Improving Novelty Detection for General Topics Using Sentence Level Information Patterns

Xiaoyan Li
Center for Intelligent Information Retrieval
Department of Computer Science

University of Massachusetts, Amherst MA 01003

W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science

University of Massachusetts, Amherst MA 0100

ABSTRACT

The detection of new information in a document stream is an important component of many potential applications. In this work, a new novelty detection approach based on the identification of sentence level *information patterns* is proposed. First, the information- pattern concept for novelty detection is presented with the emphasis on new information patterns for *general topics* (queries) that cannot be simply turned into specific questions whose answers are specific named entities (NEs). Then we elaborate a thorough analysis of sentence level information patterns on data from the TREC novelty tracks, including sentence lengths, named entities, sentence level opinion patterns. This analysis provides guidelines in applying those patterns in novelty detection particularly for the general topics. Finally, a unified pattern-based approach is presented to novelty detection for both general and specific topics. The new method for dealing with general topics will be the focus. Experimental results show that the proposed approach significantly improves the performance of novelty detection for general topics as well as the overall performance for all topics from the 2002-2004 TREC novelty tracks.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation and retrieval models

General Terms:

Algorithms, experimentation

Keywords

Novelty detection, information patterns, named entities

1. INTRODUCTION

The goal of research on novelty detection is to provide a user with a list of materials that are relevant and contain new information with respect to a 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 a tedious and time-consuming task. A variety of novelty measures have been described in the literature [6, 7, 22]. The definitions of novelty, however, are quite vague and seem only indirectly related to the intuitive notions of novelty. Usually new words appearing in an incoming sentence or document contribute to the novelty scores in various novelty measures in different ways.

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. The idea of identifying query-related named-entities (NEs) patterns in sentences has been proved very effective in our previous study [25] in significantly improving the performance in novelty detection, particularly at top ranks. This approach is inspired by question answering techniques and is similar to passage retrieval for factoid questions. 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. Instead of extracting exact answers as in typical question answering systems [14,19,20], we have proposed 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 identify novel sentences that are more likely to have new answers to the questions. The effectiveness of the pattern-based approach has been validated by the experimental results on novelty detection on TREC 2003 and 2004 novelty tracks, with significant improvements in novelty detection for those specific topics corresponding to specific NE questions.

However, queries (topics) that can be transformed into specific NE questions are only a small portion of the query sets. For example, in TREC 2003, there are only 15 (out of 50) topics that can be formulated into specific NE questions. For the rest of the topics, which will be called general topics throughout the paper since they can only be formulated into general questions, the improvement is not very significant using the pattern-based approach merely based on general NE patterns. New and effective information patterns are needed in order to significantly improve the performance of novelty detection for those general topics, and this will be the focus of this paper. Meanwhile, a unified framework of the pattern-based approach is also required to deal with both the specific and the general topics.

As one of the main contributions of this work, we have found that the detection of information patterns related to opinions 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

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
Copyright 2006 ACM 1-59593-433-2/06/0011...$5.00.

more likely to be “opinion sentences”. Let us consider the following two sentences.

***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. 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 topically related to the query. However, for the example topic given above, it is very difficult for traditional word-based approaches to separate the non-relevant sentence (sentence 2) from the relevant sentence (sentence 1). This paper tries to attack this hard problem.

The rest of the paper is organized as follows. Section 2 gives a brief overview of related work on novelty detection. Section 3 introduces the concept of the proposed information patterns for novelty detection, with emphasis on information patterns for general topics that cannot be simply turned into NE questions. Section 4 elaborates a thorough analysis of sentence level information patterns, including sentence lengths, named entities, sentence level patterns related to opinions. The analysis is performed on the data from the TREC 2002 and 2003 novelty tracks, which provides guidelines in applying those patterns in novelty detection particularly for general topics. Section 5 describes the proposed unified pattern-based approach to novelty detection for both general and specific topics. The new method for dealing with general topics will be the focus. Section 6 shows experimental results in using those information patterns for significantly improving the performance novelty detection for topics corresponding to general questions, and for improving the overall performance of novelty detection using the unified approach. Section 7 summarizes the work.

2. RELATED WORK

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/first story detection [1,2,3,4,5,16,18]. Current techniques on new event detection are usually based on clustering algorithms. Some models (vector space models, language models, lexical chains, etc.) are used 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 start a new cluster. A story which started 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 and the similarity score between the story and its closest cluster is greater than the novelty score.

Research on novelty detection at the sentence level is related to the TREC novelty track for finding relevant and novel sentences given a query and an ordered list of relevant documents [7, 8, 9, 10, 11, 12, 13, 22]. In current techniques developed for novelty detection at the sentence level or document level, new words appearing in sentences/documents usually contribute to the scores that are used to rank sentences/documents. Many similarity functions used in information retrieval are also tried in novelty detection. Usually a high similarity score between a sentence and a given query will increase the relevance rank of the sentence while a high similarity score between the sentence and all previously seen sentences will decrease the novelty rank of the sentence, for example, the Maximal Marginal Relevance model (MMR) introduced by Carbonell and Goldstein [23].Novelty detection could be also performed at the document level, for example, in Zhang et al’s work [13] on novelty and redundancy detection in adaptive filtering, in Zhai et al’s work [17] on subtopic retrieval and in Dai et al’s work [26] on minimal document set retrieval.

There are two main differences between our proposed approach and the approaches in the literature. First, none of the work described above treats new information as *new answers* to questions that represented users’ information requests. Second, in the aforementioned systems related to the TREC novelty track, either the title query or all the three sections of a query were used merely as a bag of words, while we try to form *answer patterns* from the query. Our previous work [25] made a first attempt in this direction, but novelty detection performance only increases significantly for those specific topics that can be turned into specific NE questions.

3. DEFINITIONS OF NOVELTY AND INFORMATION PATTERNS

We emphasize that the definition of novelty or “new” information is crucial for the performance of a novelty detection system. Unfortunately, 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). 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.

In our previous work [25], a new definition of novelty has been given as following statement:

Novelty Definition: “*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* that 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. 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.

The novelty definition can be applied to novelty detection at different levels – event level, sentence level and document level. In this work, we will study novelty detection via information pattern identification at the *sentence level*. Novelty detection includes two consecutive steps: first retrieving relevant sentences and then detecting novel sentences. This novelty definition is also a general one that works for novelty detection with any query that can be turned into questions.

In our previous work [25], we have shown 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 whose answers are specific *named entities* (NEs), including persons, locations, dates, time, numbers, and etc.[21]. The *general questions (*corresponding to general topics), on the other hand, require obtaining additional information patterns for effective novelty detection. This will be the focus of this paper.

The identification and extraction of information patterns is crucial in our approach. The information patterns corresponding to NE-questions are called *NE words patterns*, related to the "when", "where", "who", "what" and how many" questions. 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). For a general topic, it is very difficult (if not impossible) to identify a particular type of named entity as its answer. Any type of named entity could be an answer as long as the answer context is related to the question. In quite some relevant and novel sentences, no named entities are included. Simply using named entities seems not very helpful for improving the performance of novelty detection for these general topics, as having been shown in [25]. Therefore the focus of this work will be 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 note that the topics in TREC 2003 and 2004 novelty tracks are either classified as event topics or opinion topics. As one of the particular interesting findings, we have found that a large portion of the general questions is about opinions. Opinions can typically be identified by looking at such sentence patterns as "XXX said", "YYY reported", or as marked by quotation marks. We have identified about 20 such opinion-related sentence patterns by manually scanning through a few paragraphs related to opinion patterns (Table 1). In the following section we will provide a through data analysis to support the above observations and arguments.

Table 1. Examples of opinion patterns

" ",
said, say, according to, add, addressed, agree, affirmed, reaffirmed, argue, believe, believes, claim, concern, consider, disagreed, 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

4. INFORMATION PATTERN ANALYSIS

In this section we will perform statistics of the three types of information patterns in relevant sentences, novel sentences and non-relevant sentences. The three information patterns are: *sentence lengths*, *named entities*, and *opinion patterns*. The goal is to find out effective ways to use these information patterns in distinguishing relevant sentences from non-relevant ones, and novel sentences from non-novel ones.

4.1 Statistics of Sentence Lengths

The statistics of sentence lengths in TREC 2002 and 2003 datasets are shown in Table 2. The length of a sentence is measured in the number of words after stop words are removed from the sentence. As a very useful result, the average lengths of relevant sentences from the 2002 data and the 2003 data are 15.58 and 13.1, respectively. But the average lengths of non-relevant sentences from the 2002 data and the 2003 data are only 9.55 and 8.5, respectively.

Table 2. Statistics of sentence length

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

We have the following interesting observation:

Observation #1*: Relevant sentences on average have significantly more words than non-relevant sentences.*

This feature is simple, but is very effective since the length differences between non-relevant and relevant sentences are significant. The feature is ignored in other approaches mainly because they are doing sentence retrieval with information retrieval techniques developed for document retrieval where document lengths are usually used as a penalty factor. Thus a short document is assigned a higher rank 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 relevance, which is crucial in detecting novel (and relevant) information. The difference between novel sentences and non-relevant sentences are slightly larger, which indicate that this incorporation of the sentence length information in relevance ranking will put the novel sentences with higher ranks in relevance retrieval.

4.2 The Statistics of Opinion Patterns

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 classification of opinion and event topics in the 2002 novelty track). We classify a sentence as an *opinion sentence* if it has one or more opinion patterns. Intuitively, opinion sentences are more likely to be relevant sentences than non-opinion sentences. Opinion patterns are detected in a sentence if it includes quotation marks or one or more of the expressions indicating it states an opinion (see Table 1 for a list). These patterns are extracted from TREC 2003 novelty track by scanning through a few documents in the data collection. 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” will most probably do. If a sentence includes one or more opinion patterns, it is said to be an *opinion sentence*.

We have run statistics of opinion patterns on the 2003 novelty track in order to obtain guidelines for using opinion patterns for both 2003 and 2004 data (note that there are no classification of opinion and event topics in the 2002 novelty track). 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 3, 48.1% of relevant sentences and 48.6% novel sentences are opinion sentences, but only 28.4% of non-relevant sentences are opinion sentences. We summarize this into the following observation:

Observation #2*: There are relatively more opinion sentences in relevant (and novel) sentences than in non-relevant sentences.*

This has a significant impact on separating relevant and novel sentences from 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 1. We have noticed that some related work [28] 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 how 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 will further enlarge the difference in numbers of opinion sentences between relevant sentences and non-relevant sentences.

Table 3. 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%)

4.3 Statistics of Named Entities

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. We also want to understand the role of certain named entities and their combinations in separating relevant sentences from non-relevant sentences, for event topics and opinion topics, respectively.

The statistics of all the 21 named entities that can be identified by our system are listed in Table 4. We have found that the most frequent types of NEs are PERSON, ORGANIZATION, LCATION, DATE and NUMBER. For each of them, there are more than 25% relevant sentences, each of which has at least one named entity of the type in consideration. Among these five types of NEs, three (PERSON, LOCATION and DATE) are more important than the other two (NUMBER and ORGANIZATION) for separating relevant sentences from non-relevant sentences. The discrimination capability of the ORGANIZATION type is not as significant and this has also been validated by experiments of novelty detection on real data. The role of the NUMBER type is not consistent among three TREC datasets. This is summarized in the following observation:

Observation #3*. Named entities of the three types - PERSON, LOCATION and DATE are more effective in separating relevant sentences from non-relevant sentences.*

Therefore only the three effective types will be incorporated into the sentence retrieval step to improve the performance of relevance. Named entities of the ORGANIZAION type is not used in relevant sentence detection since they almost equally appear in both relevant and non-relevant sentences. For example, the ratio is 42%:38% in the TREC 2003 novelty track. However, the ORGANIZATION type will be also used in the new pattern detection step since an NE of this type often indicates the name of a different news agency or some other organization, and a different one in a already relevant sentence may provide new information. This is summarized in the following observation:

Observation #4*: Named entities of the POLD types - PERSON, LOCATION ORGANIZATION and DATE will be used in new pattern detection; named entities of the ORGANIZATION type may provide different sources of new information.*

Table 4 also lists the statistics of those sentences with no NEs, with no POLD (Person, Organization, Location and Date) NEs and with no PLD (Person, Location and Date) NEs. These data show that (1) there are obvious larger differences between relevant and non-relevant sentences without PLD NEs which confirms that PLD NEs are more effective in re-ranking the relevance score (Eq. 4); and (2) there are considerable large percentages of relevant sentences without NEs or without POLD NEs. The second point is summarized into the following observation:

Observation #5: *The absence of NEs cannot be used exclusively to remove sentences from the relevant sentence list. The number of the previously unseen POLD NEs only contributes part in novelty ranking.*

As we have known, 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 those sentences as its answers. Analysis in Section 4.2 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. 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 improve the performance of novelty detection for those general, event topics? Table 5 compares the difference in the statistics of NEs between event topics and opinion topics. The observation is the following:

Table 4. The statistics of named entities (2002, 2003)

TREC 2002 Novelty Track Total = 57227, Total Rel#=1365, Total Non-Rel#=55862			**TREC 2003** Novelty Track Total S# = 39820, Total Rel#=15557, Total Non-Rel#=24263		
NEs	**Rel # (%)**	**Non-Rel # (%)**	**NEs**	**Rel # (%)**	**Non-Rel # (%)**
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%)
DATE	382(27.99%)	6860(12.28%)	DATE	3926(25.24%)	4236(17.46%)
NUMBER	444(32.53%)	14035(25.12%)	NUMBER	4141(26.62%)	6573(27.09%)
ENERGY	0(0.00%)	5(0.01%)	ENERGY	0(0.00%)	0(0.00%)
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%)
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%)
LENGTH	46(3.37%)	682(1.22%)	LENGTH	103(0.66%)	29(0.12%)
TIME	9(0.66%)	495(0.89%)	TIME	140(0.90%)	1154(4.76%)
ORDEREDNUMBER	77(5.64%)	1433(2.57%)	ORDEREDNUMBER	725(4.66%)	688(2.84%)
PERCENT	62(4.54%)	1271(2.28%)	PERCENT	371(2.38%)	1907(7.86%)
PERIOD	113(8.28%)	2518(4.51%)	PERIOD	1017(6.54%)	705(2.91%)
MONEY	66(4.84%)	1775(3.18%)	MONEY	451(2.90%)	1769(7.29%)
URL	0(0.00%)	0(0.00%)	URL	0(0.00%)	62(0.26%)
SPEED	0(0.00%)	0(0.00%)	SPEED	32(0.21%)	2(0.01%)
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%)

Table 5. 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%))

Observation #6: *PERSON, LOCATION and DATE play a more important role in event topics than in opinion topics.*

This is further verified in our experiments of relevance retrieval. In the equation for NE-adjustment (Eq. 4), the best results are achieved when α takes the value of 0.5 for event topics and 0.4 for opinion topics.

5. PATTERN-BASED APPROACH

In our definition, novelty means *new answers to the potential questions* representing a user's information need. Given this definition of novelty, it is possible to detect new information patterns by 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 introduce our unified pattern-based approach for both specific and general topics (queries). Then we will focus on the new method in improving the novelty detection performance for general topics.

5.1 A Unified Pattern-Based Approach

There are two important steps in the pattern-based novelty detection approach: *query analysis* and *new pattern detection.* At the first step, an information request from users will be (implicitly) transformed into one or more potential questions that determine corresponding query-related information patterns, which are represented by combinations of query words and required answer types to the query. At the second step, sentences with the query-related patterns are retrieved as answer sentences. Then sentences that indicate potential new answers to the questions are marked novel.

5.1.1 Query Analysis

A question formulation algorithm first tries to automatically formulate multiple specific questions for a query, if possible [25]. Each potential question is represented by a query-related pattern, which is a combination of a few query words and an expected answer type. 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*.

If this is not successful, a general question will be generated. *General questions* do not require a particular type of named entities for answers. Any types of named entities as listed in Table 4 could be answers as long as the answer context is related to the questions. Answers could be in sentences without any NEs. However, from our data analysis, the NEs of POLD types (*PERSON, ORGANIZATION, LOCATION, DATE*) are the most effective in detecting novel sentences (Observation #4), and three of them (*PERSON, LOCATION, DATE*) are the most significant in separating relevant sentences from non-relevant sentences (Observation #3). 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 (Observations #1 and #2). In particular, we can use opinion patterns to identify opinion sentences that are more probably relevant to opinion topics (queries) than non-opinion sentences (Observation #2). PERSON, LOCATION and DATE play a more important role in event topics than in opinion topics (Observation #6). Therefore, for a general question, its information pattern include four entities: topic type (event or opinion, used in Eq. 4 below to adjust the α), sentence length (in Eq. 3), POLD NE types (in Eqs.1 and 4), and opinion patterns (in Eq. 5 for opinion topics only).

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 formulated multiple specific questions for 8 queries from the TREC 2002 novelty track, 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.

5.1.2 New Pattern Detection

The new pattern detection step has two main modules: relevant sentence detection and then novel sentence detection. First, a search engine takes the query words of the query-related pattern generated from a potential question of a query and searches in its data collection to retrieve sentences that are likely to have correct answers. Our relevant sentence detection module filters out those sentences that do not satisfy the query-related patterns and/or re-rank the relevance list using the required information patterns. For a specific question (topic), 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 entities will be removed from the list. Then the relevance sentences list is re-ranked by incorporating the number of different types of required NEs to answer the questions derived from the specific topic in consideration [25].

For general questions (topics), all types of named entities (including no NEs) could be potential answers (Observation #5). Therefore the required information patterns are used in re-ranking the relevance list in order to improve the relevance performance for these general topics. This means that at retrieval step, the system will revise the relevance sentence retrieval results by adjusting relevant ranking scores using sentence lengths, NEs and opinion patterns. Details will be provided in the next sub-section.

Then, the new sentence detection module extracts all query-related named entities (as possible answers) from each answer sentence and detects previously unseen "answers". For specific topics, our system will identify sentences with possible new answers to the multiple NE questions as novel sentences (details can be found in [25]). For general topics, the novelty score is calculated with the following formula based on both Observations #4 and #5:

$$S_n = \omega N_w + \gamma N_{ne} \qquad (1)$$

where S_n is the novelty score of a sentence S, N_w is the number of new words in S that do not appear in its previous sentences, and N_{ne} is the number of POLD-type named entities in S that do not appear in its previous sentences. 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 for S_n is set to 4.

We want to make two notes here.

(1). Named entities considered at this step include all POLD types, i.e., PERSON, ORANIZATION, LOCATION and DATE. The ORGANIZATION type is also considered in this step since it often refers to the name of a news agency or some other organization, which could provide new information if it is a new one.

(2) By the summation of the counts in new words and new named entities, those relevant sentences that do not include any NEs could also be selected as novel sentences.

(3). The novelty score formula given in Eq. 1 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.

5.2 Using Patterns for General Topics

Sentence-level information patterns, including sentence lengths, Person-Location-Date NEs, and opinion sentences, are incorporated in the relevance retrieval step for general topics. This sub-section details this new method of incorporating these information patterns in the relevant sentence ranking.

***5.2.1** Ranking with a TFISF model*

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 [24]. There are many different formulas to calculate TFIDF score which is used for ranking documents.

We adopt the 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 [7, 24, 27], is calculated according to the following formula

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

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 $w(t_i) = 0.4$). With pseudo 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 a preprocessing, all the sentences have passed through the stopword removal and word stemming procedures.

This score S_0 will be served as the baseline for comparing the performance increase in relevant sentence retrieval for novelty detection.

5.2.2 TFISF with Information Patterns

The TFISF score is adjusted using the following three information patterns: sentence lengths, named entities, and opinion patterns. Following Observation #1, the length-adjustment is calculated as

$$S_1 = S_0 * (L / \overline{L}) \qquad (3)$$

where L denotes the length of a sentence and $\overline{L}$ denotes the average sentence length.

Following Observation #3, the NEs-adjustment is computed as

$$S_2 = S_1 * [1 + \alpha(F_{person} + F_{location} + F_{date})] \qquad (4)$$

where $F_{person} = 1$ if a sentence has a person name, 0 otherwise; $F_{location} = 1$ if a sentence has a location, 0 otherwise; and $F_{date} = 1$ if a sentence has a date, 0 otherwise. In addition, following Observation #6, the parameter α is set to 0.4 for opinion topics and to 0.5 for event topics.

Finally, following Observation #2, the opinion-adjustment is computed as

$$S_3 = S_2 * [1 + \beta F_{opinion}] \qquad (5)$$

where $F_{opinion} = 1$ if a sentence is an opinion sentence, 0 otherwise. The final adjustment step is only performed for opinion topics. A number of patterns (i.e. “said”, “argue that”, see Table 1) are used to determine whether a sentence is an opinion sentence.

We apply the three adjustments sequentially to tune the parameters on training data for best performance, and the same parameters are used for all data sets. We have also tried different ways of adjustments, and have found that current algorithm achieves the best performance.

Incorporating information patterns at the retrieval step improves the performance of relevance and thus helps later at the novelty detection 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 example sentences shown in Section 1, the relevant (and novel) sentence (sentence 1) was ranked 14th with the original 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 on TREC 2002, 2003 and 2004 are provided in the experiment section below.

6. EXPERIMENTAL RESULTS

In this section, 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 experiments and analysis include the performance of novelty detection for general topics using the proposed information patterns, and the overall performance of novelty detection using the unified pattern based approach.

6.1 Baseline Approaches

We compared our information-pattern-based novelty detection (IPND) approach to four baselines. The first baseline (B-NN) does ***n****ot* perform any ***n****ovelty* detection but only uses the initial sentence ranking. The second baseline (B-NW) in our comparison is simply applying ***new word*** detection. Starting from the initial retrieval ranking, it keeps sentences with new words that do not appear in previous sentences as novel sentences, and removes those sentences without new words from the list. All words in the collection were stemmed and stop-words were removed. 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. The fourth baseline B-MMR is a baseline with ***m****aximal* ***m****arginal* ***r****elevance* (MMR) [6,13,24,25]. MMR starts with the same initial sentences ranking used in other baselines and our approach. In 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. We use MMR as our fourth and main baseline because MMR was reported to work well in non-redundant text summarization [23], novelty detection at document filtering [13] and subtopic retrieval [17].

For comparison, in our experiments, the same retrieval system based on the TFISF technique adopted from the LEMUR toolkit [24] is used to obtain the retrieval results of relevant sentences in both the baselines and our approach. The evaluation measure used for performance comparison is precision at rank N. It shows the fraction of correct novel sentences in the top N sentences delivered to a user (N =5, 10, 15, 20 and 30 in Tables 7-12.). The precision values at top ranks are more meaningful in real applications where uses only want to go through a small number of sentences.

6.2 Experimental Results

Before we show the overall performance of our unified IPND approach, we will first see how information patterns can significantly improve the performance of novelty detection for those general topics that cannot be easily turned into specific NE questions that can be effectively handled by our previous NE-pattern-based approach [25].

We performed two sets of experiments. The first set of experiments is to evaluate the improvement in relevant sentence retrieval. The second set of experiments is to evaluate the overall performance in novelty detection. As described in Section 5, three types of information patterns are incorporated in the relevance retrieval step of novelty detection for general topics. Table 6 gives the performance of relevance retrieval with the original TFISF ranking and our approach with sentence level features and information patterns for the TREC 2002, 2003 and 2004 data, respectively. 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. This lays a solid ground for the second step - new information detection, and therefore for improving the performance of novelty detection for those general topics.

Table 6. Performance of relevance for general topics (Note: Data with * pass significant test – same applies to Tables 7-10; (1) $\alpha = 0.4$ (2) $\alpha = 0.5$ for event topics, $\alpha = 0.4$, $\beta = 0.5$ for opinion topics)

	TREC 2002 (41 Topics)			**TREC 2003 (35 Topics)**			**TREC 2004 (39 Topics)**		
Top # Sentences	TFIDF	Length + NEs (1)		TFIDF	Length + NEs + Opinion(2)		TFIDF	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. Performance comparison in novel detection for 41 queries (general topics) from TREC 2002

Top # Sentences	**B-NN**	**B-NW**		**B-NWT**		**B-MMR**		**IPND**	
	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 8. Performance comparison in novel detection for 35 queries (general topics) from TREC 2003

Top # Sentences	**B-NN**	**B-NW**		**B-NWT**		**B-MMR**		**IPND**	
	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 9. Performance comparison in novel detection for 39 queries (general topics) from TREC 2004

Top # Sentences	**B-NN**	**B-NW**		**B-NWT**		**B-MMR**		**IPND**	
	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*

Table 10. Performance comparison in novel detection for 49 queries (all topics) from TREC 2002

Top # Sentences	B-NN	B-NW		B-NWT		B-MMR		IPND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.1878	0.1959	4.3	0.2000	6.50	0.2204	17.4	0.2367	26.1
10	0.1939	0.1918	-1.1	0.2041	5.30	0.1980	2.1	0.2102	8.4
15	0.1891	0.1946	2.9	0.1986	5.00	0.1946	2.9	0.2095	10.8
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	0.1844	6.7

Table 11. Performance comparison in novel detection for 50 queries (all topics) from TREC 2003

Top # Sentences	B-NN	B-NW		B-NWT		B-MMR		IPND	
	Precision	Precision	Chg%	Precision	Chg%	Precision	Chg%	Precision	Chg%
5	0.4480	0.4680	4.5	0.4880	8.9	0.4600	2.7	0.5880*	31.2*
10	0.4520	0.4820*	6.6*	0.5200*	15.0*	0.4880	8.0	0.5860*	29.6*
15	0.4400	0.4920*	11.8*	0.5160*	17.3*	0.4907*	11.5*	0.5680*	29.1*
20	0.4400	0.4930*	12.0*	0.5280*	20.0*	0.4700*	6.8*	0.5450*	23.9*
30	0.4247	0.4773*	12.4*	0.5267*	24.0*	0.4747*	11.8*	0.5400*	27.2*

Table 12. Performance comparison in novel detection for 50 queries (all topics) from TREC 2004

Top # Sentences	*B-NN*	*B-NW*		*B-NWT*		*B-MMR*		*IPND*	
	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	0.2420	14.2
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	0.2340	17.6
30	0.1880	0.1973*	5.0*	0.2060*	9.6*	0.1913	1.8	0.2220	18.1

Tables 7-9 show the performance comparison of our IPND approach with the four baselines on those general topics that cannot be turned into specific NE questions. We can draw the following conclusions from the results.

(1) Our IPND approach consistently outperforms all the baseline approaches across the three data sets: the 2002, 2003 and 2004 novelty tracks. The precision values for the top 20 sentences with our IPND approach for general questions of the 2002, 2003 and 2004 data are 0.21, 0.50 and 0.21, respectively (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 is increased by 11.8%, 23.9% and 14.5%, respectively and the improvements are significantly larger than the other three baselines (2-4).

(2) B-NWT achieves better performance than B-NW as expected because B-NW is a special case of B-NWT when the new word threshold is set to 1.

(3) MMR is slightly better than B-NW and B-NWT on the 2002 data but is worse than B-NWT on the 2003 and 2004 data.

The overall performance comparison of the unified pattern-based approach with the four baselines on all topics from the TREC 2002, 2003 and 2004 novelty tracks is shown in Tables 10, 11 and 12, respectively. The most important conclusion is that the unified pattern-based approach outperforms all baselines at top ranks. Significant improvements are seen with the 2003 topics. In the top 10 sentences delivered, our approach retrieves 5.9 novel sentences on average, while the four baseline approaches only retrieve 4.5, 4.8, 5.2 and 4.9 novel sentences, respectively. As anticipated, the overall performance for all topics (including specific and general ones) is slightly better than that for the general topics, since the precisions for the specific ones are slightly higher.

7. CONCLUSIONS AND FUTURE WORK

In this paper, a unified pattern-based approach was proposed for novelty detection. Here we summarize the main features 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 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 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.

Experiments were carried out on the data from the TREC novelty tracks 2002-2004. The experimental results show that the proposed approach achieves significantly better performance at top ranks than the baseline approaches on topics from all three years. The proposed unified pattern-based approach results in significant improvement for novelty detection at the sentence level.

There are more research issues in the proposed pattern-based approach. These include issues in question formulation, relevant retrieval models, and new applications. Even if we have significantly improved the performance of novelty detection for those "general" topics by using the proposed sentence level information patterns, the novelty detection precisions for the specific topics are slightly higher. Therefore there are two-fold solutions to this problem. First, exploring the possibilities of turning more topics into multiple specific questions will be of great interests. Currently, for specific topics, only NE questions 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 uncovered part of the topic, but do not contain answers to the multiple NE questions. 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. Second, for general topics, the proposed three information patterns only capture part of the characteristics of the required answers. More information patterns could be helpful in further improving the performance of novelty detection for general topics.

In terms of relevant sentence retrieval models, currently, the pattern-based approach is combined with TFISF techniques, which are very simple, common and effective techniques on sentence retrieval. The pattern-based approach starts with the retrieval results from the TFISF techniques and adjust the believe scores of sentences according to sentence lengths and query-related patterns. We need to study how to combine information patterns with other retrieval approaches in addition to TFISF techniques, such as language modeling approaches, for further performance improvement. Other future work is to extend the pattern-based approach to novelty detection in other applications, such as new event detection, document filtering, cross document summarization and minimal document set retrieval etc.

8. ACKNOWLEDGMENTS

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 sponsor.

9. REFERENCES

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

[23] 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.

[24] "Lemur Toolkit for Language Modeling and Information Retrieval", a part of the LEMUR PROJECT by CMU and UMASS, http://www-2.cs.cmu.edu/~lemur/.

[25] X. Li &W.B. Croft" Novelty Detection Based on Sentence Level Information Patterns" *Proc. ACM CIKM'*05, 2005.

[26] W. Dai. and R. Srihari, "Minimal Document Set Retrieval," *Proc. ACM CIKM'05*, pp 752-759.

[27] C. Zhai, "Notes on the Lemur TFIDF model", online notes at www.cs.cmu.edu/~lemur/1.9/tfi

[28] S.-N. Kim, D. Ravichandran and E. Hovy, "ISI novelty track system for TREC 2004," TREC 2004.