Novelty Detection via Answer Updating

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 2002 novelty track using the proposed approach. The results show that novelty detection via answer updating outperforms other novelty measures reported in the literature 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]. 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 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.

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

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]. 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.

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 distinguish 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 TREC 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:

> *<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, 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 use 101 topics where 53 topics are from the TREC 2002 novelty track and 48 topics are from a dataset collected by UMass [7]. 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 topics 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 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 by our own code.

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 two that are marked with an asterisk (one in Table 1 and one in Table 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 2. The average ratio between the named entity occurrences (in percentages) in relevant and non-relevant sentences is 1.54 in the distribution of different types of named entities (Columns 2 and 3), while the ratio is 1.45 in the distribution of named entity numbers (Columns 4 and 5). 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	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%)

(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 only 22.4%/19.4% =1.16. However the average ratio for three types of combinations of two different named entities (in Table 3) 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).

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 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	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%)

3.3 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 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 results of 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 Topics
Novel Sentences	4170	2801 (67.2%)	101
Redundant Sentences	777	355 (45.7%)	75

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 420 from TREC novelty track data is concerned about the symptoms, causes and prevention of carbon monoxide poisoning. A relevant sentence to this topic doesn't have to have any named entities to be relevant, let alone new named entities. In fact, most of the relevant sentences to this topic don't contain any named entities at all. There are about 18 such topics out of the 101 topics investigated.

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 306: "*How many civilian non-combatants have been killed in the various civil wars in Africa*?" For this topic, 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 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, either manually or automatically. We have tried two different methods for question formulation: specific question formulation and general question formulation.

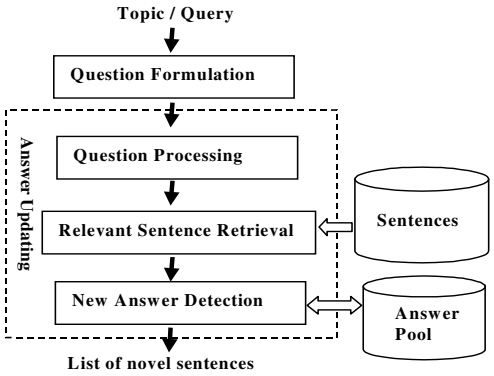

Figure 1: The proposed novelty detection system

Specific question formulation. *Specific questions* ask for specific types of named entities as answers, i.e., a topic can be transformed into NE-question(s). We note that each topic from TREC novelty track has three fields: title, description and narrative. For some of the topics, we can automatically or manually formulate specific question(s) using the key words in the title or description fields of the topics, if the topics can be transformed into NE-questions. For some topics, the right questions are readily available in the topic description. For example, for topic 306 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. Automated generation of specific questions is possible, but it is difficult. As an example, let us see topic 301, which has also been described in Section 3. Since the description field is not a question, we can manually form1ulate a specific question as: "*Which organizations participate in international criminal activity?*" We have manually formulated specific questions for 24 topics in the experiments (Section 5).

General question formulation. *General questions* ask for general information in that all types of named entities 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 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 101 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 a specific question, only a specific type of named entity that the question expects would be considered as its answer(s). For a general question, all types of named entities could be potential answers. Then a list of presumed relevant sentences (which contain answers to the question) is generated. For general questions, 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:

$$S_r = (1-\alpha)S_o + \alpha * ND \quad (9)$$

$$S_r = (1-\alpha)\, S_o + \beta * N \quad (10)$$

where S_o is the original score from the retrieval system we use, ND is the number of different type of named entities a sentence contains, N is the number of named entities and α is a weight. The weight α can be learned from some training data. In our experiments, we tuned the parameter with 52 topics. The best value of α learned in our training is 0.95. We also tried different values of 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 general question, a sentence will be marked novel if it contains a previously unseen named entity. 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

As we have mentioned earlier, 101 topics were used in our experiments. Among these topics, 53 topics are from the TREC-2002 novelty track and 48 topics are from the UMass dataset. The UMass dataset was collected by UMass researchers for their experiments in novelty detection [7]. 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.

5.2 Baselines

We could potentially compare our approach to all available approaches to novelty detection, such as simple new word count measure, set difference measure, cosine distance measure, language model measures, etc.. [6,7,8,9,10,11,12,13] Since the *set difference measure* was reported in [7] as the best novelty measures (starting with sentences returned from a well-known retrieval model that uses the vector space model with TFIDF weighting), we use this measure as our main baseline for comparison. For comparison, in our experiments, the same retrieval system based on the TFIDF technique is used to obtain the retrieval results of relevant sentences in both the baseline and our approach. The set difference measure (in the baseline) can be viewed as a more sophisticated version of the simple new word count. The novelty score of a sentence is computed as the minimum difference measure among the set of a pairwise comparison between this sentence and every previously seen sentence. Then top 10% of the sentences in each topic from the retrieval results are reranked by the computed novelty scores. Our approach starts with the same retrieval results for sentences, but then goes through two important steps for sentence retrieval and new answer detection: In the *relevant sentence retrieval module,* those sentences without any possible answers to the questions are removed, and further reranking is performed for general questions. In the *new answer detection module,* sentences are marked as novel or redundant, and the redundant sentences are removed. Because of the "hard" decision (relevant or non-relevant, novel or non-novel) instead of a "soft" ranking as in the baseline, our novelty detection approach may produce a shorter list of sentences.

Both our approach and the main baseline cannot avoid missing novel sentences, by removing non-relevant and redundant sentences in our approach and by using only the top 10% sentences from the retrieval results in the main baseline. Therefore we compare our approach with a second baseline that does not perform *any* novelty detection. The initial sentence ranking scores by the retrieval system are used directly as the novelty scores. The purpose is to see how many novel sentences our approach does not detect.

5.3 Results

We have performed four sets of experiments, which are novelty detection performance for general questions, novelty detection performance for specific questions, performance of finding relevant sentences (with answers), and performance of novelty detection using specific versus general questions.

The purpose of the first set of experiments is to compare the performance of our answer updating approach to the two baselines on *general* questions. Each topic was automatically transformed into a general question with the format "*What information is available about …* ?" Fifty-three (53) topics were used to train the parameter α in Eq. (9) and forty-nine (49) topics were used for testing. The best value of α, which is 0.95, was used in the test set. Table 6 gives the performance of our approach and two baselines for the 49 general questions. The results in Table 6 show that our approach significantly outperforms both the baselines at low recall. We have the following observations and interpretation on the experimental results.

Table 6. Performance of novelty detection for 49 general questions in the testing (S: Sentences; Chg1% / Chg2%: percentage change over Baseline 1/2; "*" indicates significant difference at 95% confidence level with t-test)

Approaches	Baseline1	Baseline2	Answer Updating		
# of Total Novel S.	1241	1241	1241		
# of Novel S Retr.	712	1204	749		
Average # of S Retr.	131	914	396	Chg1%	Chg2%
Precision at 5 S.	0.192	0.229	0.269	+40.5*	+17.9*
10	0.169	0.222	0.257	+51.8*	+15.6*
15	0.168	0.218	0.234	+39.5*	+7.5*
20	0.167	0.205	0.214	+28.1*	+4.5
30	0.154	0.195	0.189	+22.5*	-2.8
100	0.117	0.128	0.109	-6.3	-14.6*
200	0.071	0.086	0.067	-5.2	-21.8*
500	0.029	0.045	0.031	+5.2	-32.3*
1000	0.015	0.025	0.016	+5.5	-37.8*

(1). Our approach outperforms both baselines at low recall. The performance of our approach beats the first baseline by more than 20% in terms of precision at low recall (within top 30 sentences). Within top 15 sentences, our approach obtains significantly more novel sentences than the second baseline solely using IR. For example the precision of novelty detection increases 15.6% within the top 10 sentences. 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.

(2) The precision of our approach and the first baseline at high recall, which is much lower than the second baseline, 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 (and the first baseline) does not detect out of the retrieved sentences from the IR system. For example, within 1000 sentences (the last row of Table 6), the second baseline tells us there are 25 novel sentences on average for each topic; however our approach detected 16 and the first baseline detected 15 sentences. The first few rows in Table 6 show a summary of all the 49 topics. Of the 1,241 novel sentences in total for the 49 topics, our approach detected 749 correct novel sentences, whereas the number is 712 for the first baseline. The novel sentences "detected" by the second baseline is 1,204, but this simply means that 1,204 novel sentences appear in the first 914 sentences (on average) for each of the 49 topics. As a comparison, in the first baseline, the average number of sentences in top 10% of the retrieved sentences for novelty ranking is 131 for each topic. Our approach obtained 396 sentences per topic (on average) as the list of novel sentences.

In the second set of experiments (Tale 7), we compare the performance of the answer updating approach to the two baselines on 24 *specific* questions. For each topic, a specific question was manually formulated. For specific questions, the number of different types of named entities appeared in a sentence was not considered as was done for the general questions. But 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 110) than that of

the general questions in Table 6. The results of this set of experiments are shown in Table 7. Again our proposed approach has a significant performance gain at low recall. We can draw similar observations as for Table 6 but it is more interesting to see the differences between specific questions and general questions:

Due to the tighter criteria in selecting answers and new answers using specific types of named entities for a specific question, the precision at low recall further increase; Within the top 5 sentences (the first row in Table 7) the precision increases more than 10% comparing to the general question cases. On the other hand, for the same reason, more novel sentences are missing: out of 35 novel sentences in 1000 retrieved sentences, our approach only detected 8 sentences (the last row in Table 7).

Table 7. Performance of novelty detection for 24 specific questions

Approaches	**Baseline1**	**Baseline2**	**Answer Updating**	
# of Total Novel S.	977	977	977	
# of Novel S. Retr.	405	840	211	
Average # of S. Retr.	144	914	110	Chg1% Chg2%
Precision at 5 S.	0.167	0.208	0.267	+59.9* +28.4*
10	0.187	0.238	0.246	+31.6* +3.4
15	0.194	0.211	0.228	+17.5* +8.1
20	0.185	0.217	0.2	+8.1 -7.8
30	0.164	0.201	0.167	+1.8 -16.9*
100	0.133	0.139	0.082	-38.3* -41.0*
200	0.067	0.098	0.041	-38.8 -58.2*
500	0.027	0.058	0.016	-40.7* -72.4*
1000	0.013	0.035	0.008	-38.4* -77.1*

Table 8. Comparison of Performance of finding relevant sentences between for 49 general questions in the testing set

Approaches	**TFIDF**	**Answer Updating**	
# of Total Relevant S.	1365	1365	
Relevant S. Retrieved	1319	1071	
Average S. Retrieved	914	659	Chg%
Precision at 5 sentences	0.241	0.286	+18.6*
10	0.240	0.269	+11.9*
15	0.239	0.245	+2.3
20	0.225	0.233	+3.6
30	0.210	0.205	-2.6
100	0.142	0.129	-9.2
200	0.094	0.081	-14.2*
500	0.050	0.042	-15.3
1000	0.027	0.022	-18.8

The third set of experiments is designed to investigate the performance gain of finding relevant sentences for general questions with the sentence *reranking* step. Remember that, in our approach, the sentence retrieval module reranks the sentences by modifying the revised score, which is a linear combination of original belief score and the number of different types of named entities appeared in a sentence for all general questions. 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.

The last set of experiments is performed in order to compare the performance of our approach with two different ways of question formulation – specific question formulation and general question formulation. We took 16 topics out of the 101 topics. For each topic, two specific questions were manually formulated and a general question was automatically formulated. The reason for choosing two specific questions for each topic is based on our observation that two questions can often represent the topic well. The main difference between the two types of questions (specific or general) is that how named entities are treated in the answer updating step. The number of all different types of named entities appeared in a sentence was consider for general questions. But only two specific types of named entities were considered for the two specific questions associated with each topic. The results on the small set of data we selected did not show much difference between the two question formulation approaches in terms of the novelty detection performance. We will further study this issue in our future work.

5.4 Discussion

Our answer updating approach we proposed outperforms both baselines at low recall. We have been working on other approaches for improving the performance at high recall levels. However, even with the current performance, the answer updating approach is still very helpful to users who usually are more interested in finding novel information quickly and are less tolerant to redundant and non-relevant information.

A notable result shown in Table 6 and Table 7 is that the performance of our first baseline is worse than the performance of the second baseline. This indicates that, on these particular data sets we used, performing novelty detection using the set difference measure does not help in reducing the amount of redundant and non-relevant materials in general. It is possible that the gains obtained by increasing the rank of novel sentences are offset by the cost of pulling up non-relevant sentences in the ranking. There are also problems in the data sets used. There are not sufficient redundant sentences for these topics. Even worse, all relevant sentences are novel sentences in 26 topics out of the 101 topics. Therefore, any novelty detection to remove "redundant" information will decrease the performance. More appropriate data sets are desired. However, the task of generating data collection for novelty detection is very hard. The challenges of this task are discussed in the following section.

6. Challenges in Data Collection

One of the major challenges in novelty detection is collecting data for evaluating novelty detection measures. A novelty or redundancy measure is asymmetric. The novelty or redundancy of a sentence S_i depends on the order of sentences (S_1, …, S_{i-1}) that the user has seen before 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 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 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 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 novelty track data. The experimental results show that our proposed approach outperforms two baselines.

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, a linear combination of the original belief score and the number of different type of named entities was used to rerank the retrieved sentences. We are considering incorporating the distributions into a language modeling framework. Sentences with different number of named entities may be associated with different priors.

The original belief score used in the linear combination was from a retrieval system, which used the TFIDF model. The best value of α in equation (9) depends on the retrieval model used in the retrieval system; learning a parameter independent of any retrieval model is a future task. Other future work is to automatically form specific questions. Ongoing research on generating multiple questions from a high level question in QA may be applied.

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, and in part by Advanced Research and Development Activity under contract number MDA904-01-C-0984. 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 Retro-spective and On-Line event detection", *Proc. SIGIR-98*.

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

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

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

[20] X. Li, "Syntactic Features in Question Answering", *Proc. SIGIR-03,* 2003: 383-38

[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