An Answer Updating Approach to Novelty Detection

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 and novel information in a document stream is an important component of potential applications. This paper describes an *answer updating approach* to novelty detection at the sentence level. Specifically, we explore the use of question-answering techniques for novelty detection. New information is defined as new/previously unseen answers to questions representing a user's information need. A sentence is treated as novel sentence if the system believes that it may contain a previously unseen answer to the question. In our answer updating approach, there are two important steps: question formulation and new answer detection. Experiments were carried out on data from the TREC 2003 novelty track using the proposed approach. The results show that the proposed answer updating approach outperforms all three baselines in terms of precision at low recall.

Keywords

Novelty detection, question answering, 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, 23]. 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.

We give a definition of novelty as *new answers to the potential questions* representing a user's request or information need. If a new answer to the question, which represents the user's information need or part of it, appears in a sentence or story or document, then we say the sentence (story or document) has new information that the user wants. Given this definition of novelty, it is possible to detect new information by monitoring how the answer to a question changes. Therefore, we propose to perform novelty detection via *answer updating*. This approach is made even more feasible by the progress in ongoing research on question answering techniques [14].

The rest of the paper is organized as follows. Section 2 gives a short overview of related work on novelty detection. Section 3 introduces our new definition of novelty, and elaborates a new perspective of novelty understanding with an analysis of the TREC novelty track data. Section 4 describes the proposed answer updating approach for novelty detection and explains how novelty detection can be done via answer updating. Experimental design and results are shown in Section 5. Section 6 gives a brief discussion on challenges in data collections for testing various novelty detection approaches. Section 7 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 topic and an ordered list of relevant documents [7, 8, 9, 10, 11, 12, 13, 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, 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 (IR) 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

Conference'04, Month 1–2, 2004, City, State, Country.

Relevance model (MMR) introduced by Carbonell and Goldstein [24].

There are two main differences between our proposed approach and the approaches in the literature. First, none of the work described above treated 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 topic were used merely as a bag of words, while we try to *form questions* and/or to *understand the question(s)* from the sections of a topic.

3. NOVELTY UNDERSTANDING

3.1 What is Novelty?

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 answers using a question-answering system. Second, new information is obtained by detecting *new* answers from the question-answering system. Therefore, understanding novelty from the perspective of a question answering paradigm is important before we go into the methods in our answer updating approach. 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 well captured by one or more questions. Let us first explore the relationship between queries in IR (information retrieval, which most of the current novelty detection approaches are based) and questions in QA (question answering, which distinguishes our approach from others), using a few examples. This will help us understand why novelty detection via question answering is more appropriate.

Topic 306 from the TREC 2002 novelty track is a good example:

> *<title> African Civilian Deaths*
>
> *<desc> Description: How many civilian non-combatants have been killed in the various civil wars in Africa?*
>
> *<narr> Narrative: A relevant document will contain specific casualty information for a given area, country, or region. It will cite numbers of civilian deaths caused directly or indirectly by armed conflict.*

An IR system will take the title query "*African Civilian Deaths*" to retrieve relevant documents because the title/short query has more focused words and may produce better performance than long/description/narrative query does. However, the description "*How many civilian non-combatants have been killed in the various civil wars in Africa*" expresses the user's request more clearly.

Another example is topic 301 from TREC 2002:

> *<title> International Organized Crime*
>
> *<desc> Description:Identify organizations that participate in international criminal activity, the activity, and, if possible, collaborating organizations and the countries involved.*
>
> *<narr> Narrative: A relevant document must as a minimum identify the organization and the type of illegal activity (e.g., Columbian cartel exporting cocaine). Vague references to international drug trade without identification of the organization(s) involved would not be relevant.*

Although the description of topic 301 is not in the format of a question, it can be reformatted as a question "*What are the organizations that participate in international criminal activity?*" This question is a better representation of the topic than the title query consisting of the key words "*international organized crime*". As Robertson put it [15], "the object of a reference retrieval system is to predict, in response to a request, which documents the requester will find relevant to his request or useful to him in his attempt to find the answer". This implicitly suggests that a user's request can often be captured by one or more questions.

3.2 Named Entity Distribution Analysis

Our novelty definition is 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, organization locations, dates, time, numbers, and etc.[21]. We call these questions *NE-questions*. The reason for this choice is that state-of-the-art QA systems are relatively successful in dealing with NE-questions [8,9,10,14,19,20].

The novelty definition can also be applied to novelty detection at different levels – event level, sentence level and document level. In this paper we will study novelty detection via answer updating at the *sentence level*. In our novelty definition, novelty is indicated by new answers to the potential questions. Throughout the paper, sentences that contain answers to questions are called *relevant sentences*. Sentences that contain new answers are called *novel sentences*. Novelty detection includes two consecutive steps: first retrieving relevant sentences and then detecting novel sentences. Since answers and new answers to NE-questions are named entities, understanding the distribution of named entities could be very helpful both in finding relevant sentences and in detecting novel sentences. We also want to understand important factors for separating relevant sentences from non-relevant sentences, and novel sentences from non-novel sentences. These factors include the number of named entities and the number of different types of named entities in a sentence.

To learn more about this, we analyzed two kinds of 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 relevant sentences in the dataset is M_n, and the total number of non-relevant 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 two subsections, we will show and explain the results from our novelty data investigation. We used 53 topics from TREC 2002 novelty track, 48 topics from a UMass dataset [7], and 50 topics from the TREC 2003 novelty track. For each query there is a set of sentences that have been pre-marked as relevant/non-relevant, and novel/non-novel. In our experiments, 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*. Most of the named entities are identified by BBN's IdentiFinder [21], and the rest is identified by our own code. In this paper we will mainly report the results with TREC 2003 novelty track data. Similar results of the other two datasets, TREC 2002 novelty track, consisting of 54 topics, and 48 topics collected by UMass, have also been obtained, which can be found at [22]. The total number of sentences for all 50 topics in TREC 2003 is 39,870, in which the total number of relevant sentences M_r is 15,557, and the total number of non-relevant sentences M_{nr} is 24,313. The total number of novel sentences M_n is 10,226, and the number of non-novel sentences M_{nn} is 29,644.

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. In the next subsection, we are going to further study the distributions of *new* entities, which may indicate new information. 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 that are marked by asterisks (in Tables 1 and 3).

Tables 1 and Table 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 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:

(1). Relevant sentences contain more named entities than the non-relevant sentences (in 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 is greater or equal to 3. Note that the two sets of data that do not pass the t-test are in the distributions of named entity numbers (Columns 4 and 5 in Table 1 and then in Table 3).

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	2876 (18.5%)	5148 (21.2%)	2875 (18.5%)	5145 (21.2%)
1	3919 (25.2%)	7023 (28.9%)	3041 (19.5%)*	5346 (22.0%)*
2	3758 (24.2%)	6800 (28.0%)	2912 (18.7%)	5026 (20.7%)
3	2819 (18.1%)	3706 (15.2%)	2279 (14.6%)*	3436 (14.1%)*
4	1542 (9.9%)	1237 (5.1%)	1671 (10.7%)	2454 (10.1%)
5	511 (3.3%)	347 (1.4%)	1000 (6.4%)	1414 (5.8%)
>5	132 (0.8%)	52 (0.2%)	1779 (11.4%)	1492 (6.1%)

Table 2. NE combinations in relevant / non-relevant sentences

NE Combination	# of Relevant Sentences (%)	# of Non-Relevant Sentences (%)
PersonLocation	2496 (16.0%)	2286 (9.4%)
PersonDate	1911 (12.3%)	1493 (6.1%)
LocationDate	1935 (12.4%)	1235 (5.1%)
PersonLocationDate	987 (6.3%)	702 (2.9%)

(3). The particular combinations 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 24.2%/28.0% =0.86 (which does not provide any useful information). However the average ratio for three types of combinations of two different named entities (in Table 3) is 1.98 (indicating significant discriminations). The ratios for the combinations of three types of named entities (ND=3) are 1.19 in the general cases (Table 1) and 2.17 in the particular person-location-date combination (in Table 2).

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 3 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 combinations 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	1895 (18.5%)	6129 (20.7%)	1894 (18.5%)	6126 (20.7%)
1	2609 (25.5%)	8333 (28.1%)	2016 (19.7%)*	6371 (21.5%)*
2	2477 (24.2%)	8081 (27.3%)	1912 (18.7%)	6026 (20.3%)
3	1835 (17.9%)	4690 (15.8%)	1493 (14.6%)*	4222 (14.2%)*
4	1011 (9.9%)	1768 (6.0%)	1116 (10.9%)	3009 (10.2%)
5	322 (3.1%)	536 (1.8%)	648 (6.3%)	1766 (6.0%)
>5	77 (0.8%)	107 (0.4%)	1147 (11.2%)	2124 (7.2%)

Table 4. NE combinations in novel and non-novel sentences

NE Combination	**# of Novel Sentences (%)**	**# of Non-Novel Sentences (%)**
PersonLocation	1719 (16.8%)	3063 (10.3%)
PersonDate	1245 (12.2%)	2159 (7.3%)
LocationDate	1301 (12.7%)	1869 (6.3%)
PersonLocationDate	663 (6.5%)	1026 (3.5%)

3.3 New Named Entity Analysis

The next 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 answers/named entities to given questions, and novel sentences should contain new answers or previously unseen named entities. Thus a relevant sentence with no new answer/named entities is said to be redundant.

Table 5 shows that 52.7% of novel sentences do have new named entities while only 28.4% of redundant sentences have new named entities. There are two interesting questions based on these results of these statistics. First, there are 47.3% 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 28.4% 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 Topics**
Novel Sentences	10226	5389 (52.7%)	50
Redundant Sentences	5331	1514 (28.4%)	50

To answer these two questions, we did a further investigation on the novel/redundant sentences and its corresponding topics. We have found that most of the novel sentences *without* new named entities are related to some particular topics. These queries can be transformed into general questions but not NE questions that ask for certain type of named entities as answers. For example, topic N1 from TREC 2003 novelty track data is concerned about opinions about the proposed ban on partial birth abortion. A relevant sentence to this topic doesn't have to have any named entities to be relevant, let alone new named entities. In fact, about 30% of the relevant sentences to this topic don't contain any named entities at all. More than twenty percent of relevant sentences for 13 topics (out of the 50) in TREC 2003 do not have any named entities.

For the second question, all types of new named entities that could be identified by our system 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, topic N4 is about "Egyptian Air Flight 990 disaster in October of 1999" For this topic, a name, number (of passengers), date, time or location appearing in a relevant sentence could be an answer, while other named entities may not be of interest. Therefore, a relevant sentence with a previously unseen company name could be redundant.

This investigation of named entities can be used as the basis for improving the performance of finding relevant sentences and detecting novel sentences. Based on our definition of novelty and the results of novelty data investigation, we proposed an answer updating approach to novelty detection, which is detailed in Section 4.

4. AN ANSWER UPDATING APPROACH

Given the definition of novelty as new answers to potential questions that represent a user's request or information need, we propose to perform novelty detection via answer updating. There are two important steps in the proposed approach: *question formulation*, to transform each topic into one or multiple questions, and *new answer detection*, to find relevant sentences that contain the answers to a question and mark a relevant sentence as novel if it contains a new answer. The framework of our approach is shown in Figure 1.

4.1 Question Formulation

The first step of our approach is to transform each topic into one or multiple questions, automatically. We have tried three different methods for question formulation: specific question formulation, POLD question formulation and general question formulation. For typical QA systems as in [19,20], a query is generated with key words from the question. Therefore questions are represented with key words and the expected answer types. (i.e. number, person, organization, location, date and time)

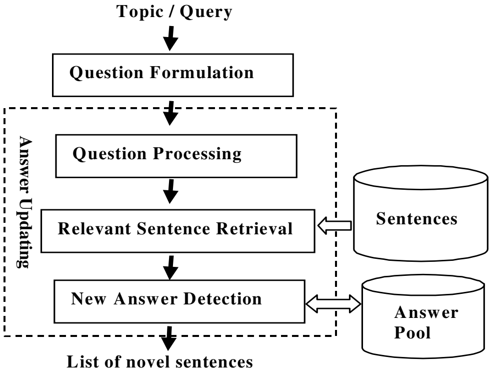

Figure 1: The proposed novelty detection system

Specific question formulation. *Specific questions* ask for specific types of named entities as answers, i.e., a topic can be transformed into multiple NE-question(s). We note that each topic from TREC 2002 and 2003 novelty track has three fields: title, description and narrative. For some of the topics, we can automatically formulate specific question(s) using the key words in the title, description and narrative fields of the topics, if the topics can be transformed into NE-questions that we have defined word patterns. For some topics, the right questions are readily available in the topic description. For example, for topic 306 of TREC 2002 we showed in Section 3, the right question is "*How many civilian non-combatants have been killed in the various civil wars in Africa?*" The question is exactly the text in the description field for this topic. For some other topics, questions are not directly available, but we have developed a simple word-pattern matching algorithm for identifying the following types of NE questions: person, organization, location, number and date (time). We found that in 15 topics (out of 50) we can formulate specific questions using the algorithm.

POLD question formulation. Novelty topics from the TREC novelty track are much broader than simple factoid questions to Question Answering Systems. A topic could possible be transformed into multiple specific NE questions. To deal with those topics for which our word-pattern matching algorithm cannot find multiple specific NE questions, we simply assume that each of them may include the following three typical questions: Who, Where and When. The topics in the TREC 2003 novelty track were of two classes: "event" and "opinion". These three questions are very important to both event topics and opinion topics. For an event topic, such as "Egyptian Air Flight 990 disaster", information about who is involved in the event, where it happened, and when it happened is very likely interesting to users. For an opinion topic, such as "opinions on prayer time in public schools", Users are probably interested in information about who made opinions related to the issue, and where or when the actions concerning this issue were taken. Therefore, we assume that each topic could be turned into these three questions (who, where and when). We call these three questions POLD questions because PERSON and ORGANIZATION are the types of named entities that who questions expect, LOCATION is the type of named entities for where questions, and DATE is for when questions.. Thus with POLD question formulation, only the four types of named entities (person, organization, location and date) in a sentence are considered as potential answers.

General question formulation. *General questions* ask for general information in that all types of named entities identified in a relevant sentence could be potential answers. We have automatically formulated general questions for each topic in the data set using the key words in the title field. We will use topic 306 in TREC 2002 again to show how general questions are formulated. The title field of topic 306 includes three key words: *African, civilian* and *deaths*. The general question we formulated for this topic is "*What information is available about African civilian deaths?*" Our system has automatically formulated general questions for all topics used in our experiments.

4.2 New Answer Detection

The new answer detection step starts with the questions generated in the process of question formulation. The task of new answer detection is carried out with an *answer updating system*, which is modified from a question answering (QA) system as in [19,20]. Once the question formulation is done, the question will be input to the answer updating system. The answer updating system has three main components: question processing module, sentence retrieval module and new answer detection module (Figure 1).

In the *question processing module*, a question is classified and the type of answer that this question expects is determined. The types of answers are characterized by the types of named entities. The next step is to find relevant sentences via the *relevant sentence* *retrieval module*. For typical QA systems as in [19,20], a query is generated with key words from the question. Then a search engine takes the query and searches in its data collection to retrieve documents that are likely to have correct answers. Our relevant sentence retrieval module takes the results in finding relevant sentences with the well-known TFIDF method as used in [7] *and* removes the sentences that do not contain any answers to the question. For specific questions, only a specific type of named entity that the question expects would be considered as its answer(s). For POLD questions, the four types of named entities (person, organization, location and date) will be considered as its answer(s). For general questions, all types of named entities could be potential answers. Then a list of presumed relevant sentences (which contain answers 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:

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

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

where S_o is the original score from the retrieval system we use, ND is the number of different type 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. The performance of finding relevant sentences using Eq. (10) is not as good as Eq. (9), which is consistent with our statistics about novelty in Section 3. Therefore, we use Eq. (9) in the sentence retrieval module for the experiments.

The *new answer detection module* then extracts answers from each sentence and marks the sentence as novel or redundant. There is an answer pool associated with each question. It is initially empty. New answers will be added to the answer pool when the answer detection module determines that the incoming answers are new. For a specific question, a sentence will be marked novel if it contains a named entity that is the type of named entity the question is asking for and the named entity is previously unseen. For a POLD (or general) question, a sentence will be marked novel if it contains previously unseen named entities of POLD (or any) types. The output of our novelty detection system is a list of sentences marked as novel.

5. EXPERIMENTS AND RESULTS

In this section, we present and discuss the main experimental results. The data used in our experiments and baselines chosen for comparison are also described.

5.1 Data

Currently, there are two sets of data available for novelty detection at the sentence level. The TREC 2002 novelty track used 54 topics and the TREC 2003 novelty track collected 50 topics. For each topic, there are up to 25 relevant documents that were algorithmically broken into sentences. A set of sentences was marked relevant, and further a subset of those sentences was marked novel. The main deference between the two sets is that the TREC 2003 novelty track collection was intended to exhibit greater redundancy and thus has less novel information [23]. Actually, 65.7% of the total relevant sentences in TREC 2003 novelty track have novel information, while 90.9% of the total relevant sentences in the 2002 track are novel sentences. The experimental results with the 2003 track data are reported in this paper because the greater redundancy in the 2003 track increases the realism of the task of novelty detection and makes more sense for comparing different novelty detection approaches. We have also used the 2002 track data in both our data analysis and experiments and have obtained very similar results to those of the 2003 track data. Results of the 2002 track data were reported in [22].

5.2 Baselines

We compared our answer updating 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 [25] is used to obtain the retrieval results of relevant sentences in both the baselines and our approach.

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 approaches (with general questions, POLD questions and specific questions) do not detect. Because of the “hard” decision (relevant or non-relevant, novel or non-novel) in the answer updating process, our novelty detection approaches 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 topics. Comparing our approaches to this baseline helps us to understand which is more important in containing new information: new words or new named entities.

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.. [6,7,8,9,10,11,12,13,24], were reported in the literature. MMR was introduced by Carbonell and Goldstein [24] 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 *i*th and *j*th sentences in the initial sentence ranking. *Q* represents the topic, *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 topic 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 [24], novelty detection at document filtering [13] and subtopic retrieval [17]. Also, MMR may incorporate various novelty measures by using different similarity matrix between sentences and 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 have performed three sets of experiments, which are (1) novelty detection performance for all 50 topics from TREC 2003 novelty track and our approaches using POLD question and general questions, compared with the three baseline approaches; (2) novelty detection performance for 15 topics with our proposed approaches using specific questions versus POLD questions and general questions; and (3) performance of finding relevant sentences (with answers).

The purpose of the first set of experiments is to compare the performance of our answer updating approaches to the three baselines on *all 50 topics*. In general question formulation, each topic was automatically transformed into a general question with the format "*What information is available about* … ?" In POLD question formulation, each topic was automatically transformed into three questions: who (person and organization), where (location) and when (date and time). The best performance of our approaches is achieved when α is 2. We tried two different similarity metrics for Sim_2 , i.e., the inner product and the cosine similarity, in MMR. Using the inner product similarity metric and setting λ to 0.3 gives the best performance for MMR. Table 6 gives the performance of our approaches and three baselines for the 50 topics. The results in Table 6 show that our approaches outperform all baselines at low recall. Our answer updating approach with POLD question formulation performs better than that with general questions. That indicates the four types of named entities (person, organization, date and location) are more important than other types of named entities in providing new information. We have the following observations and interpretation on the experimental results.

Table 6. Performance of novelty detection for all 50 topics in TREC 2003 (B-NN: Baseline with non-novelty detection; B-NW: Baseline with New word detection; B-MMR: Baseline with MMR; AU-NE: answer updating (AU) w/ new NE detection; AU-POLD: AU w/ new POLD NE detection)

Approaches	**B-NN**	**B-NW**	**B-MMR**	**AU-NE**	**AU-POLD**
# of Total Novel S.	10226	10226	10226	10226	10226
# of Novel S Retr.	8385	7385	8385	4114	3398
Average # S Retr.	513	383	513	221	178
Precision at 5 S.	0.4480	0.4640	0.4600	0.5200	0.5440
10	0.4520	0.4740	0.4880	0.5300	0.5540
15	0.4400	0.4707	0.4907	0.5293	0.5373
20	0.4400	0.4680	0.4700	0.5030	0.5230
30	0.4247	0.4567	0.4747	0.4980	0.5080
100	0.4128	0.4626	0.4014	0.4610	0.4538
200	0.3876	0.4350	0.3452	0.3695	0.3253
500	0.2948	0.2912	0.2840	0.1646	0.1359
1000	0.1677	0.1477	0.1677	0.0823	0.0680

(1). The proposed approaches outperform all baselines at low recall. The performance of our approach with POLD questions beats the baselines by 19%, 12% and 11% in terms of precision at low recall (top 20 sentences). Within top 20 sentences, our approach obtains more novel sentences than the baselines. To many users who only want to go through a small number of sentence candidates for answers, novel sentences in top 10 to 20 are more meaningful in real applications. Note that MMR performs slightly better than new word detection and the first baseline which solely uses the results from IR at low recall.

(2) The precision of our approaches at high recall, which is much lower than the three baselines, does not indicate novelty detection is worse than doing nothing, since novelty precision at high recall with more than 100 candidates does not have much practical meaning. However it indicates how many novel sentences our approach does not detect out of the retrieved sentences from the IR system. For example, within 1000 sentences (the last row of Table 6), the first baseline tells us there are 167 novel sentences on average for each topic; however our approach detected 68 sentences. The first three rows in Table 6 show a summary of all the 50 topics. Of the 10,226 novel sentences in total for the 50 topics, our approaches with general questions and POLD questions detected 4114 and 3398 correct novel sentences, respectively, whereas the number is 8385 for the first baseline B-NN. This simply means that 8,385 novel sentences appear in the 513 sentences retrieved (on average) for each of the 50 topics. As a comparison, B-MMR re-ranks all sentences in the initial retrieval ranking thus has keep all 8385 novel sentences retrieved by IR. Our approaches obtained 221 and 178 sentences per topic (on average) as the list of novel sentences. B-NW with new word detection detects 7,385 novel sentences with 383 sentences retrieved on average for each topic.

The second set of experiments is performed in order to compare the performance of our approaches with three different ways of question formulations – specific question formulation, POLD question formulation and general question formulation. We did experiments on 15 (out of 50) topics which specific questions can be automatically formulated. For each topic, specific questions were automatically formulated with key words in the title and question words, like who, where, when, how many etc., determined with our word pattern matching algorithm. Sentences without specific types of named entities that the specific questions expect were removed from the retrieval results of relevant sentences. For this reason, the average number in the novel sentence list per topic is much lower (which is 136) than that of the general questions and that of the POLD questions. The results of this set of experiments are shown in Table 7. The results indicate better performance could be achieved at low recall with specific question formulation though the performance difference on the small set of data (with 15 topics) we selected was not significant. We will further study this issue in our future work. We have also realized that challenges remain in the question formulation part. There are other potential types of specific questions that do not require named entities for answers.

Table 7. Performance of novelty detection for 15 topics in TREC 2003 with specific questions (AU-NE: answer updating (AU) w/ new NE detection; AU-POLD: AU w/ new POLD NE detection; AU-SQ: AU w/ specific questions)

Approaches	AU-NE	AU-POLD	AU-SQ
# of Total Novel S.	3990	3990	3990
# of Novel S Retr.	1607	1343	1048
Average # S Retr.	214	174	136
Precision at 5 S.	0.7067	0.7200	0.7733
10	0.7133	0.7133	0.7333
15	0.6756	0.6933	0.6889
20	0.6433	0.6700	0.6800
30	0.6244	0.6622	0.6600
100	0.5860	0.5833	0.5320
200	0.4710	0.4250	0.3460
500	0.2143	0.1791	0.1397
1000	0.1071	0.0895	0.0699

Table 8. Comparison of Performance of finding relevant sentences for 50 topics in TREC 2003

Approaches	TFIDF	Answer Updating
# of Total Relevant S.	15557	15557
Relevant S. Retrieved	12793	10563
Average # S Retr.	513	421
Precision at 5 sentences	0.7280	0.7320
10	0.6960	0.7100
15	0.6947	0.7107
20	0.7040	0.7080
30	0.6993	0.7013
100	0.6548	0.6594
200	0.6083	0.5921
500	0.4518	0.4081
1000	0.2559	0.2113

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 sentence retrieval module reranks 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 reranking process would improve the performance of finding relevant sentences. We compare the performance of finding relevant sentences with and without reranking. The comparison results are shown in Table 8, which verify our hypothesis at low recall, but the difference is not significant. But Table 6 shows that novelty detection with our approach with POLD questions significantly outperforms all three baselines at low recall. This indicates our answer updating approach makes a larger difference at the step of detecting novel sentences than at the step of finding relevant sentences.

6. Challenges in Data Collection

One of the major challenges in novelty detection is collecting data for evaluating novelty detection measures [17]. A novelty or redundancy measure is asymmetric. The novelty or redundancy of a sentence S_i depends on the order of sentences $(S_{1,}\ \ldots, S_{i-1})$ that the user has seen before it. To collect novelty judgments of each sentence with respect to all possible subsets, a human assessor has to read up to 2^{N-1} subsets. It is impossible to collect complete novelty judgments in reality. For the TREC 2002 and 2003 novelty track data, only the judgments for a particular set of sentences in a presumed order are available. There are two potential problems with this kind of data. First, it is not very accurate to evaluate a system's performance if the ranked sentences of the system have a different order from the particular set. Second, if both sentence A and sentence B are redundant but relevant sentences, A is before B in the relevant set, B will be marked redundant. However, a system might not retrieve sentence A but only B. In this case B could be considered as a novel sentence while it would still be treated as redundant using the TREC novelty judgment file.

In a novelty detection study at CMU [13], researchers initially intended to collect judgments for 50 topics, but could only get assessments for 33 topics. They provide the information on which documents before a document makes it redundant. The documents must be listed in chronological order. Thus there are problems when evaluating a novelty detection system in which documents are not output in chronological order. As research interest increases in novelty detection, more accurate and efficient data collection is crucial to the success of developing new techniques in this area.

7. 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 give 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 to use answer-updating techniques to detect new answers in incoming sentences. Thus a sentence that contains a new answer will be marked novel, which means it both is relevant to a given query and has new information. A set of experiments was performed on the TREC 2002 and 2003 novelty track data. The experimental results show that our answer updating approach outperforms the baselines with new word detection and with the MMR method. Better performance could be achieved at low recall with the specific question formulation than with the general question formulation.

We have also investigated the distributions of named entities in relevant/novel and non-relevant/non-novel sentences, and the relationship between new named entities and novelty with TREC novelty track data. 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 and some particular NE combinations. This observation has been partially incorporated in our answer-updating approach in novelty detection.

The statistics obtained from our investigation can be used to further improve the performance of finding relevant materials. In our novelty detection system, only the number of different types of named entities was considered when reranking sentences. A future effort would develop techniques to incorporate statistics on some NE combinations in order to improve the performance of novelty detection. We would also like to explore new methods to incorporate the distributions of named entities appearing in sentences. In this paper, the original belief score was adjusted with the number of different type of named entities to re-rank the retrieved sentences. We are now considering incorporating the distributions into a language modeling framework. Sentences with different number of named entities may be associated with different priors.

An important step in the proposed answer updating approach is to transform a topic into multiple specific questions. Currently, only NE questions are considered. A future work is to improve the question formulation strategy and automatically form multiple specific questions that include other types of question as well as NE questions. Ongoing research on generating multiple questions from a high level question in QA may be applied. Another future work is to apply the answer updating approach to novelty detection at the event level and the document level.

8. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, 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.

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 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] X. Li and W.B. Croft “Novelty Detection via Answer Updating”, CIIR Technical Report 334.

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

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

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