Temporal Summaries of News Topics

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

ABSTRACT

We discuss technology to help a person monitor changes in news coverage over time. We define temporal summaries of news stories as extracting a single sentence from each event within a news topic, where the stories are presented one at a time and sentences from a story must be ranked before the next story can be considered. We explain a method for evaluation, and describe an evaluation corpus that we have built. We also propose several methods for constructing temporal summaries and evaluate their effectiveness in comparison to degenerate cases. We show that simple approaches are effective, but that the problem is far from solved.

Keywords

Summarization, Experimental Design and Metrics

1. INTRODUCTION

We are interested in methods that help a person monitor changes in news coverage over time. We assume that a user has access to a stream of news stories that are on the same topic, but that the stream flows rapidly enough that no one has the time to look at every story. In this situation, a person would prefer to be kept up-to-date on the events within the topic, and to dive into the details only when the reported events trigger enough interest.

The usage model that we envision requires that the technology produce a revised summary at regular time intervals—e.g., every hour or at the start of each day. It is neither possible nor meaningful to wait until the topic is done to produce a summary. Nor does it make sense to produce an up-to-date overall summary at every time interval: the summary must indicate only what has changed. After all, the user has already been informed about everything that happened earlier.

For this study, we assume that news topics can be broken into a sequence of events, and that it is those events that interest users. In that case, it is sufficient for a system to produce a list of the events within the topic in the order those events are reported. Better summaries will list the events in the order reported, will capture all of the events, will avoid listing off-topic material as if it were an event, will avoid repeated mention of events, and will provide clear descriptions of events.

In the remainder of this study, we describe our efforts toward generating and evaluating temporal summaries that meet many of those criteria. We start by reviewing some related work on summarization and topic threading. In Section 3 we formalize the problem sufficiently that it can be evaluated. Section 4 proposes several methods for addressing the problem and presents an evaluation of those methods. We conclude in Section 5 and discuss possible directions for this research.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.

2. RELATED WORK

This research has its roots in text summarization, topic detection and tracking, and time-based summarization techniques. The following sections discuss work that is related to that discussed in this paper.

2.1 Summarization

The core technique of this temporal summarization research is to summarize a body of texts by extracting sentences that have particular properties. This work falls into a long tradition of sentence extraction, starting in the late 1950's with H.P. Luhn's classic work [19] and continuing forward [27]. Such techniques consider the words in the sentences, look for cue words and phrases [10, 32], consider even more focused features such as sentence length and case of words [18], or compare patterns of relationships between sentences [37]. Most of these approaches use statistics from the corpus itself to decide on the importance of sentences, and some leverage existing training sets of summaries to learn the properties of a summary [18, 3, 4].

Summarization techniques leverage a wide range of Natural Language Processing (NLP) and discourse information. Some focus primarily on techniques that have been developed in Information Retrieval [14], while most try to leverage both IR approaches and some aspects of NLP [16].

Of course, not all summaries are merely extracts. Some of the work already mentioned pieces together summaries from more than just sentences. Other work attempts to generate the summary directly, either from a knowledge-based representation of the content or from a statistical model of the text [39, 4].

Some summarization efforts have been focused on news stories or events. Maybury's work on event data [24] is different than this work because he was focused on events from simulations or application data rather than on events within news topics. Other work on news summarization, including work that uses the TDT corpora, focuses on single or multi-document summarization [25, 34, 12] of the stories, without attempting to capture the changes over time. Note that most multi-document summarization systems

have to include time as a component of their system to consolidate information across stories (e.g., to decide which statement is more up-to-date).

The use of Maximal Marginal Relevance (MMR) for summarization [6] is strongly related to the ideas in this paper. It shares the idea of balancing novelty and usefulness ("relevant novelty"), but focuses on query-based summarization of a static collection of stories.

This work is unlike most summarization research in its focus on summarizing changes over time. Comparative summaries of multiple documents [20] could conceivably address this problem, but we do not know of any that have.

2.2 Topic Detection and Tracking

This work also arises out of Topic Detection and Tracking (TDT), a body of research and an evaluation paradigm that addresses event-based organization of broadcast news. TDT investigation has been carried out over five years by about a dozen academic and industrial research institutions, and explored in the context of four "cooperatively competitive" evaluations sponsored by the U.S. government [1, 7, 28, 29]. The problems tackled by TDT are all story-based rather than sentence based. In many ways, the temporal summarization problem is an event- and sentence-level analogue of TDT's "first story detection" problem, where the task is to identify the first story that discusses each topic in the news.

2.3 Time-based Summarization

There has been very little work on time-based summarization to date. Some researchers have focused on how to extract temporal expressions from text, looking for and normalizing references to dates, times, and elapsed times [22]. That work is important for analyzing the content of the text, but does not directly address summarization itself.

In the summer of 1999, the Novelty Detection workshop at Johns Hopkins University's Center for Speech and Language Processing defined and explored new information detection (NID) [2]. The NID task was to identify the onset of new information within a topic by flagging the first sentence that contained it. The NID task is obviously very similar to the time-based summarization work proposed here. The summer workshop was unable to make significant progress because of problems with the definition of "new": when the team looked at an evaluation corpus they constructed, they discovered that 80% of the sentences were marked to contain new information. It turns out that almost every sentence in the news contains *some* new information—even if it is just the age of a person in the news. In this research, we have chosen a looser definition of "event" that makes this less of a problem.

This research is also related to work on automatic timeline construction [38]. That work focused on using the χ^2 measure to extract unusual words and phrases from a stream of news, and on grouping those features to isolate topics within the news. They suggested the idea of looking within the topics to create an event-level timeline, but have not yet done so. Further, since the timeline work is driven by graphical visualization, it will not take the same form as this text-driven approach.

2.4 Evaluation

Document summaries are difficult to evaluate, because for most applications there are numerous summaries that are of equally high quality. Simply rewording portions of the summary, reordering the sentences, omitting dubiously important information, etc., all result in minor variations that are still excellent summaries. Some types of evaluations that have been used are:

- Show several prototypical examples of a summarization technique's results.
- Ask a human to read the summaries and score their quality based upon some set of criteria [5, 31].
- Assume a single-document summary is the surrogate of the full document and see whether it can serve in the stead of the original. For example, is it possible to use a summary to categorize a document or to determine whether it is relevant to a query [11, 30]? Can a reader correctly answer a reading comprehension test using the summary [26]? Can a reader assign the correct keywords to a summary [36]?
- Compare agreement between sentences selected by humans and sentences selected by computer [35, 13], or compare agreement in the ranks of sentences that a system generates [9].

The first evaluation is not actually an evaluation, though it is useful for giving a sense of a summarizer's capabilities. The next set of evaluations all require that a human be part of a evaluation loop. This approach to evaluation is the most flexible, but is expensive, time-consuming, and non-repeatable. To evaluate a new summarization approach—even a minor adjustment on a previous method—requires an entirely new set of human judgments. In this work, we are focusing on evaluation methods like the last ones listed, that are based upon a fixed set of judgments and that can be repeated as often as necessary.

The initial core of our summarization approach is sentence extraction, so we can compare the sentences that a method chooses to the set of sentences that is known to be a good summary. To the extent that an approach chooses the "right" sentences, that approach is good; when it veers wildly from the ideal set, the approach is inappropriate to the task. Our approach is similar in spirit to the sentence-based evaluations listed above, but is modified significantly to take into account the time-based nature of our summaries.

3. FORMALIZING THE PROBLEM

We formalize the temporal summarization problem 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 (and are probably entirely off-topic). Those sentences are called "*off-event*" sentences and contrast with "on-event" sentences.

The order the stories are reported within the topic and the order of the sentences within each story combine to provide a total ordering on the sentences. We refer to that order as the *natural order*.

The task of a system is to assign a score to every sentence that indicates the importance of that sentence in the summary: higher scores reflect more important sentences. This scoring yields a ranking on all sentences in the topic, including off- and on-event sentences.

All sentences arriving in a specified time period can be considered together. They must each be assigned a score *before* the next set of sentences (from the next time period) is presented. For this work, we have used a time period that has one story arriving at a time.

A summary consists of all sentences scoring over some threshold θ. We will only consider the cutoff point in Section 4.5.

3.1 Measures

We will use measures that are analogues of recall and precision. Traditional Information Retrieval evaluation is concerned with finding relevant material, so the measures need consider only relevance. 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.
- *Size* of the summary is a typical measure used in summarization research and we include it here.

We define recall and precision based analogues for usefulness, novelty, and a combination of them. To do that, we assume that the entire set of sentences is broken into U, the useful sentences, and $\overline{U}$, the non-useful sentences. Let the set of v events, E be $\{e_1, ..., e_v\}$ and the set of sentences be $S = \{s_1, s_2, ...\}$. We let S_m stand for the subset of S consisting of $\{s_1, ..., s_m\}$. Finally, we let $C(X)$ represent the set of events (from E) that are mentioned in the set of sentences X.

All measures are taken after r sentences have been seen in the ranked list. $I(exp)$ is 1 if *exp* is true and 0 if not.

$$\text{u-recall} = \frac{|S_r \cap U|}{|U|}$$

$$\text{u-precision} = \frac{|S_r \cap U|}{|S_r|}$$

Intuitively, u-recall is the proportion of on-event (useful) sentences that have been retrieved, and u-precision is the proportion of retrieved sentences that are on-event for some event.

A sentence is novel if it covers one or more events that were not covered by any previous sentence. One way to evaluate novelty is:

$$\text{n-recall} = \frac{|C(S_r)|}{|E|}$$

$$\text{n-precision} = \frac{I(C(S_1) > 0) + \sum_{i=2}^{r} I(C(S_i) > C(S_{i-1}))}{|S_r \cap U|}$$

Here, n-recall is the proportion of events that have been covered so far. The first part of n-precision determines whether the top ranked sentence is novel; the summation makes the same decision for each following sentence.

One problem with n-recall and n-precision is that they do not take into account off-event ("useless") sentences: a system scores the same, no matter how the off-event sentences are ranked. We also measure n0-recall and n0-precision that treat all off-event sentences as if they were an extra event ("event 0"):

$$\text{n0-recall} = \frac{|C(S_r)| + I(S_r \neq S_r \cap U)}{|E| + 1}$$

$$\text{n0-precision} = \frac{I(C(S_1) > 0) + \sum_{i=2}^{r} I(C(S_i) > C(S_{i-1}))}{|S_r|} + \frac{I(S_r \neq S_R \cap U)}{|S_r|}$$

In this case, the n0-recall measure just includes an extra "event" and measures whether some non-useful sentence was retrieved. In parallel, n0-precision has to measure whether a non-useful sentence was retrieved, and its denominator no longer only counts useful retrieved sentences.

The final measure we use combines usefulness and novelty:

$$\text{nu-recall} = \frac{|C(S_r)|}{|E|} = \text{n-recall}$$

$$\text{nu-precision} = \frac{I(C(S_1) > 0) + \sum_{i=2}^{r} I(C(S_i) > C(S_{i-1}))}{|S_r|}$$

Note the small differences between nu-precision and the two precision measures for novelty (n-precision and n0-precision).

Just as with IR's recall and precision, those measures are set-based. To show the tradeoff between measures, we will plot the various recall and precision graphs over the entire ranked list. To average across multiple topics, the graphs will be interpolated to the standard eleven recall points (0.0, 0.1, ..., 1.0). We will also provide the exact average precision (i.e., the average of precision values at every point that recall increases). These graphs and single-number measures are analogous to those used in traditional IR evaluation.

3.2 Delay factors

How often should the system generate sentence scores? The extremes are after every sentence, and when all sentences have been seen. The evaluation measures themselves do not distinguish on this parameter: all they require is that every sentence have a score so that all sentences can be ranked.

The preferred model is that a system operates on a story basis. When a story arrives, its sentences are scored and those scores are emitted. Only then can the next story be processed.

Some of the approaches that we discuss in Section 4 process each sentence as it arrives, not even considering the rest of the story. We have also done experiments that "cheat" by waiting until all stories in a topic have been presented, but they are not reported here. Not surprisingly, those approaches work somewhat better than the more realistic delay factors.

3.3 Evaluation corpus

Our evaluation corpus is built from the TDT-2 corpus [8] of approximately 60,000 news stories covering January through June of 1998. As part of the TDT research program, about 200 news topics were identified in that period, and all stories were marked as on- or off-topic for every one of the topics. The topics have been used in several evaluations and the story–topic assignments have been confirmed by quality assurance cycles.

We selected 22 medium-sized topics from the set of 200. For each topic, two annotators independently read all on-topic stories and decided on a list of events within the topic. The annotators then worked together to decide on a common list. They then performed a second pass through the on-topic stories and assigned each sentence to zero, one, or more events. The topics were broken into 11 training and 11 test topics for this study. Table 1 lists some statistical information about the topics, events, and sentences, and shows that the training and test sets are comparable. The process of creating the evaluation corpus is described in more detail elsewhere [17].

It is interesting to note that this approach to annotating a corpus has already created a substantially different problem than that explored in the NID summer workshop [2]. Where 80% of their sentences contained new information, only 30% of ours are marked as important for some event. A perfect NID system would result in 20% reduction of the text; a summary that extracted just useful sentences would cause a 70% drop in the amount of presented text.

Table 1: Characteristics of the temporal summarization evaluation corpus used. All numbers except for the number of topics are averaged over all topics included in that column.

	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%

We used the training topics during our experimentation to select the best approaches and to do parameter fitting where needed. The remaining 11 test topics were never looked at except to present final information for this paper.

4. SOLUTIONS AND EXPERIMENTS

The evaluation measures described above capture two desirable characteristics of extractive temporal summaries: selecting useful sentences, and preferring novel (non-redundant) sentences. Our overall evaluation measures are the *nu-* measures above that combine both factors. For now, we will assume that the factors are independent, that is:

$$P(\text{useful} \wedge \text{novel}) = P(\text{useful}) \cdot P(\text{novel})$$

We will propose several alternatives for estimating each component, evaluate them separately, and then evaluate their combination.

All of the solutions that we propose here are based on "language model" representations of news topics and events [33]. Specifically, given some amount of text on a particular topic, we estimate a probabilistic model of how text from the topic is likely to be generated. Using that model, we can determine the probability that a new piece of text (sentence, story) could have been generated by the model.

For example, suppose we are given a set of stories that are on the same news topic. One way to estimate the probability that a word would appear in that topic would be,

$$P(w) = \frac{\sum_i tf(w, S_i)}{\sum_i |S_i|}$$

where $tf(w, S_i)$ represents the number of times word w occurs in story S_i. That is, a word's probability of occurrence can be estimated by the proportion of the time that it has already occurred. We make the usual assumption that word occurrences are independent, so the probability of a run of text is the product of the probability of its words. This estimator is usually smoothed using some variant of LaPlace's Law [23]. In our case, we add 0.01 to the numerator and multiply the denominator by 1.01.

4.1 Baselines

In all of the results below, we include the following baseline systems. These are provided to demonstrate that it is not better to do nothing.

- *Random* assigns a random number from 0.0 to 1.0 to each of the sentences. The process is repeated 10 times and the resulting evaluations are averaged at each standard recall point.
- *Natural order* is the baseline that assumes that sentences are ranked in the order they appear. The ranking starts with all sentences of the first story (in order), then the second story, and so on.
- *Round robin* assigns the highest score to the first sentences of stories, then the second sentence, and so on. This could be generated by assigning a score of

  $$(1000 - \text{sentence\_number}) \cdot 1000 + \text{story\_number}$$

  for example (adjusted if more than 1,000 stories are expected).
- *Worst case* is the worst possible ordering of sentences for a particular evaluation measure. For example, the worst possible measure for u-recall and u-precision would have all off-event sentences ranked first. The worst possible ordering for a novelty measure would have all sentences from one event ranked before any sentence of the next. The worst possible nu-recall and nu-precision ranking combines that by placing off-event sentences first and then grouping all stories from one event followed by all of the next, and so on.

4.2 Capturing Usefulness

Usefulness represents whether or not a sentence discusses one of the events of the topic. Sentences that are off-topic are clearly not related to any of the events. To consider whether some sentence s_k is on-topic (useful), we want to know whether it could be generated by a model created from the topic, represented by every sentence seen to date. If $\text{LM}(x)$ is used to denote the language model created from text x, then we have:

$$\begin{aligned} P(\text{useful}_1) &= P(s_k|\text{LM}(s_1, \ldots, s_{k-1})) \\ &= \left(\prod_{w \in s_k} \frac{tf(w, s_1 + \cdots + s_{k-1}) + 0.01}{1.01 \cdot \sum_i |s_i|} \right)^{\frac{1}{|s_k|}} \end{aligned}$$

The $|s_k|^{th}$ root provides length normalization so that sentences of all lengths are treated equally. Intuitively, all prior sentences are used to estimate the likelihood that a word will appear in the topic. The probability of a sentence is the probability that every word appears. We make the typical independence assumptions.

An alternate model of "useful" comes from the observation that news stories are usually predominantly about the topic in question, so that sentences that are very like their news story are more likely to be useful. If S is the story that s_k comes from, then:

$$\begin{aligned} P(\text{useful}_2) &= P(s_k|\text{LM}(S)), s_k \in S \\ &= \left(\prod_{w \in s_k} \frac{tf(w, S) + 0.01}{1.01 \cdot |S|} \right)^{\frac{1}{|s_k|}} \end{aligned}$$

Intuitively, this builds a model of the story's topic using all sentences in the story. The probability that a sentence is on-topic is then calculated from the probability that each word is part of the topic.

Figure 1 shows the effectiveness of those approaches, comparing u-precision and u-recall. The graph includes the baselines, where the theoretical worst-case performance is generated by ranking all off-event sentences first. Results for the training topics are shown in the left graph.

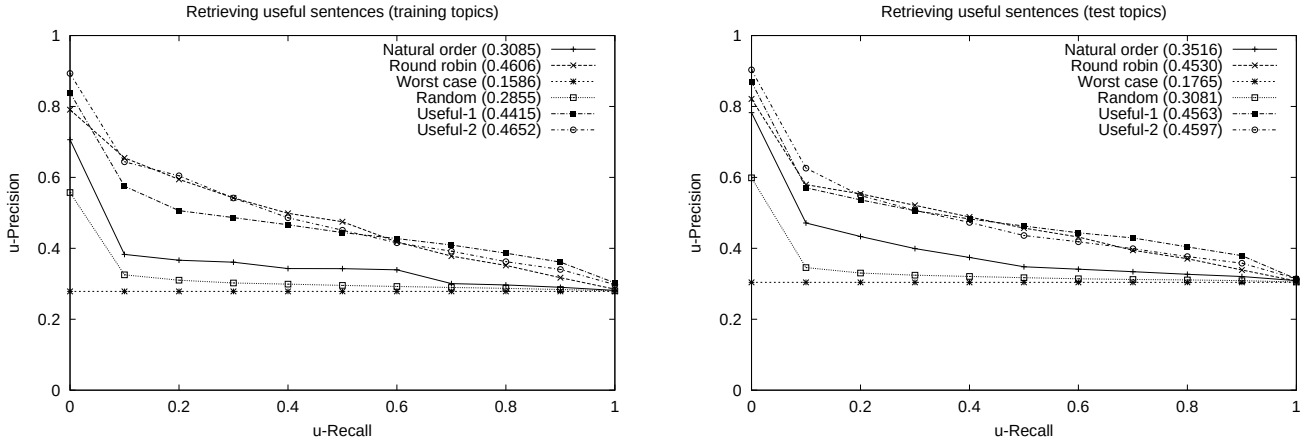

Figure 1: Shows tradeoff between measures of usefulness for various approaches. The numbers in the legend represent exact average u-precision for that approach. The left graph is the training topics; the right graph is the test topics.

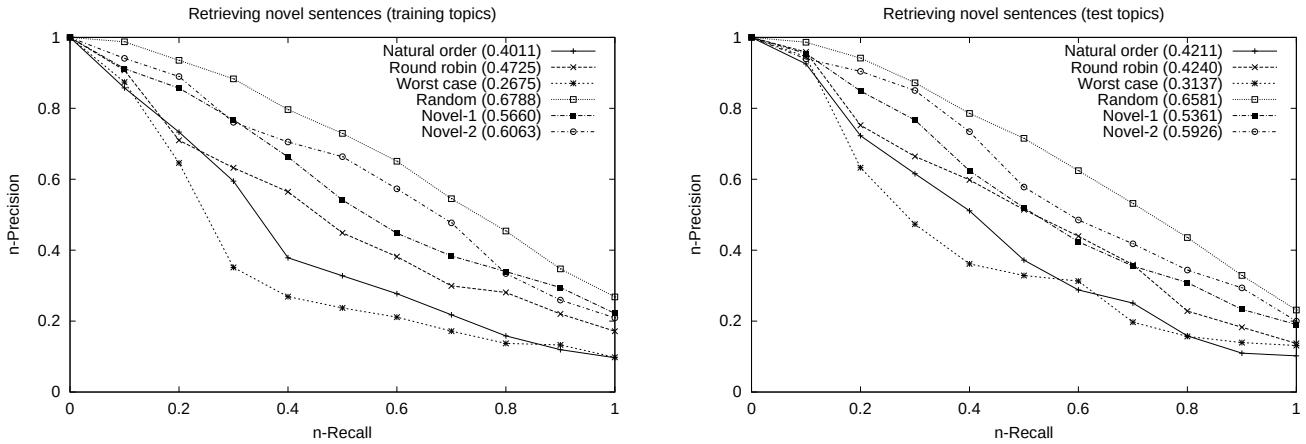

Figure 2: Shows tradeoff between measures of novelty for various approaches. The numbers in the legend represent exact average n-precision for that approach. Training topics on the left; test on the right.

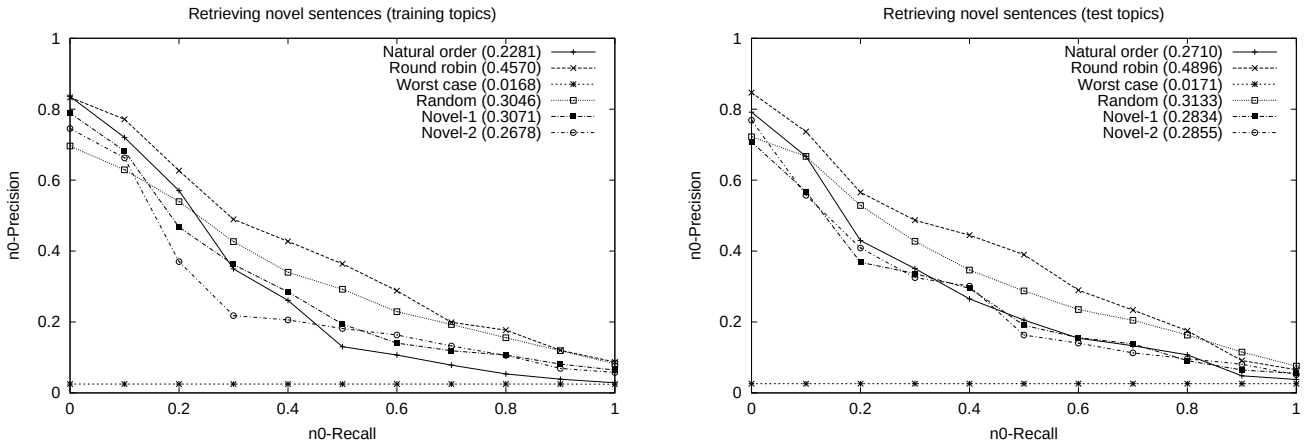

Figure 3: Shows tradeoff between measures of novelty (with off-event sentences counted as a separate topic) for various approaches. The numbers in the legend represent exact average n0-precision for that approach. Training topics on the left; test on the right.

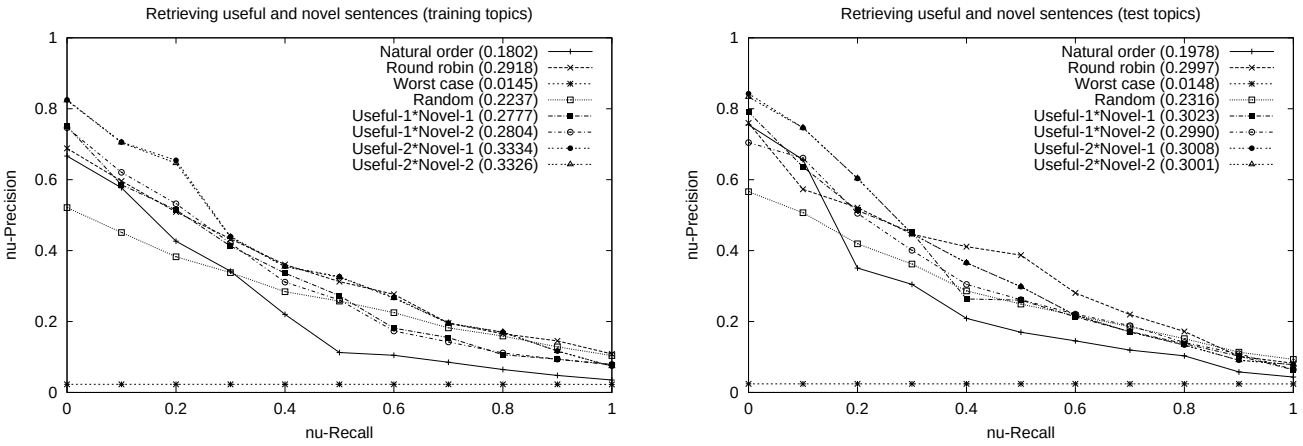

Figure 4: Shows tradeoff between measures of summarization quality based on a combination of usefulness and novelty, for various approaches. The numbers in the legend represent exact average nu-precision for that approach. Training topics are on the right; test on the right.

Of the two usefulness measures, useful$_2$ substantially outperforms useful$_1$. The difference reflects the small amount of information that is available in a single sentence. When statistics for those words are smoothed as in useful$_2$, the estimates are superior.

The surprising result for usefulness is that a round robin ranking algorithm performs almost as well as useful$_2$. We believe that reflects the pyramid nature of news reporting: important, and therefore probably on-topic, information is reported early in a story. Later material is more likely to be tangentially related to the topic, and so ranking it lower helps.

The right half of Figure 1 shows identical approaches applied to the 11 test topics. The results are comparable, except that the difference between useful$_1$ and useful$_2$ essentially disappears. The latter is still better for high-ranking sentences, but it is eventually dominated by the former. Overall there is no substantial difference between the two and round robin, but all three outperform the other baselines.

4.3 Capturing Novelty

The second characteristics of sentence selection is *novelty*. The second or third sentence about an event is less interesting than the first. To capture that property we assume that every sentence is associated with an event. When a new sentence arrives, we compare "its" event to that of all prior sentences. If it is unlike all of those events, then the new sentence is novel and should receive a high score. If $e(s_i)$ represents the event discussed by sentence s_i, then:

$$\begin{aligned} P(\text{novel}_1) &= P(e(s_k) \neq e(s_i), \forall i < k) \\ &= \left[\prod_{i<k} (1 - P(e(s_k) = e(s_i)))\right]^{\frac{1}{k-1}} \\ &= \left[\prod_{i<k} (1 - P(s_k|\text{LM}(s_i)))\right]^{\frac{1}{k-1}} \\ &= \left[\prod_{i<k} \left(1 - \left[\prod_{w \in s_k} \frac{tf(w, s_i) + 0.01}{1.01 \cdot |s_i|}\right]^{\frac{1}{|s_k|}}\right)\right]^{\frac{1}{k-1}} \end{aligned}$$

Here we are modeling the probability that two sentences discuss the same event by the probability that the later sentence could arise from the same language model as the earlier sentence. Here the model is derived from a single sentence so is probably unreliable.

That problem of sparse data to estimate the probability suggests that it might be helpful to group sentences together. For that reason, we also tried a method that clusters sentences together if they appear to be discussing the same event. Whereas in the previous approach each sentence was used to model an event, here we group sentences together and use them to model the event. If we assume that when sentence s_k arrives there are m event clusters, c_1 through c_m :

$$\begin{aligned} P(\text{novel}_2) &= P(e(s_k) \neq e(c_i), \forall i \leq m) \\ &= \left[\prod_{i \leq m} (1 - P(e(s_k) = e(c_i)))\right]^{\frac{1}{m}} \\ &= \left[\prod_{i \leq mk} (1 - P(s_k|\text{LM}(c_i)))\right]^{\frac{1}{m}} \\ &= \left[\prod_{i \leq m} \left(1 - \left[\prod_{w \in s_k} \frac{tf(w, c_i) + 0.01}{1.01 \cdot |c_i|}\right]^{\frac{1}{|s_k|}}\right)\right]^{\frac{1}{m}} \end{aligned}$$

This novel$_2$ approach is the same as novel$_1$ except that the sentence is compared to clusters and there is more information in a cluster to estimate probabilities. Note that this approach also requires a threshold for deciding whether or not a sentence should be added to a cluster. We used the training topics to find a good parameter setting, though we found that it was not very sensitive to the value chosen.

Both of these approaches may bring non-useful sentences to the top of ranking since such sentences will seem novel. The n-recall and n-precision measures take that into account by completely ignoring the ranking of off-event sentences. This choice allows us to measure the effectiveness of a novelty system without worrying about usefulness issues. We intend that our final measures—combining novelty *and* usefulness—will provide a balance between the two.

Figure 2 shows the effectiveness of this approach compared to the baselines. Worst case performance includes all sentences from

the first event, then all from the second, and so on. For this measure of effectiveness, both approaches substantially improve on the baseline cases. The novel_2 measure is also a clear improvement on novel_1, suggesting that clustering is useful for modeling the events.

Unfortunately, and initially somewhat surprisingly, the best performance by far on this problem is achieved by the "degenerate" random case. This result turns out to be almost obvious in hindsight. Because about 70% of the sentences are off-event, how they are ranked is not measured. The best way to select from the remaining sentences and be fairly confident of selecting from each event, is to *randomly* choose from them. We would expect almost perfect performance (on averaeg) from that measure if the stories were evenly distributed between the events. The imperfect performance reflects the uneven distribution of stories.

We used the n0-precision and n0-recall measures in an effort to confirm our suspicions. In Figure 3 the random baseline performance drops dramatically, because there is now a pseudo-event that includes the huge number of off-event sentences. The distribution between "events" is now wildly uneven and random samplying is a mistake. Unfortunately, the novel_1 and novel_2 measures suffer similarly, and the round robin baseline becomes the best performer.

It is clear from these results that our models of novelty are inadequate for the task and that there is substantial work that needs to be done to capture newness.

The test results on the right sides of Figures 2 and 3 are consistent with the results on the training data. That stability is encouraging and also shows that the clustering threshold may be stable across topics.

4.4 Useful novelty

In this section of our experiments we combine novelty and usefulness into a single measure of "interestingness." We choose the best measure of usefulness and the best measure of novelty and multiply their probabilities together:

$$P(\text{interesting}) = P(\text{useful}) \cdot P(\text{novel})$$

It is unlikely that the two factors are truly independent. However, we have been able to improve one without affecting the other, so they are at least not strongly related.

Figure 4 shows the effectiveness of this approach, measured by the *nu-* measures that reflect a system's ability to rank useful and novel sentences highest. We have shown the combination of both usefulness measures with each of the novelty measures. We expected that useful_2 combined with novel_2 would perform best, and were surprised to see no difference between that and a combination of useful_2 with novel_1. Figure 2 showed a clear advantage to novel_2, so it is odd that the choice of novelty measure has no impact. We believe this result is because of the poor formance that the novelty measures exhibit (in Section 4.3, and that improving that performance will help the combination.

It is quite possible that novelty and usefulness are not independent, and that other ways of combining them would be better. We tried to crudely represent an "or" of the components using a range of possible linear combinations of them:

$$\alpha \cdot P(\text{useful}) + (1 - \alpha) \cdot P(\text{novel})$$

Figure 5 shows the exact average nu-precision for a range of α values. The graph shows that no value improves over the product (exact nu-precision of 0.3334). The highest value seen with the linear combination is 0.3039.

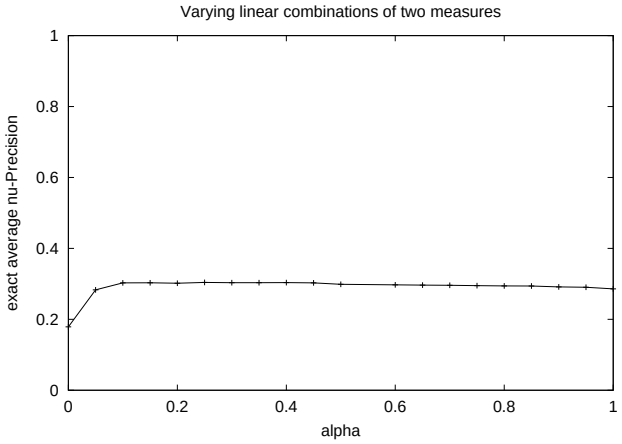

Figure 5: Depicts exact average nu-precision when a sentence's score is calculated using a linear combination of usefulness and novelty. These results are on the training data.

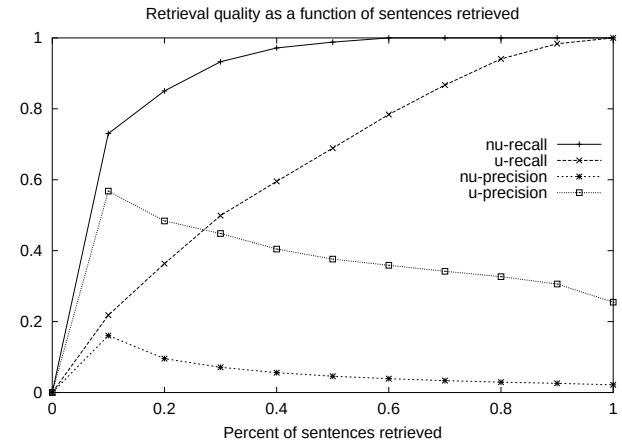

Figure 6: Shows the effect on various measures as the proportion of sentences retrieved is increased. These results are on the training data.

4.5 Summary size

A traditional measure of summary quality takes its size into account. In order to explore that question, we used the score determined by,

$$P(\text{useful}) \cdot P(\text{novel})$$

and plotted its effect using various measures compared to the amount of text retrieved. Figure 6 shows this effect for four measures. The x-axis represents the proportion of the sentences retrieved. At 20%, for example, 1600 out of 8000 sentences would have been selected (based on some threshold θ that is implicit).

The two consistently upward sloping curves show recall measures. The u-recall line is constantly increasing showing that more and more on-topic (useful) sentences are being retrieved. The other line, nu-recall moves up sharply, reflecting the approach's ability to find at least one sentence per event very quickly: almost 80% of the 162 events are represented when 10% (800) sentences have been seen.

The other two lines show the impact on precision. The upper line is u-precision and shows that about half of the sentences are on-topic at 10-20% of the full text. The nu-precision measure shows that there is still substantial redundancy, since at that same range, only 10% of the sentences are the first sentence discussing an event.

Based on these measures, this summarization technique appears to do best if it retrieves about 10% of the original sentences.

5. SUMMARY AND FUTURE WORK

We have defined temporal summarization, a new and important variant of the text summarization task. We have described it in a way that it is possible to carry out laboratory evaluations of effectiveness, avoiding costly user evaluations at every step of the process. We have built and described an evaluation corpus based on 22 topics from TDT news stories.

In this study, we also described measures of summary effectiveness that are based upon the traditional recall and precision measures. We have described several techniques for generating high-quality summaries (by those measures) and have evaluated them with our corpus.

We have shown that simple probabilistic approaches for finding "useful" sentences do not outperform the baseline round robin case. However, there is still substantial room for improvement. We have, however, shown excellent approaches for finding novel sentences to avoid redundancy. We have also shown that there are unexplained connections between the two factors that make them difficult to combine successfully. Finally we showed that by the more interesting measures, our current approach is probably optimal if it retrieves about 10% of the sentences.

Our immediate future work on this project involves a continuing investigation into modeling "interesting" sentences for temporal summarization. The current estimators for probabilities are very crude, even though they sometimes work well. We will explore better estimators for the topic and event models, possibly using smoothing techniques based upon expansion as well as backoff and mixture models. We expect that named entity tagging and possibly temporal expression normalization [22] may help match events and topics.

A property of the evaluation measures that we have chosen is that they require that the sentence scores being normalized over time. That is, a score of 0.43 must have the same meaning at any point. Because our statistical measures have more data the further into the topic the system is, our scores are not stable. We are developing measures that evaluate on a story-by-story or a day-by-day basis, reducing the impact of that problem. This will allow us to separate the problem of finding useful and novel sentences from normalizing the final scores.

All of this work is exploratory in that it was done with a "clean" set of stories for each topic—that is, every story was *known* to discuss the topic. We felt this was an important and reasonable simplification of the problem to lay the groundwork. We are now looking at the impact of completely off-topic stories. We will do that by using the topic clusters generated by TDT systems.

Acknowledgments

Some preliminary related work on the topic of language models for summarization was done under the direction of the first author by Taren Stinebrickner-Kauffman and Andrés Santiago Pérez-Bergquist. They were participating in a summer Research Experience for Undergraduates (REU) program that was supported in part by the National Science Foundation (grant number EEC-9820309).

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 authors' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[2] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. L. vrenko, R. Hoberman, and D. Caputo. Topic-based novelty detection: 1999 summer workshop at CLSP, final re port. Available at http://www.clsp.jhu.edu/ws99/tdt, 1999.

[3] C. Aone, M. Okurowski, J. Gorlinsky, and B. Larsen. A trainable summarizer with knowledge acquired from robust nlp techniques. In Mani and Maybury [21].

[4] A. Berker and V. Mittal. OCELOT: a system for summarizing web pages. In *Proceedings of SIGIR*, pages 144–151, 2000.

[5] R. Brandow, K. Mitze, and L. Rau. Automatic condensation of electronic publications by sentence selection. In Mani and Maybury [21]. Originally published in Information Processing and Management.

[6] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producting summaries. In *Proceedings of SIGIR*, pages 335–336, 1998.

[7] DARPA, editor. *Proceedings of the DARPA Broadcast news Workshop*, Herndon, Virginia, Feb. 1999.

[8] DARPA, editor. *The TDT-2 Text and Speech Corpus*, Herndon, Virginia, Feb. 1999.

[9] R. Donaway, K. Drummey, and L. Mather. A comparison of rankings produced by summarization evaluation measures. In Hahn et al. [15], pages 69–78.

[10] H. Edmundson. New methods in automatic extracting. In Mani and Maybury [21]. Originally published in Journal of the ACM.

[11] T. Firmin and M. Chrzanowski. An evaluation of automatic text summarization systems. In Mani and Maybury [21].

[12] F. Fukumoto and Y. Suzuki. Extracting key paragraph based on topic and event detection. In Hahn et al. [15], pages 31–39.

[13] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of SIGIR*, pages 121–128, 1999.

[14] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In Hahn et al. [15], pages 40–48.

[15] U. Hahn, C. Lin, I. Mani, and D. Radev, editors. *Automatic Summarization: ANLP/NAACL 2000 Workshop*, New Brunswick, NJ, 2000. Association for Computational Linguistics.

[16] E. Hovy and C. Lin. Automated text summarization in SUMMARIST. In Mani and Maybury [21].

[17] V. Khandelwal, R. Gupta, and J. Allan. An evaluation scheme for summarizing topic shifts in news streams, 2001. Submitted to Human Language Technology Conference.

[18] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In Mani and Maybury [21]. Originally published in Proceedings of SIGIR.

[19] H. Luhn. The automatic creation of literature abstracts. In Mani and Maybury [21]. Originally published in IBM Journal of R&D.

[20] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. In Mani and Maybury [21]. Originally published in Information Retrieval.

[21] I. Mani and M. Maybury, editors. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, Massachusetts, 1999.

[22] I. Mani and G. Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 69–76, New Brunswick, New Jersey, 2000. Association for Computational Linguistics.

[23] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[24] M. Maybury. Generating summaries from event data. In Mani and Maybury [21]. Originally published in Information Processing and Management.

[25] K. McKeown and D. Radev. Generating summaries of multiple news articles. In Mani and Maybury [21].

[26] A. Morris, G. Kasper, and D. Adams. The effects and limitations of automated text condensing on reading comprehension preformance. In Mani and Maybury [21]. Originally published in Information Systems Research.

[27] S. Myaeng and D. Jang. Development and evaluation of a statistically based document summarization system. In Mani and Maybury [21].

[28] NIST. Proceedings of the tdt 1999 workshop. Notebook publication for participants only, Mar. 2000.

[29] NIST. Proceedings of the tdt 2000 workshop. Notebook publication for participants only, Nov. 2000.

[30] M. Oka and Y. Ueda. Evaluation of phrase-representation summarization based on information retrieval task. In Hahn et al. [15], pages 59–68.

[31] M. Okurowski, H. Wilson, J. Urbina, R. Clark, T. Taylor, and F. Krapcho. Text summarizer in use: Lessons learned from real world deployment and evaluation. In Hahn et al. [15], pages 49–58.

[32] J. Pollock and A. Zamora. Automatic abstracting research at chemical abstracts service. In Mani and Maybury [21]. Originally published in Journal of Chemical Information and Computer Science.

[33] J. Ponte and W. Croft. A language modeling approach to information retrieval. In *Proceedings of SIGIR*, pages 275–281, 1998.

[34] D. Radev, H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In Hahn et al. [15], pages 21–30.

[35] G. Rath, A. Resnick, and T. Savage. The formation of abstracts by the selection of sentences. In Mani and Maybury [21]. Originally published in American Documentation (now JASIS).

[36] H. Saggion and G. Lapalme. Concept identification and presentation in the context of technical text summarization. In Hahn et al. [15], pages 1–10.

[37] G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. In Mani and Maybury [21]. Originally published in Information Processing and Management.

[38] R. Swan and J. Allan. Automatic generation of overview timelines. In *Proceedings of SIGIR*, pages 49–56, Athens, Greece, 2000. ACM.

[39] M. Witbrock and V. Mittal. Ulta-summarization : A statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of SIGIR*, pages 315–316, 1999. Poster description.