Novelty Detection Based on Sentence Level 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 01003

ABSTRACT
The detection of new information in a document stream is an important component of many potential applications. In this paper, a new novelty detection approach based on the identification of sentence level patterns is proposed. Given a user's information need, some patterns in sentences such as combinations of query words, named entities and phrases, may contain more important and relevant information than single words. Therefore, the proposed novelty detection approach focuses on the identification of previously unseen query-related patterns in sentences. Specifically, a query is preprocessed and represented with patterns that include both query words and required answer types. These patterns are used to retrieve sentences, which are then determined to be novel if it is likely that a new answer is present. An analysis of patterns in sentences was performed with data from the TREC 2002 novelty track and experiments on novelty detection were carried out on data from the TREC 2003 and 2004 novelty tracks. The experimental results show that the proposed pattern-based approach significantly outperforms all three baselines in terms of precision at top ranks.

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]. These 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/story/document contribute to the novelty scores in various novelty measures though in different ways.

CIKM'05, October 31–November 5, 2005, Bremen, Germany.

We believe that patterns such as combinations of query words, named entities, phrases and etc, 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. For example, 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 relevant to the query. Let us consider the following four sentences given below.

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 bold) that match the key words from the query. However, only the first two sentences are relevant sentences. Besides the two matching words, the first two sentences also have a number 100 and 11 (underlined), respectively. Hence, the first two sentences are both topically and typically relevant to the query. The third sentence and the forth sentence are not relevant to the query mainly because they do not contain a number which is 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). Furthermore, 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. Therefore, a new approach that can identify query-related information patterns beyond single words is desired. This motivates our work in this paper.

The idea of identifying query-related patterns in sentences 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 propose to first extract interesting sentences with certain 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 rest of the paper is organized as follows. Section 2 gives a brief overview of related work on novelty detection. Section 3 introduces our understanding of novelty, and elaborates an analysis of sentence level patterns, focusing on named entities, with the data from the TREC 2002 novelty track data. Section 4 describes the proposed pattern-based approach to novelty detection. The experimental design and results are shown in Section 5. Section 6 summarizes the paper with conclusions and future work.

2. RELATED WORK

Novelty detection has been done at three different levels: event level, sentence level and document 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/first story detection [1,2,3,4,5,16,18]. Current techniques on new event detection are usually based on clustering algorithms. Some model (vector space model, language model, lexical chain, etc.) is used to represent each incoming news story/document. 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 queryand an ordered list of relevant documents [7, 8, 9, 10, 11, 12, 13, 22]. 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, and in Zhai et al's work [17] on subtopic retrieval. 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].

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, which we believe is essential in novelty detection. Second, in the aforementioned systems related to the TREC novelty track, either the title query or all the three sections of a querywere used merely as a bag of words, while we try to form *answer patterns* from the query.

3. NOVELTY UNDERSTANDING

We argue 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.

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 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. As shown in the example given in the introduction, the answer type in that query-related information pattern is NUMBER and the potential answer is 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.

3.1 Named Entity Pattern Analysis

Our novelty definition can be applied to novelty detection at different levels – event level, sentence level and document level. In this paper we will study novelty detection via information pattern identification at the *sentence level*. Throughout the paper, sentences that contain query-related patterns are called *relevant sentences*. Sentences that contain new patterns are called *novel sentences*. Novelty detection includes two consecutive steps: first retrieving relevant sentences and then detecting novel sentences.

Our novelty definition is also a general one that works for novelty detection with any query that can be turned into questions. In this paper we focus on one type of question whose answers are *named entities* (NEs), including persons, locations, dates, time, numbers, and etc.[21]. We call these questions *NE-questions*. The information patterns that will be discussed in this paper are NE patterns, which is a combination of both query words (of potential questions) and answer types (which requires named entities as potential answers). Since answers and new answers to NE-questions are named entities, 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 certain NE combinations (patterns) for separating relevant sentences from non-relevant sentences, and novel sentences from non-novel sentences. These NE patterns in consideration are the number of named entities and the number of different types of named entities in a sentence.

We analyzed two kinds of NE pattern distributions on the four classes of sentences: relevant, non-relevant, novel and non-novel. First 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. 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

$$P_r(N) = O_r(N)/M_r \quad (1)$$

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

$$P_{nr}(N) = O_{nr}(N)/M_{nr} \quad (2)$$

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

$$P_r(ND) = O_r(ND)/M_r \quad (3)$$

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

$$P_{nr}(ND) = O_{nr}(ND)/M_{nr} \quad (4)$$

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

$$P_n(N) = O_n(N)/M_n \quad (5)$$

$$P_{nn}(N) = O_{nn}(N)/M_{nn} \quad (6)$$

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

$$P_n(ND) = O_n(ND)/M_n \quad (7)$$

$$P_{nn}(ND) = O_{nn}(ND)/M_{nn} \quad (8)$$

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 and 48 queries are from the dataset collected by UMass. 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 the followings: *person, location, organization, money, date, time, number, percentage, temperature, ordered number, mass, height, length, period, energy, power, area, space, distance* and *object*. Most of the named entities are identified by BBN's IdentiFinder [21] and the rest by our own heuristic extractor.

In this subsection, 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 (one in Table 1 and one in Table 3).

Tables 1 and 2 show the results of the first set of statistical analyses. In Table 1, the second and third columns show the distributions of relevant sentences and non-relevant sentences with different types of named entities, indicated 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, indicated by the number in the first row (N). Table 2 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 results in Tables 1 and 2 indicate the following conclusions for NE patterns:

(1). Relevant sentences contain more named entities than the non-relevant sentences (as a percentage).

(2). The number of different types of named entities is more significant than the number of entities in discriminating relevant form non-relevant sentences, particularly when ND or N ≥ 2.

(3). The particular NE combinational patterns we select (in Table 2) have more impact on relevant sentence retrieval. For general combinations of two types of named entities (ND = 2 in Table 1), 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 2) is 2.41. The ratios for the combinations of three types of named entities (ND=3) are 1.85 in the general cases (Table 1) and 3.21 in the particular person-location-date combination (in Table 2).

Table 1. Named Entities(NE) distributions in relevant/non-relevant sentences (symbols are defined in Eqs. (1) – (4))

	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 2. 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%)

In the second set of analysis, we further study the distributions of named entities in *novel* and *non-novel* sentences. Tables 3 and 4 show the results. The design of the "novelty distribution" experimental analysis in Tables 3 and 4 is the same as the design in Tables 1 and 2, 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: (1) there are relatively more novel sentences (as a percentage)

than non-novel sentences that contain at least 2 different types of named entities (Table 3); and (2) there are relatively more novel sentences (in percentiles) than non-novel sentences that contain the four particular NE combinational patterns of interest (Table 4).

Table 3. Named Entities in novel and non-novel sentences (symbols are defined in Eqs. (5) – (8))

	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 4. 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%)

3.2 New Named Entity Pattern Analysis

The second step of our 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 shows that 67.2% of novel sentences do have new named entities while only 45.7% of redundant sentences have new named entities. There are two 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?

Table 5. Previously unseen NEs and Novelty/Redundancy

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

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 be transformed into general questions but not NE questions that ask for certain type of named entities/patterns as answers. For example, query 420 from TREC novelty track data 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.

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: "*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, while 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 indicates that only certain types of named entities may contain important information for a query.

4. AN 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 patterns at the sentence level. 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.

4.1 Query Analysis

The first step of the proposed pattern-based approach is to analysis the user's query and determine the possible query-related patterns that correspond to one or more potential specific questions or one general question, transformed from the query. A question formulation algorithm first tries to automatically formulate multiple specific questions for a queryif 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 paper, we deal with NE-questions that expect some type of named entities for answers. 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*. For a question like "*How many civilian non-combatants have been killed in the various civil wars in Africa*", the query analysis component formulates a query-related information pattern with both query words and an answer type. It first determines that the type of answer is NUMBER. Then it extracts *civilian, non, combatant, kill, various, civil, war, Africa* as query words of the question after question words (how many) and stopwords (have, been, in, the) are removed and stemming.

General questions do not require a particular type of named entities for answers. Any types of named entities could be answers as long as the answer context is related to the questions. The types of named entities 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*.

Named entities are identified with an algorithm based on BBN's IdentiFinder [21]. Each query from the TREC novelty tracks has three fields: title, description and narrative. Even though not explicitly provided in the format of a question, a significant number of the queries can be transformed into multiple specific questions. There are many approaches that can be used for question formulation and pattern determination. In our current implementation, we used a simple word-pattern matching algorithm to formulate questions and corresponding information patterns from queries. For each type of the five NE-questions, a number of word patterns were 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]. Some patterns are listed in Table 6. 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. 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.

Table 6. Word patterns for the five types of NE questions

Answer types	Word patterns
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
Number	how many, how much, length, number, polls, death tolls, injuries, how long,
Date	when, date, time, which year, which month, which day

There are 50 queries in the TREC 2003 novelty track and 50 queries in the TREC 2004 novelty track. The word-pattern matching algorithm formulated multiple specific questions 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 because of the lack of matched word patterns in their description and narrative fields.

4.2 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-relate 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. In another words, it first takes the results in finding relevant sentences with the TFIDF model implemented in LEMUR [24], and then removes the sentences that do not contain any "answers" to the potential question. For a specific question, only a specific type of named entities 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. This is the main difference between our pattern-based approach and other word-based approaches. For general questions, all types of named entities could be potential answers. Therefore only sentences without any named entities are removed from the list. In both cases, a list of presumed *answer sentences* (which contain expected named entities to the question) is generated. To improve the performance of finding relevant sentences and increase the rank for sentences with more named entities, the sentence retrieval module will further re-rank the sentences by a revised score S_r, which is calculated according to one of the following equations:

$$S_r = S_o + \alpha * ND \quad (9)$$

$$S_r = S_o + \beta * N \quad (10)$$

where S_o is the original score from the retrieval system we use, ND is the number of different types of named entities a sentence contains, N is the number of named entities and α is a parameter. We tried various values of α and β. It turned out that Equation (9) with measurements of different *types* of named entities is more effective than Equation (10) with merely measurements of number of named entities in finding relevant sentences and identifying novel sentences. This observation is consistent to our findings in named entity pattern distribution analysis shown in Section 3. Therefore, we use Equation (9) in the sentence retrieval module for the experiments reported in this paper. This is the second main difference between our pattern-based approach and the previous word-based approaches.

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 a query that is transformed into a general question, all named entities identified in an answer sentence will be extracted as potential answers. For a query with multiple specific questions formulated, an answer sentence may have answers to one or more specific questions formulated from the query. So named entities related to any one of the specific questions in the answer sentences should be extracted. There is an answer pool associated with each question, which is initially empty. As sentences come in, new answers 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.

5. EXPERIMENTS AND RESULTS

In this section, we present and discuss the main experimental results. The data used in our experiments and the comparison of our approach and several baseline approaches are also described.

5.1 Data

Currently, there are three sets of data officially available for novelty detection at the sentence level. The TREC 2002 novelty track generated 54 queries. Each of the TREC 2003 novelty track and 2004 novelty tracks collected 50 queries. 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, there are zero or more non-relevant documents in addition to 25 relevant documents as well. A set of sentences was marked relevant, and further a subset of those sentences was marked novel. The main difference between the three sets is that the TREC 2003 and 2004 novelty track collections exhibited greater redundancy and thus has less novel sentences [22]. 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, while 90.9% of the total relevant sentences in the 2002 track are novel sentences. Our pattern-based approach was trained with the data from the TREC 2002 novelty track and tested on the data from the TREC 2003 and 2004 novelty tracks.

5.2 Baselines

We compared our pattern-based novelty detection (PBND) approach to three baselines: B-NN: baseline with initial retrieval ranking (without novelty detection), B-NW: baseline with new word detection, and B-MMR: baseline with maximal marginal relevance (MMR). For comparison, in our experiments, the same retrieval system based on the TFIDF technique implemented in 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 novel sentences in the top N sentences (N =5, 10, 15 … in Tables 7-11.) delivered to a user. The precision at top ranks is more meaningful in real applications where uses only want to go through a small number of sentences.

5.2.1 B-NN: Initial Retrieval Ranking

The first baseline does not perform any novelty detection but only uses the initial sentence ranking scores by the retrieval system directly as the novelty scores. One purpose of using this baseline is to see how much novelty detection processes may help in removing redundancies. Another purpose is to see how many novel sentences in the initial retrieval ranking list that our approach does not detect. Because of the "hard" decision (relevant or non-relevant, novel or non-novel) in the new pattern detection process, our novelty detection approach may produce a shorter list of sentences.

5.2.2 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 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 stopwords were removed.

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

5.2.3 B-MMR: Maximal Marginal Relevance (MMR)

Many approaches to novelty detection, such as maximal marginal relevance (MMR), simple new word count measure, set difference measure, cosine distance measure, language model measures, etc. [6-13,24], were reported in the literature. MMR was introduced by Carbonell and Goldstein [23] in 1998, which was used for reducing redundancy while maintaining query relevance in document reranking and text summarization. MMR starts with the same initial sentences ranking used in other baselines and our approaches. 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. The process stops until all sentences in the initial ranking list are selected. MMR is calculated by Eq. (11)

$$MMR = \arg \max_{S_i \in R/N} \left[\lambda (Sim_1(S_i, Q) - (1-\lambda) \max_{S_j \in N} Sim_2(S_i, S_j) \right] \quad (11)$$

where S_i , and S_j are the ith and jth sentences in the initial sentence ranking. 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 a different similarity metric between sentences.

We use MMR as our third 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]. Also, MMR may incorporate various novelty measures by using different similarity matrix between sentences and 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].

5.3 Results and Discussions

We tested the pattern-based novelty detection (PBND) approach on the data from the TREC 2003 and 2004 novelty tracks and compared it to the aforementioned three baselines. Three sets of experimental results are shown here, which are (1) performance of identifying novel sentences for queries that were transformed into multiple specific questions (with query words and specific NE answer types); (2) performance of identifying novel sentences for queries that were transformed into general questions (with any NEs as answers); and (3) performance of finding relevant sentences for all queries. From Table 7 to Table 11, Chg% denotes the percent change compared to the first baseline and starts indicate statistically significant difference at a 95% confidence level by the Wilcoxon test.

The purpose of the first set of results, shown in Tables 7 and 8, is to compare the performance of our pattern-based approach to the three baselines for queries with specific question formulation. Our query analysis algorithm formulated multiple specific questions 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. We have the following observations and interpretations on the experimental results.

(1). The proposed approach outperforms all baselines at top ranks. The performance of our approach with specific questions beats the first baseline by more than 20% at rank 30 on both the data from both the TREC 2003 novelty track and the 2004 novelty track. Within the top 30 sentences, our approach obtains more novel sentences than the baselines. For many users who only want to go through a small number of sentences for answers, novel sentences in the top 10, 20 or 30 ranks are more meaningful in real applications. Note that MMR performs slightly better than both the new word detection baseline and the first baseline which solely uses the results from IR at low recall.

(2) The precision of our approach at rank 1000 is significantly lower than the three baselines. Although retrieving this number of sentences would be impractical for novelty-based applications, this result does indicate that are very precision-oriented. For example, in the top 1000 sentences (the last row of Table 7), the first baseline indicates that there are 218 novel sentences on average for each

query; however our approach only detected 111 sentences. The first three rows in Table 7 show a summary of the 15 queries reported. Of the 3,990 novel sentences in total for the 15 queries with specific question formulation, our approach detected 1,079 correct novel sentences, whereas the number is 3268 for the first baseline B-NN.

Table 7. Performance comparison in identifying novel sentences for 15 queries from TREC 2003 with specific question formulation (#TNS: # of Total Novel Sentences; #NSR: Total # of Novel Sentences Retrieved; #ASR: Average # of Sentences Retrieved per Query)

	B-NN	B- NW		B-MMR		PBND-Specific	
# TNS	3990	3990		3990		3990	
# NSR	3268	2862		3268		1079	
# ASR	499	356	Chg%	499	Chg%	219	Chg%
Prec.at 5 S.	0.507	0.520	+2.6	0.520	+2.6	0.693	+36.8*
10	0.540	0.567	+5.0	0.560	+3.7	0.720	+33.3*
15	0.498	0.560	+12.4	0.600	+20.5*	0.680	+36.6*
20	0.520	0.567	+9.0	0.587	+12.9*	0.643	+23.7*
30	0.513	0.564	+9.9	0.598	+16.6*	0.631	+22.9*
50	0.496	0.580	+16.9*	0.576	+16.1*	0.636	+28.2*
100	0.499	0.578	+15.8*	0.517	+3.6	0.606	+21.4*
1000	0.218	0.191	-12.4*	0.218	0	0.111	-49.1*

Table 8. Performance comparison in identifying novel sentences for 11 queries from TREC 2004 w/ specific questions

	B-NN	B- NW		B-MMR		PBND-Specific	
# TNS	866	866		866		866	
# NSR	801	677		801		349	
# ASR	596	430	Chg%	596	Chg%	196	Chg%
Prec. at 5 S.	0.200	0.200	0	0.218	+9.0	0.273	+36.5*
10	0.245	0.245	0	0.245	0	0.300	+22.4
15	0.230	0.230	0	0.236	+2.5	0.303	+31.7*
20	0.246	0.250	+1.6	0.250	+1.6	0.309	+25.6
30	0.239	0.245	+2.5	0.273	+14.2	0.303	+26.8*
50	0.240	0.260	+8.3*	0.236	-1.7	0.296	+23.3*
100	0.229	0.235	+2.6	0.175	-22.3*	0.219	-4.4
1000	0.073	0.062	-15.1*	0.073	0	0.032	-56.2*

The second set of experimental results compares the performance of our PBND approach to the three baselines for remaining queries that were transformed into general questions. The results of this set of experiments are given in Tables 9 and 10. The results show that the performance of our approach on these queries are slightly better than the baselines but the performance difference for these queries with general question formulations was not as significant as that for those queries with specific question formulations reported in Tables 7 and 8. This indicates that simply identifying new named entities in sentences does not produce a significant performance gain for novelty detection for general queries. Other types of questions that do not require named entities for answers also need to be considered in order to get better performance, especially for the queries that could not be transformed into multiple specific questions in the current implementation of the pattern-based approach (reported in Tables 9 and 10). This is a major focus of our future work.

The third set of experiments is designed to investigate the performance gain of finding relevant sentences with the sentence *re-ranking* step in our approaches. Remember that, in our approach, the relevant sentence retrieval module re-ranks the sentences by the revised scores that incorporate the number of different types of named entities appeared in a sentence. Our hypothesis is that this re-ranking process would improve the performance of finding relevant sentences. We compare the performance of finding relevant sentences with and without re-ranking. The comparison results are shown in Table 11, which verify our hypothesis at low recall, even if the difference is not significant. But the results in Tables 7 and 8 have shown that the pattern-based approach significantly outperforms all three baselines at low recall for identifying novel sentences. 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.

Table 9. Performance comparison in identifying novel sentences for 35 queries from TREC 2003 w/ general questions

	B-NN	B- NW		B-MMR		PBND	
# TNS	6236	6236		6236		6236	
# NSR	5117	4523		5117		2451	
# ASR	520	397	Chg%	520	Chg%	222	Chg%
Prec. at 5 S.	0.423	0.440	+4.0	0.434	+2.6	0.445	+5.4
10	0.414	0.434	+4.8	0.457	+10.4	0.449	+8.3
15	0.415	0.432	+4.1	0.443	+6.7	0.465	+11.9*
20	0.406	0.426	+4.9*	0.420	+3.4	0.443	+9.2*
30	0.387	0.410	+4.7*	0.422	+9.0	0.441	+14.0*
50	0.366	0.405	+10.7	0.398	+8.7*	0.427	+14.3*
100	0.376	0.413	+9.8*	0.352	-6.4*	0.399	+6.2*
1000	0.146	0.129	-11.6*	0.146	0	0.070	-52.1*

Table 10. Performance comparison in identifying novel sentences for 39 queries from TREC 2004 w/ general questions

	B-NN	B- NW		B-MMR		PBND	
# TNS	2588	2588		2588		2588	
# NSR	2312	1889		2312		1183	
# ASR	711	499	Chg%	711	Chg%	447	Chg%
Prec. at 5 S.	0.236	0.236	0	0.236	0	0.241	+2.1
10	0.203	0.203	0	0.203	0	0.215	+5.9
15	0.195	0.198	+1.5	0.195	0	0.198	+1.5
20	0.186	0.186	0	0.185	-0.5	0.183	-0.5
30	0.174	0.179	+2.9	0.168	-3.4	0.174	0
50	0.162	0.167	+3.1	0.146	-9.9	0.167	+3.1
100	0.143	0.151	+5.6*	0.131	-8.4*	0.146	+2.1
1000	0.059	0.048	-18.6*	0.059	0	0.030	-49.2*

Table 11. Performance comparison in finding relevant sentences for 50 queries in TREC 2003 & TREC 2004 each

Dataset	**50 Queries in TREC 2003**			**50 Queries in TREC 2004**		
Mtds	TFIDF	PBND		TFIDF	PBND	
# TNS	15557	15557		8343	8343	
# NSR	12793	10563		7610	6730	
# ASR	513	421	Chg%	685	580	Chg%
Prec. at 5S.	0.728	0.732	+0.5	0.508	0.512	+0.8
10	0.696	0.710	+2.0	0.484	0.488	+0.8
15	0.695	0.712	+2.3	0.469	0.475	+1.1
20	0.705	0.708	+0.6	0.458	0.467	+2.0
30	0.699	0.702	+0.3	0.447	0.445	-0.3
50	0.679	0.689	+1.5	0.427	0.432	+1.2
100	0.655	0.659	+0.7	0.396	0.390	-1.5
1000	0.256	0.211	-17.4*	0.152	0.135	-12.7*

6. CONCLUSIONS AND FUTURE WORK

The motivation of this work is to explore new methods for novelty detection, an important task to reduce the amount of redundant as well as non-relevant material presented to a user. In this paper, we introduce a new definition of novelty (or new information) as *new answers to the potential questions* representing a user's request or information need. Based on this definition, we have proposed a pattern-based approach to identify novel sentences, i.e. sentences with certain patterns that indicate the presence of potential new answers to the questions related to a query. The proposed pattern-based approach was trained with the data from the TREC 2002 novelty track and tested on 100 queries from the TREC 2003 and 2004 novelty tracks. The experimental results show that the pattern-based approach significantly outperforms all three baselines in terms of precision at low recall, but only for queries where specific answer-related patterns can be formulated. For general queries, there is small but not significant improvement.

We have also investigated the distributions of named entities and patterns in relevant/novel and non-relevant/non-novel sentences. The important observation is that there are relatively more novel/relevant sentences than non-novel/non-relevant sentences that contain *multiple types* of named entities. This observation has been partially incorporated in the pattern-based approach.

An important step in the proposed pattern-based approach is to determine information patterns that correspond to multiple specific questions (implicitly) transformed from a query. Currently, only NE-questions and NE-patterns are considered. In future work, we will improve the pattern-based approach to explore general patterns for the improvement of performance of general questions. Other future work will extend the pattern-based approach to novelty detection in other applications, such as new event detection and document filtering, etc.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsor.

8. 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 Retrospective 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: 383-38

[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 Reranking 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/