Fact-Focused Novelty Detection: a Feasibility Study

Jahna Otterbacher*
Department of Public and
Business Administration
University of Cyprus
P.O. Box 20537
CY-1678 Nicosia, Cyprus
jahna@ucy.ac.cy

Dragomir Radev
School of Information and
Department of EECS
University of Michigan
1085 South University Ave.
304 West Hall
Ann Arbor, MI 48109-1107
radev@umich.edu

ABSTRACT

Methods for detecting sentences in an input document set, which are both relevant and novel with respect to an information need, would be of direct benefit to many systems, such as extractive text summarizers. However, satisfactory levels of agreement between judges performing this task manually have yet to demonstrated, leaving researchers to conclude that the task is too subjective. In previous experiments, judges were asked to first identify sentences that are relevant to a general topic, and then to eliminate sentences from the list that do not contain new information. Currently, a new task is proposed, in which annotators perform the same procedure, but within the context of a specific, factual information need. In the experiment, satisfactory levels of agreement between independent annotators were achieved on the first step of identifying sentences containing relevant information relevant. However, the results indicate that judges do not agree on which sentences contain novel information.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software; H.1.2 [**User/Machine Systems**]: Human factors

General Terms

Design, Experimentation, Human Factors

Keywords

Novelty, Summarization

1. INTRODUCTION

A core challenge for retrieval systems is to find information that is not only relevant to a user's need, but is also novel [1]. To this end, the task of "novelty detection," or identifying the textual units that express interesting and previously unseen information, has been put forward. In contrast to systems that retrieve all relevant information (e.g. standard search engines), systems incorporating novelty detection aim to reduce the amount of redundant information seen by the user.

A major effort in sentence-level novelty detection was the TREC Novelty Track [1]. In this evaluation, the goal was to train systems that perform a two-stage task. Given a TREC topic query and a set of relevant documents, the systems should first retrieve all sentences that are relevant to the stated topic. In the second step, the systems should choose, from the list of relevant sentences, the novel sentences, defined as those containing "previously unseen information" [3, 5]. Several problems were noted by the organizers in creating the manually-labeled data sets for the training and evaluation of the systems. For example, in the first year of the track, the assessors chose few relevant sentences, resulting in many negative and few positive relevance examples available for training systems. While this was resolved in later years, one persistent problem was the large assessor effect, such that the judges did not reach a high level of agreement as to which sentences were relevant and novel.

Recently, Schiffman created his own corpus of novelty judgments, and tried to improve on the TREC annotation results by showing judges only two news articles at a time [4]. One article was presented as background information, while the other was considered a new document. The judges were then to choose the spans of text in the new document that presented novel information not contained in the background article. However, the new annotation task also yielded a low rate of agreement between independent judges.

Currently, we propose a new sentence-level annotation task - fact-based relevance and novelty detection. We assume that a user has a general topic of interest, and has identified a set of relevant documents. Next, we assume that the user has a set of specific facts of interest about the topic. For simplicity, the user states each fact as a natural language question. Our task is, for a given fact, to first identify the set of sentences in the document set that contain relevant information. A sentence contains relevant information only if it provides an answer to the question. In a second step, only the sentences containing previously unseen information about the fact of interest are kept. In the current paper, we evaluate the reproducibility of the two steps of the proposed task.

2. EXPERIMENTAL SETUP

The data for our experiment come from the 2003 TREC Novelty track test data [5]. Novelty track clusters consist of a topic query and a set of 25 relevant news articles. We chose two of the "event"

*This work was conducted while the first author was at the University of Michigan's School of Information.

[1] $http://trec.nist.gov$

SIGIR'06, August 6–10, 2006, Seattle, Washington, USA.

clusters for our experiment and show their attributes in Table 1. We read through all of the documents in each of the two stories, and created a list of ten factual questions that are central to each story. The questions ask about key facts in the stories, that may change with time as news sources publish additional information, and that expect atomic answers such as a number, name of a person, or a place.

Cluster	Subject	Example question
N4	Egyptian Air disaster 990	How many people were on board?
N33	Sinking of Russian submarine Kursk	Where did the Kursk sink?

Table 1: Data clusters used in the experiment.

Six paid subjects were hired for the study. Three were randomly assigned to the test (fact-based) setting and three to the control (topic-based) setting. Each subject performed the respective task on both clusters. In both settings, the judges were given the 25 documents in a given cluster, ordered chronologically. Judges in the control setting were given the TREC topic query, while those in the test setting were given the set of factual questions. Both sets of subjects were asked to first familiarize themselves with the news story by reading either the given query or set of questions, as well as by skimming through the news articles. Judges in the control setting were asked to complete the TREC task - that is, to first identify those sentences that are relevant to the topic, and then to reread the relevant sentences, eliminating those that do not contain information that has not been previously seen [5]. Annotators in the test setting were to find, for each question, the set of relevant sentences containing an answer to the question. They were then asked to reread through their set of sentences in chronological order, eliminating those that do not provide a new answer to the given question.

3. RELEVANCE JUDGMENTS

Table 2 shows the agreement on the task of finding sentences that are relevant to the topic (in the control setting) and relevant to the factual questions (in the test setting). For the topic-based judgments, over the two clusters N4 and N33, there were a total of 1,636 sentence judgments. The three judges agreed on the relevance status of 39% of the sentences. The corresponding Kappa statistic, which factors out the expected (chance) agreement, is also shown [2]. In the fact-based setting, there are a total of 16,360 sentence-level relevance judgments (i.e. 10 questions for each document cluster). The three subjects' judgments were in agreement in 99% of these cases, which corresponds to a Kappa of 0.67.

4. NOVELTY JUDGMENTS

The agreement on the novelty judgments is shown in Table 3. In order to calculate novelty agreement, we first found the union of the judges' sets of relevant sentences. In other words, we considered the agreement on novelty status among the sentences that any judge had labeled as being relevant. As can be seen in the table, the three judges agreed on 21% of the sentences in the control setting. This level of agreement is actually less than what is expected by chance (corresponding to a negative Kappa of -0.06). In the fact-based setting, we see that the judges agreed on 52% of the sentences. However, while agreement is better in the test setting as compared to the topic-based setting, a Kappa of 0.18 certainly does not indicate a sufficiently high level of agreement on the fact-based novelty annotation task.

	Prop. agree	Kappa
Topic-based	0.39	0.15
Fact-based	0.99	0.67

Table 2: Interjudge agreement on relevance.

	Prop. agree	Kappa
Topic-based	0.21	-0.06
Fact-based	0.52	0.18

Table 3: Interjudge agreement on novelty.

5. CONCLUSION

We proposed the problem of fact-focused relevance and novelty detection at the sentence level. We conducted a preliminary evaluation of its reproducibility, both on the relevance and novelty stages of the task. In addition, we reproduced the TREC Novelty annotation experiment, in which the judges found relevant and novel sentences with respect to a general topic query. Our results thus far suggest that a higher level of agreement can be reached on relevance judgments when they are constrained to apply to a set of factual questions, as compared to the case of topic-based judgments. In addition, while the agreement on novelty judgments in the fact-based task was better than in the topic-based case, the level of agreement was not very satisfactory. In our immediate future work, we will further analyze the data from the current experiment in a number of ways. In particular, we plan to qualitatively analyze the annotations of our judges in order to determine if there are features that can be used to identify sentences upon which judges are unlikely to agree as to relevance and novelty status.

6. ACKNOWLEDGMENTS

This work was partially supported by the U.S. National Science Foundation under the following grant: 0329043 "Probabilistic and link-based Methods for Exploiting Very Large Textual Repositories" administered through the IDM program. All opinions, findings, conclusions, and recommendations in this paper are made by the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank the members of the CLAIR research group and the anonymous SIGIR reviewers for their feedback and comments on this work.

7. REFERENCES

[1] J. Allan, B. Carterette, and J. Lewis. When Will Information Retrieval Be "Good Enough"? In *28th Annual ACM SIGIR (SIGIR '05)*, Salvador, Brazil, August 2005.

[2] J. Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. 22(2):249–254, 1996.

[3] D. Harman. Overview of the TREC 2002 novelty track, 2002.

[4] B. Schiffman. *Learning to Identify New Information*. PhD thesis, Department of Computer Science, Columbia University, 2005.

[5] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty Track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, NIST, Gaithersburg, ML, 2003.