Novelty Detection: The TREC Experience

Ian Soboroff and **Donna Harman**
National Institute of Standards and Technology
Gaithersburg, MD
(ian.soboroff,donna.harman)@nist.gov

Abstract

A challenge for search systems is to detect not only when an item is relevant to the user's information need, but also when it contains something new which the user has not seen before. In the TREC novelty track, the task was to highlight sentences containing relevant and new information in a short, topical document stream. This is analogous to highlighting key parts of a document for another person to read, and this kind of output can be useful as input to a summarization system. Search topics involved both news events and reported opinions on hot-button subjects. When people performed this task, they tended to select small blocks of consecutive sentences, whereas current systems identified many relevant and novel passages. We also found that opinions are much harder to track than events.

1 Introduction

The problem of novelty detection has long been a significant one for retrieval systems. The "selective dissemination of information" (SDI) paradigm assumed that the people wanted to be able to track new information relating to known topics as their primary search task. While most SDI and information filtering systems have focused on similarity to a topical profile (Robertson, 2002) or to a community of users with a shared interest (Belkin and Croft, 1992), recent efforts (Carbonell and Goldstein, 1998; Allan et al., 2000; Kumaran et al., 2003) have looked at the retrieval of specifically *novel* information.

The TREC novelty track experiments were conducted from 2002 to 2004 (Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004). The basic task was defined as follows: given a topic and an ordered set of documents related to that topic, segmented into sentences, return those sentences that are both relevant to the topic and novel given what has already been seen previously in that document set. This task models an application where a user is skimming a set of documents, and the system highlights new, on-topic information.

There are two problems that participants must solve in this task. The first is identifying relevant sentences, which is essentially a passage retrieval task. Sentence retrieval differs from document retrieval because there is much less text to work with, and identifying a relevant sentence may involve examining the sentence in the context of those surrounding it. The sentence was specified as the unit of retrieval in order to standardize the task across a variety of passage retrieval approaches, as well as to simplify the evaluation.

The second problem is that of identifying those relevant sentences that contain new information. The operational definition of "new" here is information that has not appeared previously in this topic's set of documents. In other words, we allow the system to assume that the user is most concerned about finding new information in this particular set of documents, and is tolerant of reading information he already knows because of his background knowledge. Since each sentence adds to the user's knowledge, and later sentences are to be retrieved only if they contain new information, novelty retrieval resembles a filtering task.

Novelty is an inherently difficult phenomenon to operationalize. Document-level novelty detection, while intuitive, is rarely useful because nearly every document contains something new, particularly when the domain is news. Hence, our decision to use sentences as the unit of retrieval. Moreover, determining ground truth for a novelty detection task is more difficult than for topical relevance, because one is forced not only to face the idiosyncratic na-

ture of relevance, but also to rely all the more on the memory and organizational skills of the assessor, who must try and remember everything he has read. We wanted to determine if people could accomplish this task to any reasonable level of agreement, as well as to see what computational approaches best solve this problem.

2 Input Data

The first year of the novelty track (Harman, 2002) was a trial run in several ways. First, this was a new task for the community and participating groups had no training data or experience. But second, it was unclear how humans would perform this task and therefore creating the "truth" data was in itself a large experiment. NIST decided to minimize the cost by using 50 old topics from TRECs 6, 7, and 8.

The truth data was created by asking NIST assessors (the humans performing this task) to identify the set of relevant sentences from each relevant document and then from that set of relevant sentences, mark those that were novel. Specifically, the assessors were instructed to identify a list of sentences that were:

1. relevant to the question or request made in the description section of the topic,
2. their relevance was independent of any surrounding sentences,
3. they provided new information that had not been found in any previously picked sentences.

Most of the NIST assessors who worked on this task were not the ones who created the original topics, nor had they selected the relevant documents. This turned out to be a major problem. The assessors' judgments for the topics were remarkable in that only a median of 2% of the sentences were judged to be relevant, despite the documents themselves being relevant. As a consequence, nearly every relevant sentence (median of 93%) was declared novel. This was due in large part to assessor disagreement as to relevancy, but also that fact that this was a new task to the assessors. Additionally, there was an encouragement not to select consecutive sentences, because the goal was to identify relevant and novel sentences minimally, rather than to try and capture coherent blocks of text which could stand alone. Unfortunately, this last instruction only served to confuse the assessors. Data from 2002 has not been included in the rest of this paper, nor are groups encouraged to use that data for further experiments because of these problems.

In the second year of the novelty track (Soboroff and Harman, 2003), the assessors created their own new topics on the AQUAINT collection of three contemporaneous newswires. For each topic, the assessor composed the topic and selected twenty-five relevant documents by searching the collection. Once selected, the documents were ordered chronologically, and the assessor marked the relevant sentences and those relevant sentences that were novel. No instruction or limitation was given to the assessors concerning selection of consecutive sentences, although they were told that they did not need to choose an otherwise irrelevant sentence in order to resolve a pronoun reference in a relevant sentence. Each topic was independently judged by two different assessors, the topic author and a "secondary" assessor, so that the effects of different human judgments could be measured. The judgments of the primary assessor were used as ground truth for evaluation, and the secondary assessor's judgments were taken to represent a ceiling for system performance in this task.

Another new feature of the 2003 data set was a division of the topics into two types. Twenty-eight of the fifty topics concerned events such as the bombing at the 1996 Olympics in Atlanta, while the remaining topics focused on opinions about controversial subjects such as cloning, gun control, and same-sex marriages. The topic type was indicated in the topic description by a `<toptype>` tag.

This pattern was repeated for TREC 2004 (Soboroff, 2004), with fifty new topics (twenty-five events and twenty-five opinion) created in a similar manner and with the same document collection. For 2004, assessors also labeled some documents as irrelevant, and irrelevant documents up through the first twenty-five relevant documents were included in the document sets distributed to the participants. These irrelevant documents were included to increase the "noise" in the data set. However, the assessors only judged sentences in the relevant documents, since, by the TREC standard of relevance, a document is considered relevant if it contains any relevant information.

3 Task Definition

There were four tasks in the novelty track:

Task 1. Given the set of documents for the topic, identify all relevant and novel sentences.

Task 2. Given the relevant sentences in all documents, identify all novel sentences.

Task 3. Given the relevant and novel sentences in the first 5 documents **only**, find the relevant

and novel sentences in the remaining documents. Note that since some documents are irrelevant, there *may not be* any relevant or novel sentences in the first 5 documents for some topics.

Task 4. Given the relevant sentences from all documents, and the novel sentences from the first 5 documents, find the novel sentences in the remaining documents.

These four tasks allowed the participants to test their approaches to novelty detection given different levels of training: none, partial, or complete relevance information, and none or partial novelty information.

The test data for a topic consisted of the topic statement, the set of sentence-segmented documents, and the chronological order for those documents. For tasks 2-4, training data in the form of relevant and novel "sentence qrels" were also given. The data was released and results were submitted in stages to limit "leakage" of training data between tasks. Depending on the task, the system was to output the identifiers of sentences which the system determined to contain relevant and/or novel relevant information.

4 Evaluation

Because novelty track runs report their relevant and novel sentences as an unranked set, traditional measures of ranked retrieval effectiveness such as mean average precision can't be used. One alternative is to use set-based recall and precision. Let M be the number of matched sentences, i.e., the number of sentences selected by both the assessor and the system, A be the number of sentences selected by the assessor, and S be the number of sentences selected by the system. Then sentence set recall is $R = M/A$ and precision is $P = M/S$.

However, set-based recall and precision do not average well, especially when the assessor set sizes A vary widely across topics. Consider the following example as an illustration of the problems. One topic has hundreds of relevant sentences and the system retrieves 1 relevant sentence. The second topic has 1 relevant sentence and the system retrieves hundreds of sentences. The average for both recall and precision over these two topics is approximately .5 (the scores on the first topic are 1.0 for precision and essentially 0.0 for recall, and the scores for the second topic are the reverse), even though the system did precisely the wrong thing. While most real systems wouldn't exhibit this extreme behavior, the fact remains that set recall and set precision averaged over a set of topics is not a robust diagnostic indicator

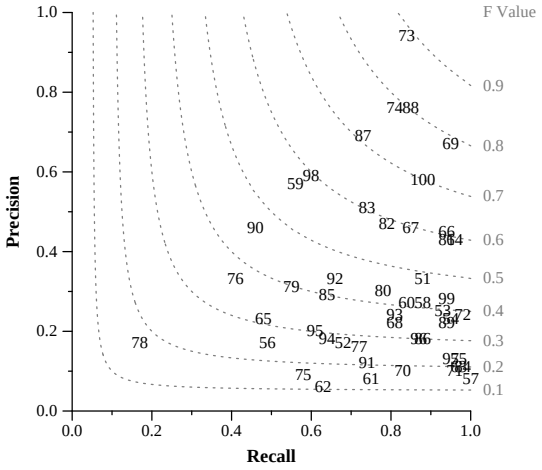

Figure 1: The F measure, plotted according to its precision and recall components. The lines show contours at intervals of 0.1 points of F. The black numbers are per-topic scores for one TREC system.

of system performance. There is also the problem of how to define precision when the system returns no sentences ($S = 0$). Leaving that topic out of the evaluation for that run would mean that different systems would be evaluated over different numbers of topics. The standard procedure is to define precision to be 0 when $S = 0$.

To avoid these problems, the primary measure used in the novelty track was the F measure. The F measure (which is itself derived from van Rijsbergen's E measure (van Rijsbergen, 1979)) is a function of set recall and precision, together with a parameter β which determines the relative importance of recall and precision:

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

A β value of 1, indicating equal weight, is used in the novelty track:

$$\mathrm{F}_{\beta=1} = \frac{2PR}{P + R}$$

Alternatively, this can be formulated as

$$\mathrm{F}_{\beta=1} = \frac{2 \times (\#\text{ relevant retrieved})}{(\#\text{ retrieved}) + (\#\text{ relevant})}$$

For any choice of β, F lies in the range $[0, 1]$, and the average of the F measure is meaningful even when the judgment sets sizes vary widely. For example, the F measure in the scenario above is essentially 0, an intuitively appropriate score for such behavior. Using the F measure also deals with the problem of

what to do when the system returns no sentences since recall is 0 and the F measure is legitimately 0 regardless of what precision is defined to be.

Note, however, that two runs with equal F scores do not indicate equal precision and recall. The contour lines in Figure 1 illustrate the shape of the F measure in recall-precision space. An F score of 0.5, for example, can describe a range of precision and recall scores. Figure 1 also shows the per-topic scores for a particular TREC run. It is easy to see that topics 98, 83, 82, and 67 exhibit a wide range of performance, but all have an F score of close to 0.6. Thus, two runs with equal F scores may be performing quite differently, and a difference in F scores can be due to changes in precision, recall, or both. In practice, if F is used, precision and recall should also be examined, and we do so in the analysis which follows.

5 Analysis

5.1 Analysis of truth data

Since the novelty task requires systems to automatically select the same sentences that were selected manually by the assessors, it is important to analyze the characteristics of the manually-created truth data in order to better understand the system results. Note that the novelty task is both a passage retrieval task, i.e., retrieve relevant sentences, and a novelty task, i.e., retrieve only relevant sentences that contain new information.

In terms of the passage retrieval part, the TREC novelty track was the first major investigation into how users select relevant parts of documents. This leads to several obvious questions, such as what percentage of the sentences are selected as relevant, and do these sentences tend to be adjacent/consecutive? Additionally, what kinds of variation appear, both across users and across topics. Table 1 shows the median percentage of sentences that were selected as relevant, and what percentage of these sentences were consecutive. Since each topic was judged by two assessors, it also shows the percentage of sentences selected by assessor 1 (the "official" assessor used in scoring) that were also selected by assessor 2. The table gives these percentages for all topics and also broken out into the two types of topics (events and opinions).

First, the table shows a large variation across the two years. The group in 2003 selected more relevant sentences (almost 40% of the sentences were selected as relevant), and in particular selected many consecutive sentences (over 90% of the relevant sentences were adjacent). The median length of a string of consecutive sentences was 2; the mean was 4.252 sentences. The following year, a different group of assessors selected only about half as many relevant sentences (20%), with fewer consecutive sentences. This variation across years may reflect the group of assessors in that the 2004 set were TREC "veterans" and were more likely to be very selective in terms of what was considered relevant.

The table also shows a variation across topics, in particular between topics asking about events versus those asking about opinions. The event topics, for both years, had more relevant sentences, and more consecutive sentences (this effect is more apparent in 2004).

Agreement between assessors on which sentences were relevant was fairly close to what is seen in document relevance tasks. There was slightly more agreement in 2003, but there were also many more relevant sentences so the likelihood of a match was higher. There is more agreement on events than on opinions, partially for the same reason, but also because there is generally less agreement on what constitutes an opinion. These medians hide a wide range of judging behavior across the assessors, particularly in 2003.

The final two rows of data in the table show the medians for novelty. There are similar patterns to those seen in the relevant sentence data, with the 2003 assessors clearly being more liberal in judging. However, the pattern is reversed for topic types, with more sentences being considered relevant and novel for the opinion topics than for the event topics. The agreement on novelty is less than on relevance, particularly in 2004 where there were smaller numbers of novel and relevant sentences selected.

Another way to look at agreement is with the kappa statistic (Cohen, 1960). Kappa computes whether two assessors disagree, with a correction for "chance agreement" which we would expect to occur randomly. Kappa is often interpreted as the degree of agreement between assessors, although this interpretation is not well-defined and varies from field to field (Di Eugenio, 2000). For relevant sentences across all topics in the 2004 data set, the kappa value is 0.549, indicating statistically significant agreement between the assessors but a rather low-to-moderate degree of agreement by most scales of interpretation. Given that agreement is usually not very high for relevance judgments (Voorhees, 1998), this is as expected.

5.2 Analysis of participants results

Most groups participating in the 2004 novelty track employed a common approach, namely to measure relevance as similarity to the topic and novelty as

		2003	2004
Relevant	all topics	0.39	0.20
	events only	0.47	0.25
	opinions only	0.38	0.15
Consecutive	all topics	0.91	0.70
	events only	0.93	0.85
	opinions only	0.91	0.65
Relevant agreement	all topics	0.69	0.60
	events only	0.82	0.68
	opinions only	0.63	0.50
Novelty	all topics	0.68	0.40
	events only	0.61	0.38
	opinions only	0.73	0.42
Novelty agreement	all topics	0.56	0.35
	events only	0.65	0.45
	opinions only	0.48	0.29

Table 1: Median fraction of sentences which were relevant and novel, fraction of consecutive relevant sentences, and proportion of agreement by the secondary assessor.

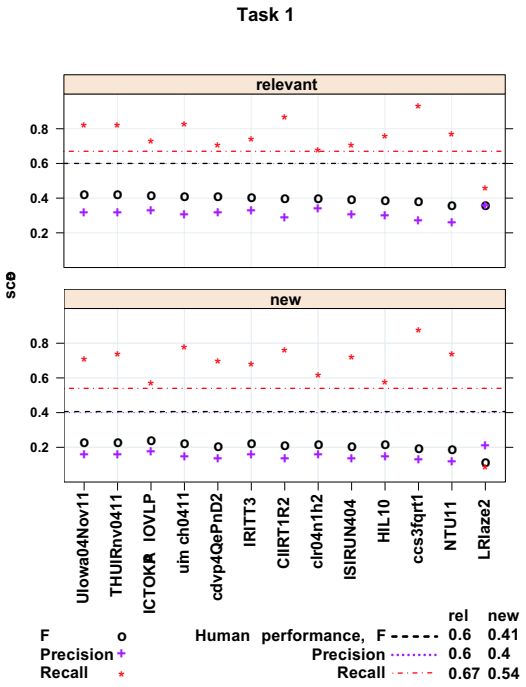

Figure 2: Task 1 precision, recall, and F scores for the top run from each group in TREC 2004

dissimilarity to past sentences. On top of this framework the participants used a wide assortment of methods which may be broadly categorized into statistical and linguistic methods. Statistical methods included using traditional retrieval models such as tf.idf and Okapi coupled with a threshold for retrieving a relevant or novel sentence, expansion of the topic and/or document sentences using dictionaries or corpus-based methods, and using named entities as features. Some groups also used machine learning algorithms such as SVMs in parts of their detection process. Semantic methods included deep parsing, matching discourse entities, looking for particular verbs and verb phrases in opinion topics, coreference resolution, normalization of named entities, and in one case manual construction of ontology's for topic-specific concepts.

Figure 2 shows the Task 1 results for the top run from each group in TREC 2004. Groups employing statistical approaches include UIowa, CIIR, UMich, and CDVP. Groups employing more linguistic methods include CLR, CCS, and LRI. THU and ICT took a sort of kitchen-sink approach where each of their runs in each task tried different techniques, mostly statistical.

The F scores for both relevance and novelty retrieval are fairly uniform, and they are dominated by the precision component. The top scoring systems by F score are largely statistical in nature; for example, see (Abdul-Jaleel et al., 2004) (CIIR) and (Eichmann et al., 2004) (UIowa). CLR (Litkowski, 2004) and LRI (Amrani et al., 2004), which use much stronger linguistic processing, achieve the highest precision at the expense of recall. Overall, precision is quite low and recall is high, implying that most systems are erring in favor of retrieving many sentences.

A closer comparison of the runs among themselves and to the truth data confirms this hypothesis. While the 2004 assessors were rather selective in choosing relevant and novel sentences, often selecting just a handful of sentences from each document, the systems were not. The systems retrieved an average of 49.5% of all sentences per topic as relevant, compared to 19.2% chosen by the assessor. Furthermore, the runs chose 41% of all sentences (79% of their own relevant sentences) as novel, compared to the assessor who selected only 8.4%. While these numbers are a very coarse average that ignores differences between the topics and between the documents in each set, it is a fair summary of the data. Most of the systems called nearly every sentence relevant and novel. By comparison, the person attempting this task (the second assessor, scored as a run and shown as horizontal lines in Figure 2) was much more effective than the systems.

The lowest scoring run in this set, LRIaze2, actually has the highest precision for both relevant and

novel sentences. The linguistics-driven approach of this group included standardizing acronyms, building a named-entity lexicon, deep parsing, resolving coreferences, and matching concepts to manually-built, topic-specific ontologies (Amrani et al., 2004). A close examination of this run's pattern shows that they retrieved very few sentences, in line with the amounts chosen by the assessor. They were not often the correct sentences, which accounts for the low recall, but by not retrieving too many false alarms, they managed to achieve a high precision.

Our hypothesis here is that the statistical systems, which are essentially using algorithms designed for document retrieval, approached the sentences with an overly-broad term model. The majority of the documents in the data set are relevant, and so many of the topic terms are present throughout the documents. However, the assessor was often looking for a finer-grained level of information than what exists at the document level. For example, topic N51 is concerned with Augusto Pinochet's arrest in London. High-quality content terms such as Pinochet, Chile, dictator, torture, etc appear in nearly every sentence, but the key relevant ones — which are very few — are those which specifically talk about the arrest. Most systems flagged nearly every sentence as relevant, when the topic was much narrower than the documents themselves.

One explanation for this may be in how thresholds were learned for this task. Since task 1 provides no data beyond the topic statement and the documents themselves, it is possible that systems were tuned to the 2003 data set where there are more relevant sentences. However, this isn't the whole story, since the difference in relevant sentences between 2003 and 2004 is not so huge that it can explain the rates of retrieval seen here. Additionally, in task 3 some topic-specific training data was provided, and yet the effectiveness of the systems was essentially the same.

Of those systems that tried a more fine-grained approach, it appears that it is complicated to learn exactly which sentences contain the relevant information. For example, nearly every system had trouble identifying relevant opinion sentences. One might expect that those systems which analyzed sentence structure more closely would have done better here, but there is essentially no difference. Identifying relevant information at the sentence level is a very hard problem.

We see very similar results for novel sentence retrieval. Rather than looking at task 1, where systems retrieved novel from their own selection of relevant sentences, it's better to look at runs in task 2 (Figure 3). Since in this task the systems are given all relevant sentences and just search for novelty, the baseline performance for comparison is just labeling all the sentences as novel. Most systems, surprisingly including the LRI run, essentially do retrieve nearly every sentence as novel. The horizontal lines show the baseline performance; the baseline recall is 1.0 and is at the top of the Y axis. All the runs except clr04n2 are just above this baseline, with cdvp4NTerFr1 and novcolrcl the most discriminating.

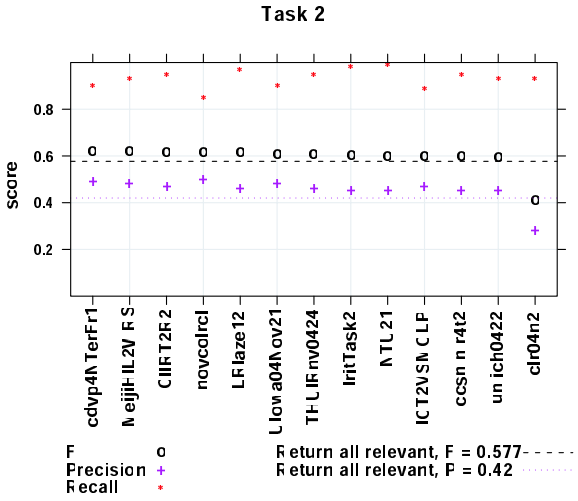

Figure 3: Task 2 scores for the top run from each group in TREC 2004

The approach of Dublin City University (cdvp4NTerFr1) is essentially to set a threshold on the tf.idf value of the unique words in the given sentence, but their other methods which incorporate the history of unique terms and the difference in sentence frequencies between the current and past sentences perform comparably (Blott et al., 2004). Similarly, Columbia University (novcolrcl) focuses on previously unseen words in the current sentence as the main evidence of novelty (Schiffman and McKeown, 2004). As opposed to the ad hoc threshold in the DCU system, Columbia employs a hill-climbing approach to learning the threshold. This particular run is optimized for recall; another optimized for precision achieved the highest precision of all task 2 runs, but with very low recall. In general, we conclude that most systems achieving high scores in novelty detection are recall-oriented and as a result still provide the user with too much information.

As was mentioned above, opinion topics proved much harder than events. Every system but one did better on event topics than on opinions in task 1

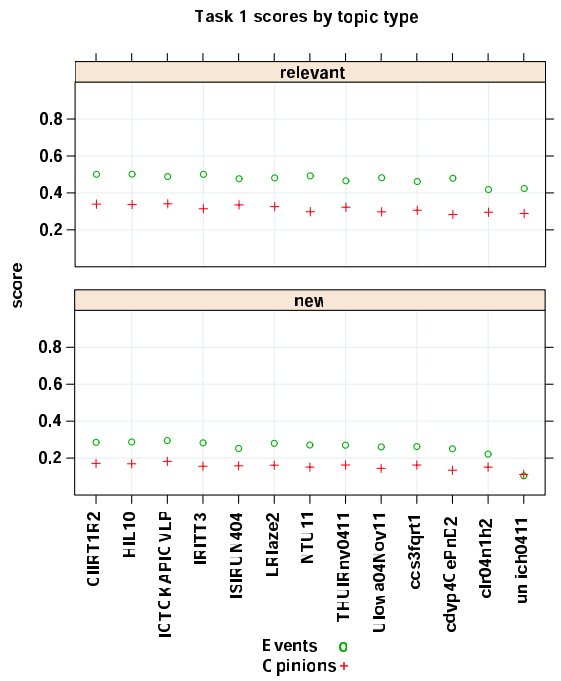

Figure 4: F scores for event and opinion topics in task 1.

(Figure 4). In task 2, where all relevant sentences were provided, many runs do as well or better on opinion topics than events. Thus, the complexity for opinions is more in finding which sentences contain them, than determining which opinions are novel.

6 Conclusion

The novelty track in TREC examined a particular kind of novelty detection, that is, finding novel, on-topic sentences within documents that the user is reading. Both statistical and linguistic techniques, as well as filtering and learning approaches can be applied to detecting novel relevant information within documents, but nevertheless it is a hard problem for several reasons. First, because the unit of interest is a sentence, there is not a lot of data in each unit on which to base the decision. When the document as a whole is relevant, techniques designed for document retrieval seem unable to make fine distinctions about which sentences within the document contain the relevant information. Initial threshold setting is critical and difficult.

When we examined human performance on this task, it is clear that users do make very fine distinctions. Looking particularly at the 2004 set of relevant and novel sentences, less than 20% of the sentences in relevant documents were marked as relevant, and only 40% of those (or 8% of the total sentences) were marked as both relevant and novel.

The TREC novelty data sets themselves support some interesting uses outside of the novelty track. Whereas the data from 2002 is clearly flawed and should not be used, the data from 2003 and 2004 can be regarded as valid samples of user input in terms of relevant sentence selection, and further reduction of those sentences to those presenting new information. One obvious use is in the passage retrieval arena, e.g., using the relevant sentences for testing passage retrieval, either at the single sentence level or using the consecutive sentences to test when to retrieve multiple sentences. A second use is for summarization, where the relevant AND novel sentences can serve as the truth data for the extraction phase (and then compressed in some manner). Other uses of the data include manual analysis of user behavior when processing documents in response to a question, or looking further into the user agreement issues, particularly in the summarization area.

The novelty data is also unique in that it deliberately contains a mix of topics on events and on opinions regarding controversial subjects. The opinions topics are quite different in this regard than other TREC topics, which have historically focused on events or narrative information on a subject or person. This exploration has been an interesting and fruitful one. By mixing the two topic types within each task, we see that identifying opinions within documents is hard, even with training data, while detecting new opinions (given relevance) seems analogous to detecting new information about an event.

References

Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, `http://trec.nist.gov/pub/trec13/papers/umass.novelty.hard.pdf`.

James Allan, Victor Lavrenko, and Hubert Jin. 2000. First story detection in TDT is hard. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM 2000)*, pages 374–381.

Ahmed Amrani, Jérôme Azé, Thomas Heitz, Yves Kodratoff, and Mathieu Roche. 2004. From the texts to the concepts they contain: a chain of linguistic treatments. In *Proceedings of the Thirteenth Text REtrieval Confer-*

ence (TREC 2004), http://trec.nist.gov/pub/trec13/papers/uparis.novelty2.pdf.

Nicholas J. Belkin and W. Bruce Croft. 1992. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38, December.

Stephen Blott, Oisin Boydell, Fabrice Camous, Paul Ferguson, Georgina Gauhan, Cathal Gurrin, Gareth J. F. Jones, Noel Murphy, Noel O'Connor, Alan F. Smeaton, Barry Smyth, and Peter Wilkins. 2004. Experiments in terabyte searching, genomic retrieval and novelty detection for TREC-2004. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, http://trec.nist.gov/pub/trec13/papers/dcu.tera.geo.novelty.pdf.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (SIGIR '98)*, Melbourne, Australia, August, pages 335–336. ACM Press.

J. A. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Barbara Di Eugenio. 2000. On the usage of kappa to evaluate agreement on coding tasks. In *Proceedings of the Second International Conference on Language Resources and Evaluatation (LREC 2000)*, Athens, Greece, pages 441–446.

David Eichmann, Yi Zhang, Shannon Bradshaw, Xin Ying Qui, Li Zhou, Padmini Srinivasan, Aditya Kumar Sehgal, and Hudon Wong. 2004. Novelty, question answering and genomics: the University of Iowa response. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, http://trec.nist.gov/pub/trec13/papers/uiowa.novelty.qa.geo.pdf.

Donna Harman. 2002. Overview of the TREC 2002 novelty track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, NIST Special Publication 500-251, pages 46–55, Gaithersburg, MD, November.

Girindhar Kumaran, James Allan, and Andrew McCallum. 2003. Classification models for new event detection. Technical Report IR-362, CIIR, University of Massachusetts, Amherst.

Kenneth C. Litkowski. 2004. Evolving XML and dictionary strategies for question answering and novelty tasks. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, http://trec.nist.gov/pub/trec13/papers/clresearch.qa.novelty.pdf.

Stephen E. Robertson. 2002. Introduction to the special issue: Overview of the TREC routing and filtering tasks. *Information Retrieval*, 5:127–137.

Barry Schiffman and Kathleen R. McKeown. 2004. Columbia university in the novelty track at TREC 2004. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, http://trec.nist.gov/pub/trec13/papers/columbiau.novelty.pdf.

Ian Soboroff and Donna Harman. 2003. Overview of the TREC 2003 novelty track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, NIST Special Publication 500-255, Gaithersburg, MD, November.

Ian Soboroff. 2004. Overview of the TREC 2004 novelty track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, http://trec.nist.gov/pub/trec13/papers/NOVELTY.OVERVIEW.pdf.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworths.

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (SIGIR '98)*, Melbourne, Australia, August, pages 315–323. ACM Press.