The time-based language model by Li and Croft [48] was designed to improve performance for queries that favor recent documents. They studied the relationship between time and relevance and showed how time could be incorporated into both query likelihood language model and relevance-based language model. Their approach was to change the uniform priors *p(d)* in query likelihood model (in Equation 8.1) and *P(D)* in relevance-based model (in Equation 8.3) with an exponential distribution (given in Equation 8.4) for queries where recency is a major requirement of a user's information need.

$$p(D) = P(d) = \lambda e^{-\frac{(T_C - T_D)^2}{2}} \tag{8.4}$$

Here T_C is the most recent date (in month) in the whole collection and T_D is the creation date of a document. The exponential distribution assigns higher priors to more recent documents.

In a similar fashion to the time-based language model, some information patterns identified in a sentence can by used to change the prior of the sentence in both query likelihood language model and relevance-based language model. For instance, the priors of sentences in a collection can be assigned according to sentences lengths. One way of incorporating sentence lengths into the prior assignments is shown in Equation 8.5.

$$p(S) = P(s) = \frac{|s|}{\sum |s_i|} \tag{8.5}$$

where p(s) and p(S) are the priors of a sentence in query likelihood language model and relevance-based language model, respectively. |s| is the length of a sentence and $\sum |s_i|$ is the

total length of all sentences in the collection (or top N sentences used for estimating the relevance model).

Equation 8.5 shows one way to incorporate sentence lengths into language models. This is only one of the information patterns that have been considered in our approach. The other two types of information patterns, named entities and opinion patterns, may need difference ways to be incorporated into language models. The Indri retrieval model proposed by Metzler and Croft [56, 57] combines the language modeling approaches [46] and inference network approaches in the Inquery system [54] to information retrieval. This model allows structured queries to be evaluated using language modeling estimates within the network, rather than TFIDF estimates in the Inquery system. The inference network model in the Indri retrieval model is composed of four types of node: document nodes d_i, representation nodes r_k, query nodes q_j and information need node *I*. A document node d_i corresponds to the event that a document is observed. There is a document node for each document in the collection. Representation nodes r_k can be any easily indexed features of a document. The beliefs of representation nodes in Indri are estimated with probabilities from smoothed language models. Query nodes q_j lie between representation nodes r_k and the information node *I*. They define how the belief of document relevance should depend on the document representation. The query nodes can be single terms or proximity operators. The *I* node represents the event the information need is met. It is a query node that combines all of the evidence from other query nodes into a single belief. The belief score of a document can be calculated by starting from the d_i node it corresponds to and propagating beliefs through the network all the way down to the information need node.

The Indri query language is composed of operators, each of which can be considered as a query node in an inference network. The operators allow users to construct detailed queries that are more expressive and complex than short keyword queries. There are some operators that correspond to representation nodes, such as single terms, ordered window, unordered window,

phrase and field operators, etc. These operators can combine the beliefs of representation nodes. The other operators are belief operators that correspond to query nodes, such as NOT, AND, OR, MAX, SUM, Weighted SUM and Weighted AND operators. These belief operators allow users to combine many different kinds of evidence in the network.

As a future work, the named entity patterns (person, location, and date) and opinion patterns proposed in this thesis could be incorporated into the Indri model with the following approach. First, the index of sentences is built with person, location, date, and opinion tags. Then four representation nodes, which represent person, location, date, and opinion, respectively, are added into the inference network of the Indri model. The field operator "#any" can be used for search the existence of a named entity pattern or opinion pattern in a sentence. Last, the belief operator "#WAND", weighted AND operator, will combine the beliefs at the query nodes corresponding to the patterns with the beliefs of other query nodes. Using this approach, more information patterns discovered in future study could be incorporated into the Indri model.

We want to note here that a major problem of using language modeling approaches in sentence retrieval is the data sparsity problem. Sentences are much shorter than documents. With a short sentence, it is difficult to approximate the true language model behind the sentence. Therefore, directly applying language modeling approaches to sentence retrieval may not achieve good performance. Murdock & Croft [49] made a first attempt and showed that smoothing a sentence from its local context could improve retrieval over using the query likelihood model alone.

8.2.3. Generating Dynamic Novelty Judgment Files for Performance Evaluation

As we have discussed earlier in Chapter 7, relevance or novelty judgment files by human assessors are used as the basis for evaluating the performance of different systems. In the TREC

novelty tracks, the judgment of novel sentences was based on a particular set of relevant sentences in a presumed order. There are two potential problems with this judgment file. First, it is not very accurate to evaluate a system's performance if the ranked sentences of a novelty detection system have a different order from the particular set. Second, a relevant sentence, pre-marked redundant by human assessors, could be treated as a novel sentence if it has new information that is not covered by previous sentences in the list provided by a novelty detection system. That is the case where the sentences that make it redundant in the list by human assessors are simply not retrieved by the system.

One possible solution to novel sentence judgment is to generate dynamic novelty judgment files at evaluation time. Given a topic, instead of creating a fixed set of novel sentences based on the set of relevant sentences in a presumed order, relevant sentences can be classified into *novelty groups*. In each group, all sentences basically contain the same information with respect to the topic. Therefore, if a user reads any sentence from a *novelty group*, the rest of the sentences in the group will become redundant sentences to the user because they do not contain any new information.

To evaluate the performance of a novelty detection system, a dynamic novelty judgment file can be generated at evaluation time. Given the ranked list of sentences from the system and the set of relevant sentences classified in novelty groups for a topic, the dynamic novelty judgment file can be generated by simply scanning the sentences in the ranked list and selecting sentences that are in the novelty groups of the topic and are the first sentence appearing in the ranked list from each novelty group. With the dynamic judgment file, the standard evaluation tools provided by TREC can still be used to evaluate the performance of the novelty detection system. Evaluations with the dynamic novelty judgment file should be more accurate than using a fixed novelty judgment file based on a presumed order of sentences. An ideal system should select one sentence from each novelty group and deliver these sentences to the user. The ranked list of sentences provided by

such an ideal system will be a complete set of novel sentences for the topic in the sense that it contains all relevant information related to the topic. The performance of the ideal system is then 100% for both precision and recall.

Appendix

A.1. Experiments for Incorporating Information Patterns in Relevance Re-Ranking

We apply the three adjustments sequentially to tune the parameters on training data for best performance of relevant sentence retrieval, and the same parameters are used for all data sets. We have also tried different ways of adjustments as will discussed below and found that above adjustment mechanism (Eqs. 6.8 - 6.10) achieves the best performance.

A.1.1. Sentence Lengths Adjustments

In incorporating the information of sentences lengths into relevant sentence re-ranking, we have tried the following methods: +Length method, *ClippedLength method and *Length method. They are all compared with the baseline, in which the original TFISF ranking score S_0 is simply normalized by the maximum possible ranking score for all sentences in consideration.

(1) **+Length** method:

The length adjustment with this method is to "add" length factor to the original score, and it is calculated as

$$S_1 = S_0/S_{max} + \alpha\ (L/\underline{L}) \qquad (A.1)$$

where S_0 is the original ranking score for a sentence, S_{max} is the maximum possible ranking score for all sentences in consideration, L is the length of the sentence and $\underline{L}$ is the average sentence length on a data collection. We have also tried different values of the parameter α, and find that $\alpha = 0.2$ gives the best performance when only sentence length adjustment is applied to relevant sentence re-ranking.

(2) *ClippedLength method

The length adjustment with this method is to "multiple" a roof-clipped length factor and it is calculated as

$$S_1 = S_0/S_{max} * \min(1.0, L/\underline{L}) \quad (A.2)$$

in which only penalty is given to sentences shorter than average.

(3) ***Length** method

The length adjustment with this method is "multiple" a length factor and it is calculated as

$$S_1 = S_0/S_{max} * (L/\underline{L}) \quad (A.3)$$

This is essentially the same as the method we put into real use for our ip-BAND approach (Eq. 6.8).

Table A.1 shows the results of the performance comparison with different sentence length adjustment methods to 49 topics from TREC 2002. It is obvious that *Length method outperforms the other two significantly. Therefore we use the *Length method in our ip-BAND approach.

Table A.1. Performance of relevant sentence retrieval with sentence length adjustments for 49 topics from TREC 2002

Top # Sentences	(0) Original	(1) +Length		(2) *Clipped Length		(3) *Length	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.2000	0.2286	14.3	0.2041	2.0	0.2286	14.3
10	0.2020	0.1878	-7.1	0.1980	-2.0	0.2041	1.0
15	0.2000	0.1891	-5.4	0.1946	-2.7	0.2136	6.8**
20	0.1939	0.1847	-4.7	0.1878	-3.2	0.1969	1.6
30	0.1810	0.1741	-3.8	0.1850	2.3**	0.1844	1.9

** significant test at 90% confidence level by the Wilcoxon test; * significant test at 95% level

A.1.2. Named Entities Adjustments

The following three methods are compared with the baseline with normalized relevant ranking: +ALL, +POLD, +PLD, *ALL, *POLD, *PLD. The *PLD method is the same as Eq. 6.9 we use in our ip-BAND approach.

(1) **+ALL** method

This method counts how many different types of all the named entities (NEs) in a sentence, and "add" the all-NE factor to the original ranking score. It is calculated as

$$S_2 = S_0/S_{max} + \alpha N_{all} \quad (A.4)$$

where N_{all} is the occurrence of different types of all the named entities (NEs, as listed in Table 5.4) in the sentence, and the parameter a is 0.05 for the best performance (we used the same parameters for all the methods in this comparison).

(2) **+POLD** method

This method counts how many different types of four types of named entities (PERSON, ORGANIZATION, LOCATION and DATE) in a sentence, and "add" the POLD-NE factor to the original ranking score. It is calculated as

$$S_2 = S_0/S_{max} + \alpha N_{POLD} \quad (A.5)$$

where N_{POLD} is the occurrence of different types of all the POLD-type named entities (NEs) in the sentence.

(3) **+PLD** method

This method counts how many different types of three types of named entities (PERSON, LOCATION and DATE) in a sentence, and "add" the PLD-NE factor to the original ranking score. It is calculated as

$$S_2 = S_0/S_{max} + \alpha N_{PLD} \quad (A.6)$$

where N_{PLD} is the occurrence of different types of all the PLD-type named entities (NEs) in the sentence.

(4) ***ALL** method

This method counts how many different types of all the named entities (NEs) in a sentence, and "multiple" the all-NE factor to the original ranking score. It is calculated as

$$S_2 = S_0/S_{max} \ (1+ \alpha N_{all}) \quad (A.7)$$

where N_{all} is the occurrence of different types of all the named entities (NEs) in the sentence.

(2) ***POLD** method

This method counts how many different types of four types of named entities (PERSON, ORGANIZATION, LOCATION and DATE) in a sentence, and "multiple" the POLD-NE factor to the original ranking score. It is calculated as

$$S_2 = S_0/S_{max} \ (1+ \alpha N_{POLD}) \quad (A.8)$$

where N_{POLD} is the occurrence of different types of all the POLD-type named entities (NEs) in the sentence.

(3) ***PLD** method

This method counts how many different types of three types of named entities (PERSON, LOCATION and DATE) in a sentence, and "multiple" the PLD-NE factor to the original ranking score. It is calculated as

$$S_2 = S_0/S_{max} + (1+\alpha N_{PLD}) \quad (A.9)$$

where N_{PLD} is the occurrence of different types of all the PLD-type named entities (NEs) in the sentence. The *PLD method is essentially the one we use in our ip-BAND approach.

Table A.2 shows the results of the performance comparison with different named entities adjustment methods to 49 topics from TREC 2002. It is obvious that *PLD method outperforms the all the other five methods significantly. Note that we not only compare different ways to apply named entities, i.e., "adding" or "multiplying", but we have tried different combinations of named entities. It is clear that using only Person/Location/Data-type named entities gives the best performance, with the two different calculations. Between the two ways of calculations, "multiplying" method is clearly better. Therefore we use the *PLD method in our ip-BAND approach.

Table A.2. Performance of relevant sentence retrieval with named entity adjustments for 49 topics from TREC 2002

Top # Sentences	(0) Originnal	(1) +ALL		(2) +POLD		(3) +PLD	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.2000	0.2200	10.2	0.2204	10.2	0.2327*	16.3*
10	0.2020	0.2041	1.0	0.2082	3.0	0.2000	-1.0
15	0.2000	0.2054	2.7	0.1973	-1.4	0.2041	2.0
20	0.1939	0.1959	1.1	0.1959	1.1	0.2051	5.8**
30	0.1810	0.1905	5.3	0.1857	2.6	0.1837	1.5

** significant test at 90% confidence level by the Wilcoxon test; * significant test at 95% level

Top # Sentences	(0) Originnal	(4) *ALL		(5) *POLD		(6) *PLD	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.2000	0.2082	4.1	0.2286*	14.3*	0.2449*	22.4*
10	0.2020	0.2082	3.0	0.2082	3.0	0.2122	5.1
15	0.2000	0.2095	4.8	0.2000	0.0	0.2190	9.5**
20	0.1939	0.1990	2.6	0.1929	-0.5	0.2102	8.4**
30	0.1810	0.1891*	4.5*	0.1857	2.6	0.1857	2.6

** significant test at 90% confidence level by the Wilcoxon test; * significant test at 95% level

A.1.3. Opinion Patterns Adjustments

In incorporating the information of opinion patterns into relevant sentence re-ranking, we have tried the following methods: +Opinion method and *Opinion method. They are both compared with the baseline with original TFISF ranking score S_0 normalized by the maximum possible ranking score for all sentences in consideration.

(1) **+Opinion** method

The opinion-adjustment with this method is computed as

$$S_3 = S_0 / S_{\max} *[\beta F_{opinion}] \quad \text{(A.10)}$$

where $F_{opinion}$ = 1 if a sentence is identified as an opinion sentence with one or more opinion patterns, 0 otherwise.

(2) ***Opinion** method

The opinion-adjustment with this method is computed as

$$S_3 = S_0 / S_{\max} *[1 + \beta F_{opinion}] \quad \text{(A.11)}$$

where $F_{opinion}$ = 1 if a sentence is identified as an opinion sentence with one or more opinion patterns, 0 otherwise. This is essentially the same we used in our ip-BAND approach. In both methods we set $\beta = 2$.

Table A.3 shows the results of the performance comparison with the two different sentence length adjustment methods to 49 topics from TREC 2002 with the parameter $\beta = 2$. It can be seen that the precisions in relevant sentence retrieval by using the two methods are very close. And we have also found that different values of this parameter do not change the performance very much. Therefore either one of them could be chosen into our ip-BAND approach. However, to be consistent with the other two adjustments (the length adjustment and the NE adjustment), we select the *Opinion method in our current implementation.

Table A.3. Performance of relevant sentence retrieval with opinion pattern adjustments for 22 topics from TREC 2003

Top # Sentences	(0) Original	(1) +Opinion		(2) *Opinion	
	Precision	Precision	Chg%	Precision	Chg%
5	0.5909	0.6455	9.2	0.6455	9.2
10	0.6000	0.6227	3.8	0.6227	3.8
15	0.6030	0.6576	9.0**	0.6576	9.0**
20	0.6068	0.6477	6.7**	0.6455	6.4**
30	0.6000	0.6545	9.1**	0.6530	8.8

** significant test at 90% confidence level by the Wilcoxon test; * significant test at 95% level

Finally we apply all the three adjustment and tune the parameters for the best performance of relevant sentence retrieval. To simplify the computation, we apply the three adjustments sequentially. The sentence length adjustment is applied first using Eq. 6.8 (or Eq. A.3). No parameters need to be tuned for this adjustment.

Then the NE adjustment is applied with different values of the parameter α, using Eq. 6.9 (or Eq. A.9). We found that for general, opinion topics, the best performance was achieved when $\alpha = 0.5$, and for general, event topics, or specific topics, $\alpha = 0.4$ gave the best results.

Finally the opinion pattern adjustment is applied with different values of β, using Eq. 6.10 (or Eq. A.11). We found that for specific topics, we could tune a parameter for the best performance of relevant sentence retrieval. However, the performance in the final novelty detection was better without the opinion pattern adjustment. Therefore, the opinion pattern adjustment is not used for specific topic. For general opinion topics, we found that the best performance was achieved when $\beta=0.5$. We did not apply the opinion pattern adjustment to those general event patterns.

Note that the final parameters, (α = 0.4 or 0.5, β = 0.5 or 0.0), for best performance when applying all the adjustments together are different from the best parameters (α = 0.05, β >= 2) when we applying them individually.

A.2. List of the Candidate's Publications

1. Li, X. and Croft, W.B., "Novelty Detection Based on Sentence Level Patterns," *Proceedings of ACM Fourteenth Conference on Information and Knowledge Management (CIKM),* Bremen, Germany, 31st October - 5th November, 2005, pp 744-751, acceptance rate 17.8%

2. Li, X. and Croft, W.B., "Improving the Robustness of Relevance-Based Language Models," *CIIR Technical Report, IR-401*, Department of Computer Science, University of Massachusetts Amherst, 2005.

3. Abdul-Jaleel, N. Allan, J. Croft, W. B., Diaz, F. Larkey, L. Li, X, Smucker, M. D. and Wade, C. "UMass at TREC 2004: Novelty and HARD," *Proceedings of 2004 Text REtrieval Conference (TREC 2004).*

4. Li, X. and Croft, W. B., "An Answer Updating Approach to Novelty Detection," *CIIR Technical Report, IR-359*, Department of Computer Science, University of Massachusetts Amherst, 2004.

5. Li, X., "Syntactic Features in Question Answering," Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada July 28th – August 1st, 2003, pp.455-456.

6. Li, X. and Croft, W.B., "Time-Based Language Models," *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management*, New Orleans, Louisiana, November 2-8, 2003, pp. 469-475, acceptance rate 15%

7. Li, X. and Croft, W. B., "The Impact of Syntactic Evidence on the Effectiveness of Question Answering," *CIIR Technical Report, IR-246*, Department of Computer Science, University of Massachusetts Amherst, 2002.

8. Li, Xiaoyan and Croft, W.B., "Incorporating Syntactic Information in Question Answering," *CIIR Technical Report, IR-239*, Department of Computer Science, University of Massachusetts Amherst, 2001.

9. Li, X. and Croft, W.B., "Evaluating Question Answering Techniques in Chinese," *Proceedings of Human Language Technology Conference (HLT-2001)*, San Diego, March 18-21, 2001, pp. 201-206.

10. Allan, J., Connell, M., Croft, W.B., Feng, F., Fisher, D. and Li, X., "INQUERY and TREC-9," *Proceedings of 2000 Text Retrieval Conference (TREC 2000),* pp. 551-577.

BIBLIOGRAPHY

[1] J. Allan, R. Paka, and V. Lavrenko, "On-line New Event Detection and Tracking", Proc. SIGIR-98, 1998: 37-45.

[2] Y. Yang, J. Zhang, J. Carbonell and C. Jin, "Topic-conditioned Novelty Detection", SIGKDD, 2002: 688-693.

[3] N. Stokes and J. Carthy, "First Story Detection using a Composite Document Representation", *Proc. HLT01*, 2001.

[4] M. Franz, A. Ittycheriah, J. S. McCarley and T. Ward, "First Story Detection, Combining Similarity and Novelty Based Approach", *Topic Detection and Tracking Workshop*, 2001.

[5] J. Allan, V. Lavrenko and H. Jin, "First Story Detection in TDT is Hard", *Proc. CIKM*, 2000.

[6] D. Harman, "Overview of the TREC 2002 Novelty Track", *TREC 2002*.

[7] J. Allan, A. Bolivar and C. Wade, "Retrieval and Novelty Detection at the Sentence Level", *Proc. SIGIR-03,* 2003.

[8] H. Kazawa, T. Hirao, H. Isozaki and E. Maeda, "A machine learning approach for QA and Novelty Tracks: NTT system description", *TREC-10*, 2003.

[9] H. Qi, J. Otterbacher, A. Winkel and D. T. Radev, "The University of Michigan at TREC2002: Question Answering and Novelty Tracks", *TREC* 2002.

[10] D. Eichmann and P. Srinivasan. "Novel Results and Some Answers, The University of Iowa TREC-11 Results", *TREC* 2002.

[11] M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin and L. Zhao, "Expansiion-Based Technologies in Finding Relevant and New Information: THU TREC2002 Novelty Track Experiments", *TREC* 2002.

[12] K.L. Kwok, P. Deng, N. Dinstl and M. Chan, "TREC2002, Novelty and Filtering Track Experiments using PRICS", *TREC* 2002.

[13] Y. Zhang, J. Callan and T. Minka, "Novelty and Reduncancy Detection in Adaptive Filtering", *Proc. SIGIR,* 2002.

[14] E. M. Voorhees, "Overview of the TREC 2002 Question Answering Track", *TREC* 2002.

[15] S. E. Robertson, "The Probability Ranking Principle in IR", *Journal of Documentation*, 33(4):294-304, December 1977.

[16] Y. Yang, T. Pierce and J. Carbonell, "A Study on Retro-spective and On-Line event detection", *Proc. SIGIR-98*.

[17] C. Zhai, W. W. Cohen and J. Lafferty, "Beyond Independent Relevance: Method and Evaluation Metrics for Subtopic Retrieval", *Proc. SIGIR-03*, 2003: 10-17.

[18] T. Brants, F. Chen and A. Farahat, “A System for New Event Detection”, *Proc. SIGIR-03*, 2003: 330-337.

[19] X. Li and W. B. Croft, “Evaluating Question Answering Techniques in Chinese”, *Proc. HLT01, 2001*: 96-101.

[20] X. Li, “Syntactic Features in Question Answering”, *Proc. SIGIR-03,* 2003: 383-384.

[21] Daniel M. Bikel and Richard L. Schwartz and Ralph M. Weischedel, “An Algorithm that Learns What's in a Name”, *Machine Leaning*, vol 3, 1999. pp221-231.

[22] X. Li and W.B. Croft “Novelty Detection Based on Sentence Level Information Patterns”, *Proc. ACM CIKM’*05, 2005.

[23] I. Soboroff and D. Harman, “Overview of the TREC 2003 Novelty Track”, TREC 2003.

[24] J. Carbonell and J. Goldstein, “The Use of MMR, Diversity-Based Re-ranking for Reordering Documents and Producing Summaries”, Proc. SIGIR-98, 1998: 335-336.

[25] “Lemur Toolkit for Language Modeling and Information Retrieval”, a part of the LEMUR PROJECT by CMU and UMASS, http://www.lemurproject.org.

[26] C. Aone and M. Ramos-Santacruz, “REES: A Large-Scale Relation and Event Extraction System”, in the Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'00). May, 2000. Seattle, Washington. http://acl.ldc.upenn.edu/A/A00/A00-1011.pdf.

[27] R. Basili, M.T. Pazienza and M. Vindigni, “Corpus-driven learning of Event Recognition Rules”, in Proc. of Machine learning for Information Extraction workshop, held jointly with the ECAI2000, Berlin, Germany, 2000.

[28] S. Huttunen, R. Yangarber, and R. Grishman, “Complexity of Event Structure in IE Scenarios”, in Proceedings of the 19th International Conference on Computational Linguistics: COLING-2002 (August 2002) Taipei, Taiwan.

[29] R. Grishman, S. Huttunen and R. Yangarber, “Real-Time Event Extraction for Infectious Disease Outbreaks”, in Proceedings of the 3rd Annual Human Language Technology Conference: HLT-2002 (March 2002) San Diego, CA.

[30] R. Grishman, S. Huttunen and R. Yangarber, **“**Automatic Acquisition of Domain Knowledge for Information Extraction”, in Proceedings of the 18th International Conference on Computational Linguistics: COLING-2000 (August 2000) Saarbrücken, Germany.

[31] T. Brants, F. Chen, and A. Farahat, “A System for New Event Detection”, in Proceedings of ACM SIGIR2003.

[32] M. Tsai, M. Hsu and H. Chen, “Approach of Information Retrieval with Reference Corpus to Novelty Detection”, TREC 2003.

[33] Q. Jin, J. Zhao and B. Xu, “NLPR at TREC 2003: Novelty and Robust”, TREC 2003.

[34] J. Sun, J. Yang, W. Pan, H. Zhang, B. Wang and X. Cheng, “TREC-2003 Novelty and Web Track at ICT”, TREC 2003.

[35] K.C. Litkowski, “Use of Metadata for Question Answering and Novelty Tasks”, TREC 2003.

[36] M. Zhang, C. Lin, Y. Liu, L Zhao and S. Ma, “THUIR at REEC 2003: Novelty, Robust and Web”, TREC 2003.

[37] T. Dkaki and J. Mothe, “TREC Novelty Track at IRIT-SIG”, TREC 2003.

[38] D. Eichmann, P. Srinivasan, M. Light, H. Wang, X. Y. Qiu, R. J. Arens, and A. Sehgal, “Experiments in Novelty, Genes and Questions at the University of Iowa”, TREC 2003.

[39] G. Kumaran and J. Allan, “Text Classification and Named Entities for New Event Detection”, in Proceedings of ACM SIGIR 2004, pp297-304.

[40] W. Dai. and R. Srihari, “Minimal Document Set Retrieval,” *Proc. ACM CIKM’05*, pp 752-759.

[41] C. Zhai, “Notes on the Lemur TFIDF model”, online notes at www.cs.cmu.edu/~lemur/1.9/tfidf.ps

[42] I. Soboroff, “Overview of the TREC 2004 Novelty Track”, TREC 2004.

[43] R. Krovetz, 1993: "Viewing morphology as an inference process," in R. Korfhage et al., Proc. 16th ACM SIGIR Conference, Pittsburgh, June 27-July 1, 1993; pp. 191-202.

[44] G. Salton and M. J. McGill 1983 *Introduction to modern information retrieval*. McGraw-Hill, ISBN 0070544840.

[45] S.-N. Kim, D. Ravichandran, E. Hovy, “ISI novelty track system for TREC 2004,” TREC 2004.

[46] J. Ponte. *A Language Modeling Approach to Information Retrieval.* PhD thesis, Univ. of Massachusetts at Amherst, 1998.

[47] V. Lavrenko and W. Bruce Croft. Relevance-based Language Models. In 24th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’01) 2001.

[48] X. Li and W. B. Croft. “Time-based Language Models”, Proceedings of 12th International Conference on Information and Knowledge Management, 2003.

[49] V. Murdock and W. B. Croft, “A Translation Model for Sentence Retrieval,” in Proceedings of the Conference on Human Language Technologies and Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 684-691, 2005.

[50] C.J. van Rijsbergen, *Information retrieval*, Second edition, Butterworths, 1979.

[51] E.H. Hovy, U. Hermjakob, D. Ravichandran. “A Question/Answer Typology with Surface Text Patterns,” in Proceedings of the DARPA Human Language Technology conference (HLT). San Diego, CA, 2002

[52] M. Hu and B. Liu, “Mining and Summarizing Customer Reviews”, in Proceedings of KDD’04, the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp168-17, Seattle, 2004.

[53] S. Kim and E. Hovy, “Determining the Sentiment of Opinions”, in Proceedings of COLING-04, the Conference on Computational Linguistics, Geneva, CH, 2004.

[54] H. Turtle, and W.B. Croft, “Evaluation of an Inference Network-Based Retrieval Model”, ACM Transactions on Information System, in 9(3), 187-222, 1991.

[55] D. Metzler and W.B. Croft, “Combining the Language Model and Inference Network Approaches to Retrieval,” Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 40(5), 735-750, 2004

[56] T. Strohman, D. Metzler, H. Turtle, and W.B. Croft, “Indri: A Language-model based Search Engine for Complex Queries (extended version)”, http://ciir.cs.umass.edu/pubfiles/ir-407.pdf, CIIR technical report-407, 2004.

[57] J. Allan, R Gupta and V. Khandelwal, "Temporal Summaries of News Topics," in the Proceedings of the 24th Annual International ACM SIGIR Conference, pp. 10-18, 2001

[58] G. Kumaran, and J. Allan, , "Simple Questions to Improve Pseudo-Relevance Feedback Results," to appear in the Proceedings of the 29th Annual International ACM SIGIR Conference, 2006

[59] E. M. Voorhees, “Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness”, *Information Processing and Management*, 36(5): 697-716,2000