Topic Models for Summarizing Novelty

James Allan, Rahul Gupta, and Vikas Khandelwal
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst

Abstract

We define temporal summaries of news stories as extracting as few sentences as possible 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 outline an evaluation strategy that we have developed for this task and describe simple language models for capturing novelty and usefulness in the context of summarization. We show that the simple approaches work moderately well, and outline our ideas for moving forward.

1 Introduction

We are interested in methods that help a person monitor changes in news coverage over time. We aim to do that by providing a streaming summary of the news topic, selecting sentences that describe the key events within the topic as they arrive. Additional sentences on each event and off-topic sentences should be suppressed. We call the resulting selection of sentences a "temporal summary."

To do this, we will model the topics that news stories discuss, as well as the events within those topics. Topic models will be useful for finding sentences that are useful (on-topic) and the event models will be used to determine whether the sentence talks about a previously unseen event within the topic.

The usage 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.

Our intent in this work is to use a language model-based approach for modeling topics and events, and for selecting sentences to include in the summary. Because evaluation is such a difficult problem in text summarization research, we have started by developing an evaluation framework. We sketch the main ideas behind that framework in Section 3; it is described in detail elsewhere [10, 2].

Section 4 presents the very simple modeling that we used to capture novelty and usefulness. These models are intended to represent baseline performance and to illustrate the evaluation framework. Our future work involves building and evaluating more accurate models.

2 Related work

In addition to language modeling, this research has its roots in text summarization, topic detection and tracking, and time-based summarization techniques.

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 [11] and continuing forward [16]. The use of Maximal Marginal Relevance (MMR) for summarization [4] 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 [12] could conceivably address this problem, but we do not know of any that have.

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 [1, 5, 17, 18]. 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.

There has been very little work on time-based summarization to date. 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) [3]. 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.

3 Evaluation

Document summaries are difficult to evaluate, because for most applications there are numerous summaries that are of equally high quality. In this work, we are focusing on evaluation methods that are based upon a fixed set of judgments and that can be repeated as often as necessary.

The 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 other sentence-based evaluations [20, 8, 7], but is modified significantly to take into account the time-based nature of our summaries.

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 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 of 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.

3.1 Evaluation measures

We will use measures that are analogs of recall and precision, but that capture not only usefulness (relevance), but also novelty.[1] For example, "useful recall" is the proportion of retrieved sentences that are useful (relevant) and "useful precision" is the proportion of retrieved sentences that are useful.

Extending that, "novel recall" is a measure of the proportion of retrieved sentences that are novel—i.e., discuss events that have not been seen before. Novelty is somewhat slippery because sentences can discuss multiple events. That means that whether sentences are novel depends on which sentences have already been seen. "Novel precision" has the same awkwardness.

The measures can also be combined to create something called "useful novelty" (or "novel usefulness") that measures how many sentences were *both* useful *and* novel.

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 Evaluation corpus

Our initial experiments have been done using the TDT-2 corpus [6] of approximately 60,000 news stories covering January through June of 1998. We selected 22 medium-sized topics from the set of 200 that are provided with the corpus. 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. 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 as part of the final evaluation [2] that is reported here. Additional details about constructing the evaluation corpus are provided elsewhere [10].

[1] Detailed descriptions of all measures can be found elsewhere [2].

4 Modeling topics and events

The goal of this research was to model the topics and events that sentences describe, and to look for the occurrence of new events (novelty) within the topic (usefulness). All of the solutions that we propose here are based on "language model" representations of news topics and events [19]. 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 maximum likelihood estimator is usually smoothed using some variant of LaPlace's Law [15]. In our case, we add 0.01 to the numerator and multiply the denominator by 1.01.

4.1 Topic models for 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 $\mathrm{LM}(x)$ is used to denote the language model created from text x, then we have:

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

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 each of its words 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:

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

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.

The left of Figure 1 shows the effectiveness of those two approaches on our 11 test topics, 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.

Of the two usefulness measures, useful_2 outperforms useful_1 at the low recall portion of the graph. This is likely because high quality sentences from an on-topic story are not "diluted" by the language of earlier stories. This technique is problematic because it clearly will fail if the stories themselves are not on-topic.

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.

Overall there is no substantial difference between the two usefulness models and round robin, but all three outperform the other baselines.

4.2 Event models for 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:

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

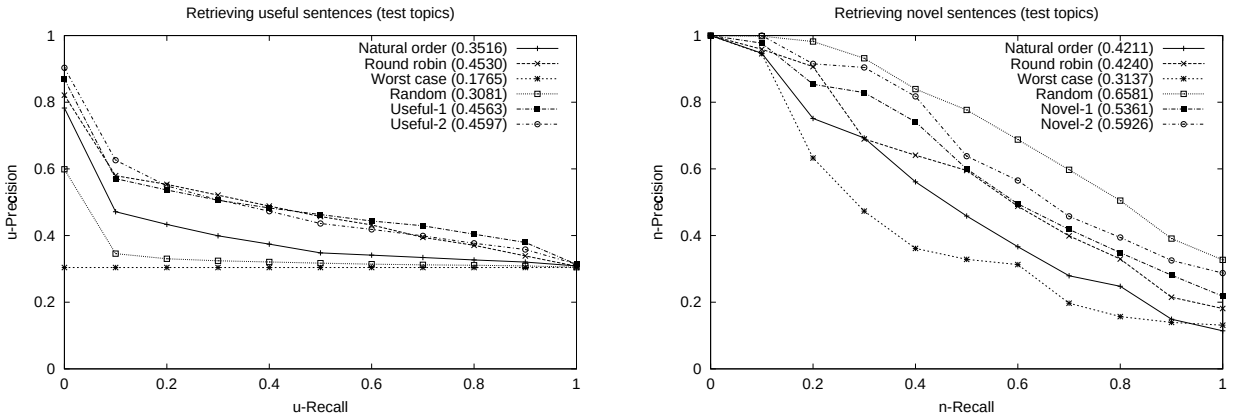

Figure 1: Shows the tradeoff between measures of usefulness (on the left) and novelty (on the right). The numbers in the legend represent the appropriate exact average precision for that approach. We include the following baseline systems: random ranking of sentences, their natural order, first sentences of all stories followed by second then third and so on, and a worst case that represents the worst possible ordering of sentences for the particular evaluation measure.

$$= \left[\prod_{i<k}\left(1-\left[\prod_{w\in s_k}\frac{tf(w,s_i)+0.01}{1.01\cdot|s_i|}\right]^{\frac{1}{|s_k|}}\right)\right]^{\frac{1}{k-1}}$$

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 :

$$P(\text{novel}_2) = P(e(s_k) \neq e(c_i), \forall i \leq m)$$

$$= \left[\prod_{i\leq m}(1-P(e(s_k)=e(c_i)))\right]^{\frac{1}{m}}$$

$$= \left[\prod_{i\leq mk}(1-P(s_k|\text{LM}(c_i)))\right]^{\frac{1}{m}}$$

$$= \left[\prod_{i\leq m}\left(1-\left[\prod_{w\in s_k}\frac{tf(w,c_i)+0.01}{1.01\cdot|c_i|}\right]^{\frac{1}{|s_k|}}\right)\right]^{\frac{1}{m}}$$

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 they 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.

The right of Figure 1 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.

The surprising result is that random sentence ranking substantially outperforms all other approaches, including the more carefully modeled ones: novel$_1$ and novel$_2$. This effect is because off-event sentences are totally ignored. Most sentences are off-event (71% of them discuss no event), but these measures do not penalize a ranking, no matter where those off-event sentences are ranked. Modifying the measures to treat "no event" as a special event on its own (i.e., the first off-event sentence is novel, but

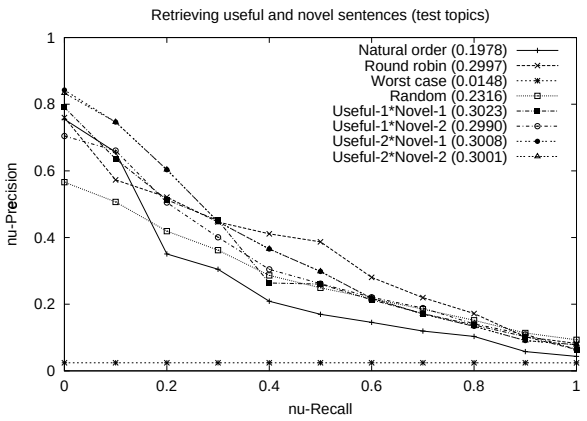

Figure 2: Shows tradeoff between measures of summarization quality based on a combination of usefulness and novelty, for various approaches. The numbers in the legend represent exact average nu-precision for that approach.

the second and subsequent are errors) changes random so that it is approximately as good as the novelty models. It is clear that our models of novelty are weak.

4.3 Combining the models

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 2 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. The novelty graph of Figure 1 showed a clear advantage to novel_2, so it is odd that the choice of novelty measure has no impact.

We have also explored using a linear combination of usefulness and novelty to combine them, but the results are similar.

5 Summary and future work

We have defined temporal summarization and described the framework we developed for evaluating system effectiveness on this task. We showed simple ways of modeling news topics and the events they comprise. Finally, we compared the performance of very simple models to baseline approaches to the problem.

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 [14] may help match events and topics.

We have already tried using a multinomial approach for measuring the probability that a sentence is useful. That is, rather than merely measuring the probability that each word comes from the topic model, we consider the entire set of words in all their possible orderings. The goal of this is to prevent a "sentence" that contains the most probable word dozens of times from scoring well if it does not contain other probable words. We found that there was no difference by our evaluation measures (useful recall and precision), even though the multinomial is intuitively a superior model. We believe this is because the multinomial is "correcting" for a problem that does not typically occur: sentences in news are generally well formed and do not suffer from overly repeated words. We are continuing to investigate this issue.

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

We are grateful to Victor Lavrenko for his thoughtful comments and suggestions regarding the topic and event models.

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.

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, R. Gupta, and V. Khandelwal. Temporal summaries of news topics. In *Proceedings of SIGIR*, 2001. Forthcoming.

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

[4] 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.

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

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

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

[8] 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.

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

[10] V. Khandelwal, R. Gupta, and J. Allan. An evaluation scheme for summarizing topic shifts in news streams. In *Proceedings of the Human Language Technology (HLT) Conference*, 2001. Forthcoming.

[11] H. Luhn. The automatic creation of literature abstracts. In Mani and Maybury [13]. Originally published in IBM Journal of R&D.

[12] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. In Mani and Maybury [13]. Originally published in Information Retrieval.

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

[14] 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.

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

[16] S. Myaeng and D. Jang. Development and evaluation of a statistically based document summarization system. In Mani and Maybury [13].

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

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

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

[20] G. Rath, A. Resnick, and T. Savage. The formation of abstracts by the selection of sentences. In Mani and Maybury [13]. Originally published in American Documentation (now JASIS).