An Evaluation Corpus For Temporal Summarization

Vikash Khandelwal, Rahul Gupta, and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{vikas,rgupta,allan}@cs.umass.edu

ABSTRACT

In recent years, a lot of work has been done in the field of Topic Tracking. The focus of this work has been on identifying stories belonging to the same topic. This might result in a very large number of stories being reported to the user. It might be more useful to a user if a summary of the main events in the topic rather than the entire collection of stories related to the topic were presented. Though work on such a fine-grained level has been started, there is currently no standard evaluation testbed available to measure the accuracy of such techniques. We describe a scheme for developing a testbed of user judgments which can be used to evaluate the above mentioned techniques. The corpus that we have created can also be used to evaluate single or multi-document summaries.

1. THE PROBLEM

In recent years, a lot of progress has been made in the field of Topic Tracking ([2], [3], [8], etc). The focus of this work has been on identifying news stories belonging to the same topic. This might result in a very large number of stories being reported to the user. It would be more useful to a user if a summary of the main events/developments in the topic rather than the entire collection of stories related to the topic were presented. We can formulate the problem as follows.

We are given a stream of chronologically ordered and topically related stories. We strive to identify the shifts in the topic which represent the developments within the topic. For example, consider the topic "*2000 Presidential Elections*". On the night of November 7, there were reports of Gore conceding defeat to Bush. The next morning, there were reports claiming his retraction of the previous concession. Most of the stories on the next day would also contain old information including details of Gore's first phone call to Bush. We want to present only the new development (i.e., Gore's retraction) on the next day.

We assume that sentence extracts can identify such topic shifts. At the very least, they can convey enough information to a user to keep track of the developments within that topic. For example, in Figure 1, the mappings indicate how the sentences in a story correspond to events.

Human judgments are required to evaluate accuracy of extracts. The approach usually taken is to have each such extract evaluated by human beings but such a process is expensive and time consuming. We need an evaluation corpus similar to the TDT or TREC corpora that can be used over and over again to do such evaluations automatically. We propose a new scheme for building such a corpus.

Summarization evaluation is difficult because summaries can be created for a range of purposes. The Tipster SUMMAC evaluation [7] required human assessors to evaluate each summary, and most other evaluations have also required human checking of every summary [6]. There are others who have attempted automatic evaluations ([5], [9]) but none of these evaluation schemes captures all the desirable properties in a summary.

The particular problem of summarizing shifts in a news topic was attacked slightly differently at a Summer 1999 workshop on Novelty Detection [4]. Those efforts towards "*new information detection*" were a dead end because the granularity of new information was too small, e.g., a mention of a person's age might count as new information even when it is not the focus of the story. Swan and Allan also created an event-level summary "timeline" ([10], [11]) but they did not develop any evaluation corpus for their work.

This paper is organised as follows. In Section 2, we discuss the desirable properties of such an evaluation corpus. Section 3 discusses the entire annotation process, as well as the interesting practical issues, the problems faced and then the statistics of the corpus that we have built. Finally, in Section 4, we discuss one possible way of utilising this corpus.

2. DESIRABLE PROPERTIES OF THE EVALUATION CORPUS

Any evaluation corpus of sentence extracts and events which is to be used for the purpose of evaluating summaries of topic shifts in a news stream should have the following properties:

- It should be possible to identify all new events on a periodic basis. This would be required to estimate the recall of a system.

.

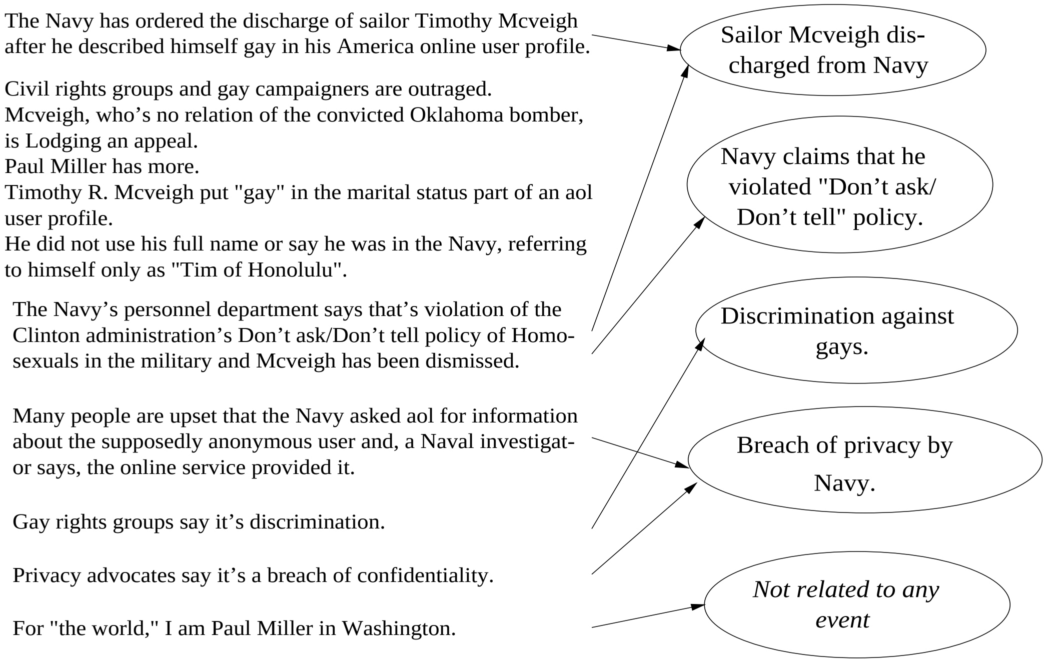

The Navy has ordered the discharge of sailor Timothy Mcveigh after he described himself gay in his America online user profile.
Civil rights groups and gay campaigners are outraged.
Mcveigh, who's no relation of the convicted Oklahoma bomber, is Lodging an appeal.
Paul Miller has more.
Timothy R. Mcveigh put "gay" in the marital status part of an aol user profile.
He did not use his full name or say he was in the Navy, referring to himself only as "Tim of Honolulu".
The Navy's personnel department says that's violation of the Clinton administration's Don't ask/Don't tell policy of Homosexuals in the military and Mcveigh has been dismissed.
Many people are upset that the Navy asked aol for information about the supposedly anonymous user and, a Naval investigator says, the online service provided it.
Gay rights groups say it's discrimination.
Privacy advocates say it's a breach of confidentiality.
For "the world," I am Paul Miller in Washington.
Sailor Mcveigh discharged from Navy
Navy claims that he violated "Don't ask/ Don't tell" policy.
Discrimination against gays.
Breach of privacy by Navy.
Not related to any event

Figure 1: An example showing how sentence extracts can indicate events

- It should be possible to quantify the precision of a summary, i.e., it should be possible to find the proportion of relevant sentences in the summary,
- It should be possible to identify redundancy in the system output being evaluated. There should be some way of assigning a marginal utility to sentences containing relevant but redundant information
- It should be possible to quantify the "usefulness" of a summary taking recall, precision as well as redundancy into account.
- Sentence boundaries should be uniquely identified (though they need not be perfect) because the aim of the system is to identify the relevant portions in the summary.

3. BUILDING AN EVALUATION CORPUS

3.1 The annotation process

We collect a stream of stories related to a certain topic from the TDT-2 corpus of stories from January 1 to June 30 1998. We used stories that were judged "on-topic" by annotators from LDC. The topics were selected from the TDT 1998 and 1999 evaluations. The stories are parsed to obtain sentence boundaries and all the sentences are given unique identifiers. We proceed with collecting the human judgments in the following four steps.

1. Each judge reads all the stories and identifies the important events.
2. The judges sit together to merge the events identified by them, to form a single list of events for that topic. All the events are given unique identifiers.
3. Each judge goes through the stories again, connecting sentences to the relevant events. Obviously, not all sentences need to be related to any event. However, if some sentence is relevant to more than one event, it is linked to all those events.
4. Another judge now verifies the mapping between the sentences and the events. This gives us the final mapping from sentences to events.

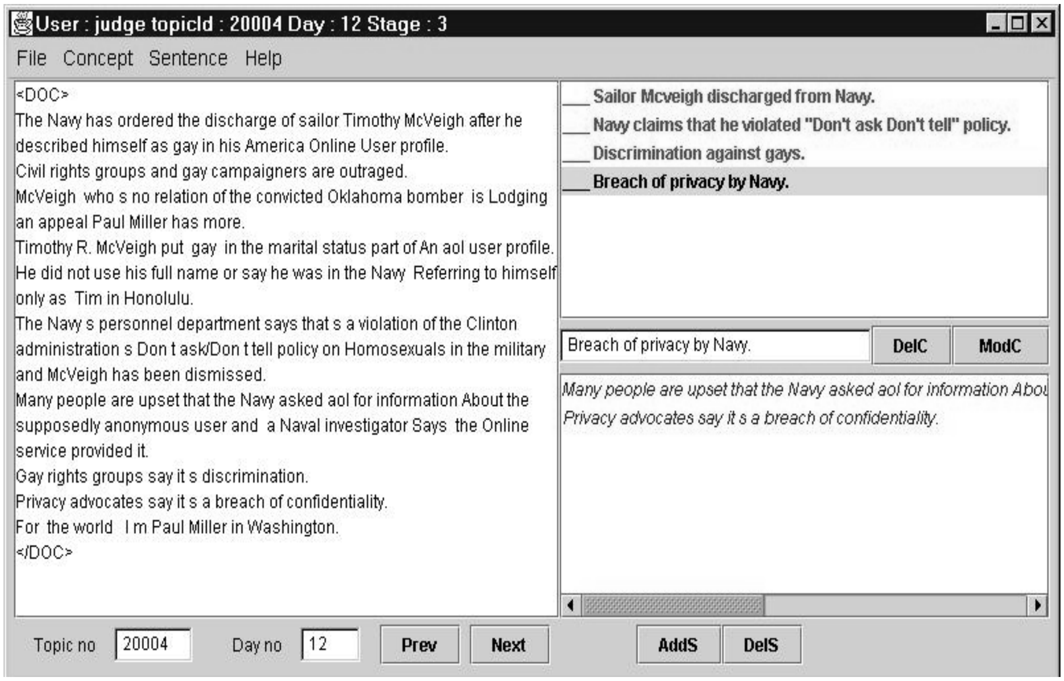
This way we obtain all the events mentioned within a story and we can also find out the events which find their first mention within this story. The advantage of building the evaluation corpus in this way is that these judgments can be used both for summarizing topic shifts as well as summarizing any given story by itself.

We have built a user interface in Java to allow judges to do the above work systematically. Figure 2 shows a snapshot of the interface used by the judges.

3.2 Statistics of the judgments obtained

We have obtained judgments for 22 topics. Three judges worked on each topic. We summarize the results of the annotation process for a subset of the topics in Table 1. We define the interjudge agreement for an event to be the ratio of the number of sentences linked to that event, as agreed upon by the third judge, to the number of sentences in the union of the sentences individually marked by the first two judges for that event. For a topic, the interjudge agreement is defined to be the average of the agreement for all the events in that topic. It is to be noted that the Kappa statistic is not applicable here in any standard form.

We found a large variance in the number of sentences linked to different events. As an example, in Table 2, we show the statistics for a group of news stories describing



<DOC>
The Navy has ordered the discharge of sailor Timothy McVeigh after he described himself as gay in his America Online User profile.
Civil rights groups and gay campaigners are outraged.
McVeigh who s no relation of the convicted Oklahoma bomber is Lodging an appeal Paul Miller has more.
Timothy R. McVeigh put gay in the marital status part of An aol user profile.
He did not use his full name or say he was in the Navy Referring to himself only as Tim in Honolulu.
The Navy s personnel department says that s a violation of the Clinton administration s Don t ask/Don t tell policy on Homosexuals in the military and McVeigh has been dismissed.
Many people are upset that the Navy asked aol for information About the supposedly anonymous user and a Naval investigator Says the Online service provided it.
Gay rights groups say it s discrimination.
Privacy advocates say it s a breach of confidentiality.
For the world I m Paul Miller in Washington.
</DOC>

___ Sailor Mcveigh discharged from Navy.
___ Navy claims that he violated "Don't ask Don't tell" policy.
___ Discrimination against gays.
___ Breach of privacy by Navy.

Breach of privacy by Navy.

Many people are upset that the Navy asked aol for information Abou
Privacy advocates say it s a breach of confidentiality.

20004

12

Figure 2: A snapshot of the user interface used for annotating the topics

Topic id	# of stories	# of events	Time taken (in hours)	Inter-judge Agreement
20008	49	10	4.5	0.91
20020	34	23	4.5	0.98
20021	48	9	2.5	0.97
20022	27	10	3.5	0.85
20024	38	12	2.75	0.98
20026	68	11	2.5	0.87
20031	34	15	2.5	0.62
20041	24	11	2	0.94
20042	28	14	2.5	0.66
20057	19	9	2	0.66
20065	57	16	2.33	0.94
20074	51	13	3	0.96
Average	39.75	12.75	2.88	0.86

Table 1: Annotation statistics for some of the topics

Event id	# of sentences	Inter-judge Agreement
1	43	1.0
2	9	1.0
3	33	0.97
4	8	1.0
5	4	0.8
6	5	1.0
7	14	1.0
8	19	1.0
9	9	1.0

Table 2: Variance in the number of sentences linked to different events for topic 20021

the damage due to tornados in Florida. We see that event 5 ("Relief agencies needed more than $300,000 to provide relief") is linked to 4 sentences while event 1 ("At least 40 people died in Florida due to 10-15 tornados.") is linked to 43 sentences. We may be able to use the number of sentences linked to a event as an indicator of the weight/importance of the event.

We have divided our corpus into two parts - one each for training and testing respectively. Each part consists of 11 topics. Care was taken to ensure that both the parts had topics of roughly the same size and time of occurrence. The statistics of both parts of the corpus are given in Table 3.

3.3 Problems faced

- Sometimes our sentence parser broke up a valid sentence into multiple parts. One judge linked only the relevant part of the sentence to the corresponding event while another linked all the parts to that event. This happened in the case of three of the topics (topics 20031, 20042 and 20057) before we detected the problem.
- Sometimes when similar sentences occur in different stories, one of the judges neglected the later occurrences of the sentence.

3.4 Interesting issues/judges' comments

We asked the judges for feedback on the annotation process and the difficulties faced. Here are some of the interesting issues which cropped up :

- Some ideas/events cannot be covered by any single sentence but only by a group of sentences. By themselves, none of the sentences might be relevant to the event. For example, Suppose, the event is *The Navy and AOL contradict each other* and we have two sentences - *"the navy has said in sworn testimony that*

	Training	Test	All
Number of topics	11	11	22
Number of stories	474	470	944
per topic	43.1	42.7	42.9
Number of events	162	181	343
per topic	14.7	16.5	15.6
Number of sentences	8043	9006	17049
per topic	731.2	818.7	775.0
per story	17.0	19.2	18.1
Off-event sentences	72%	70%	71%
Single-event sentences	24%	26%	25%
Multi-event sentences	4%	4%	4%

Table 3: Characteristics of the corpus. All numbers except for the number of topics are averaged over all topics included in that column.

this did happen." and *"america online is saying this never happened."* Clearly, any one sentence does not adequately represent the event. This can be easily taken care of by considering groups of sentences rather than single sentences.

- Abstract ideas : Sometimes the meaning of individual sentences is totally different from overall idea they convey. Satirical articles are an example of this. These kind of ideas cannot be represented by sentence extracts. We omitted such events.
- Sometimes different stories totally contradict each other. For example, some stories (on the same day) claim a lead for Bush while others claim Gore to be far ahead. This is more of a summarization issue though and need not be dealt with while building the evaluation corpus.

4. USING THE EVALUATION CORPUS

We have used the corpus for evaluating our system which produces temporal summaries in news stream ([1]). The problem of temporal summarization can be formalized as follows. A news topic is made up of a set of events and is discussed in a sequence of news stories. Most sentences of the news stories discuss one or more of the events in the topic. Some sentences are not germane to any of the events. Those sentences are called *"off-event"* sentences and contrast with "on-event" sentences. The task of the system is to assign a score to every sentence that indicates the importance of the sentence in the summary. This scoring yields a ranking on all sentences in the topic, including off- and on-event sentences.

We will use measures that are analogues of recall and precision. We are interested in multiple properties:

- *Useful* sentences are those that have the potential to be a meaningful part of the summary. Off-event sentences are not useful, but all other sentences are.
- *Novel* sentences are those that are not redundant—i.e., are new in the presentation. The first sentence about an event is clearly novel, but all following sentences discussing the same event are not.

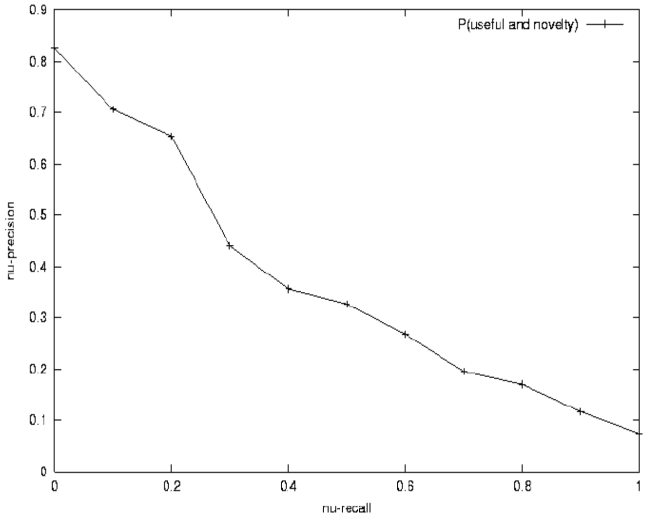


Figure 3: nu-recall vs nu-precision plot for the task of summarizing topic shifts in a news stream

- *Size* of the summary is a typical measure used in summarization research and we include it here.

Based on those properties, we could define the following measure to capture the combination of usefulness and novelty:

$$\begin{aligned} nu - recall &= \frac{\sum I(r(e_i) > 0)}{E} \\ nu - precision &= \frac{\sum I(r(e_i) > 0)}{S_r} \end{aligned}$$

where S_r is the number of sentences retrieved, E is the number of events in the topic, e_i is event number i ($1 \leq i \leq E$), $r(e_i)$ is the number of sentences retrieved for event e_i, $I(exp)$ is 1 if *exp* is true and 0 if not. All summations are as i ranges over the set of events. Note that $S_r \neq \sum r(e_i)$ since completely off-topic sentences might be retrieved.

The nu-recall measure is the proportion of the events that have been mentioned in the summary, and nu-precision is the proportion of sentences retrieved that are the first mentions of an event.

We used this measure to evaluate the performance of our system over the entire training corpus. The results for the training corpus are shown in the nu-recall/nu-precision graph in figure 3. This work is described in detail elsewhere([1]).

This is just one of the possible ways of using the corpus. We can define a number of other similar measures which could be easily computed using the data provided by such a corpus. These same measures can also be used to evaluate a system producing single or multi-document summaries too.

5. FUTURE WORK

We intend to complete collecting user judgments for more topics soon. After analyzing the reliablity of these judgments and correcting the few mistakes that we had made initially, we will collect annotations for more topics. Initially, we had used a simple barebones sentence parser, since that is mostly sufficient for the work such a corpus would be put to. Nevertheless, in future annotations, we will need to improve the sentence parser. We intend to continue using these judgments to evaluate the performance of the systems that we are currently building to identify and summarize topic shifts in news streams.

Acknowledgements

This material is based on work supported in part by the Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623 and in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal Summaries of News Topics. *Proceedings of SIGIR 2001 Conference, New Orleans, LA.*, 2001.

[2] J. Allan, V. Lavrenko, D. Frey, and V. Khandelwal. UMASS at TDT2000. *TDT 2000 Workshop notebook*, 2000.

[3] J. Allan, R. Papka, and V. Lavrenko. On-line New Event Detection and Tracking. *Proceedings of SIGIR 1998, pp. 37-45*, 1998.

[4] J. Allan et al. Topic-based novelty detection. *1999 Summer Workshop at CLSP Final Report. Available at http://www.clsp.jhu.edu/ws99/tdt*, 1999.

[5] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence Selection and Evaluation Metrics. *Proceedings of SIGIR 1999*, August 1999.

[6] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization Evaluation Methods: Experiments and Analysis. *Working notes, AAAI Spring Symposium on Intelligent Text Summarization, Stanford, CA*, April, 1998.

[7] Inderjeet Mani and et al. The TIPSTER SUMMAC Text Summarization Evaluation Final Report. 1998.

[8] R. Papka, J. Allan, and V. Lavrenko. UMASS Approaches to Detection and Tracking at TDT2. *Proceedings of the DARPA Broadcast News Workshop, Herndon,VA, pp. 111-125*, 1999.

[9] D. R. Radev, H. Jing, and M. Budzikowska. Summarization of multiple documents: clustering, sentence extraction, and evaluation. *ANLP/NAACL Workshop on Summarization, Seattle, WA*, 2000.

[10] R. Swan and J. Allan. Extracting Significant Time Varying Features from Text. *Proceedings of the Eighth International Conference on Information and Knowledge Management, pp.38-45*, 1999.

[11] R. Swan and J. Allan. Automatic Generation of Overview Timelines. *Proceedings of SIGIR 2000 Conference, Athens, pp.49-56*, 2000.