First Story Detection In TDT Is Hard

James Allan, Victor Lavrenko, and Hubert Jin*

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, Massachusetts USA

*MapQuest.Com, Inc.
Mountville, Pennsylvania USA

ABSTRACT

We discuss two Topic Detection and Tracking (TDT) event-based information organization tasks: tracking and first story detection. We show that when a first story detection system is based upon tracking technology, we can expect that performance will be poor. That prediction is consistent with actual performance in TDT evaluations. We then show that to achieve high-quality first story detection, tracking effectiveness must improve to a degree that experience suggests is unlikely. We conclude that effective first story detection is either impossible or requires substantially different approaches.

1. INTRODUCTION

Research into Topic Detection and Tracking (TDT) began in 1996 with a pilot study [DARPA, 1997, Allan et al., 1998a]. The purpose of the study was to determine the effectiveness of state-of-the-art technologies toward addressing several event-based information organization tasks. Because event-based organization is similar to subject-based organization, most approaches to the TDT problems, in the pilot study as well as the following TDT-2 [DARPA, 1999, Papka et al., 1999, Allan et al., 1998b, Yang et al., 1998, Papka, 1999] and TDT-3 [NIST, 1999] efforts, were derived from or were very similar to Information Retrieval (IR) methods. We are interested in two of the tasks for this study:

1. Topic tracking is an analog of the Information Filtering task in TREC [Voorhees and Harman, 1999]. In Tracking, the system is given a small number, N_t, of stories that are known to be on the same event-based news topic (e.g., Oklahoma City bombing, earthquake in Kobe, etc.). The system then monitors the stream of subsequent news stories for ones that are on the same topic.

2. The TDT first story detection (FSD) task also monitors a stream of arriving news stories. In this case, however, the task is to mark each story as "first" or "not first" to indicate whether or not it is the first one discussing a news topic. The system provides a score for each story, where a high score indicates confidence that a story is first.

In order to find out whether current IR technologies could address the TDT tasks, researchers adopted typical methodology: they defined the tasks, built training and test collections, ran a system against them, and evaluated the results. Though they were able to draw solid conclusions about the effectiveness of current technology [Allan et al., 1998a], the TDT efforts have done little to date to examine whether the performance is what would be expected of IR approaches, and whether there are any bounds on effectiveness.

The strong relationship between tracking and filtering raise the obvious question of whether their effectiveness is comparable. The slightly different task definitions, evaluation corpora, and evaluation methodologies make it somewhat difficult to draw a conclusion in that area. However, we have shown suggestive evidence elsewhere [Allan et al., 2000] that tracking effectiveness is approximately what filtering predicts.

In this study, we use a relationship between tracking and FSD to determine whether or not FSD performance is as expected. We use measured effectiveness on one task to predict likely effectiveness on another. It is common practice in complexity analysis, for example, to show that since problem A can be used to solve problem B, if problem B is known to be NP-complete, then A must be similarly difficult [Garey and Johnson, 1979]. In this study, we will reduce First Story Detection to tracking and show that it is almost certainly impossible to realize an effective FSD system with the approaches typically used.

*Much of Hubert Jin's work on this study was performed while at BBN Technologies in Cambridge, Massachusetts.

This paper is organized as follows. The next section presents more details about the tasks that we use to illustrate these ideas. In Section 3 we present an overview of several evaluation measures common in TDT and IR. In Section 4 we show that the First Story Detection task of TDT is as effective as we would anticipate, and argue that only substantially different approaches will improve it. Conclusions and discussion of future directions for research follow in Section 5.

2. TRACKING, FILTERING, AND FIRST STORY DETECTION

We will use two TDT tasks—tracking and first story detection—in this study. This section discusses each of those tasks in some detail. We also describe the corpora that are used for training and evaluation in this study.

2.1 Tracking

The TDT tracking task is fundamentally similar to IR's information filtering task [Voorhees and Harman, 2000]. Each begins with a representation of a topic and then monitors a stream of arriving documents, making decisions about documents as they arrive (without a deferral period). Documents are assigned a score for that topic and, if the score is high enough, are retrieved. The specifics of the tasks are slightly different:

- The topic in filtering is a subject-based query. It is represented by an explicit query, though sometimes is augmented with sample relevant (and non-relevant) documents. Evaluation is usually done over many topics, all evaluated on the same set of documents.
- The topic in tracking is an event-based news topic. It is never represented by an explicit query, but only by a small number of training stories (e.g., $N_t = 4$) that are known to be on the same topic. Evaluation in TDT always starts with the story immediately following the last training story. (This choice has the unusual effect of yielding a different evaluation set for every topic.) Unlike TREC's filtering task, tracking also requires that scores across topics be comparable (i.e., a score of 0.75 represents comparable "relevance" no matter which topic/story pair generates it).
- There is no user feedback after tracking begins. Systems may adapt based on their "guesses" that a story is on topic, but they do not get human confirmation that they were correct.

In this study, we used the TDT-2 corpus, approximately 60,000 news stories running from January through June of 1998. The stories were either newswire text or closed-caption quality transcriptions of radio and television speech.

All parameter tuning for tracking was done using the first four months of data and evaluation was done on the final two months. That split corresponds to the development/evaluation breakdown used in TDT-2, meaning that our results can be compared to those of other TDT-2 sites.[DARPA, 1999] We used 92 topic sets from the TDT-2 workshop [Cieri et al., 1999], and an additional 92 topic sets that were created for the summer workshop [Allan et al., 1999b]. 119 of those topics had at least one on-topic story in the two-month evaluation set.

The tracking task starts with N_t training stories and then processes the *remainder* of the evaluation set looking for other stories on the topic (each topic therefore has a slightly different evaluation set). If a topic has fewer than N_t on-topic stories, then it is not considered during evaluation. Because all 119 topics in the evaluation set have at least one story, the $N_t = 1$ tracking case evaluates 119 topics. There are only 78 topics used in the $N_t = 4$ evaluation.

Our tracking system[1] uses a vector model for representing stories—i.e., we represent each story as a vector in term-space, where coordinates are weighted by the Inquery version of Okapi's tf-idf function. [Allan et al., 1999a]. Terms (or features) of each vector are single words, reduced to their root form by a dictionary-based stemmer.

We represent topics by a vector centroid created by averaging the initial N_t stories' vectors. Incoming stories are compared to the the topic centroid, and if the similarity of the story to the centroid exceeds a threshold, we declare the story on-topic. If the similarity does not exceed the threshold, we declare the story off-topic. Similarity is measured by the well-known cosine similarity function. The threshold was set during parameter tuning on the training data. The choice of threshold does not have any impact on the DET curve; it merely selects a single point on the curve, representing specific miss and false alarm error rates.

2.2 First Story Detection

The TDT first story detection (FSD) task also monitors a stream of arriving news stories. In this case, however, the task is to mark each story as "first" or "not first" to indicate whether or not it is the first one discussing a news topic. In fact, the system provides a score for each story, where a high score indicates confidence that a story is first.

The FSD runs in this study were carried out using the same training and evaluation split of the TDT-2 data as was used for tracking. However, because there is no notion of N_t in first story detection, the evaluation is on the entire two-month evaluation corpus. This means that 119 first stories are known and can be judged as possible misses. An additional 2748 stories are on-topic for one of the 119 topics and are therefore known to be

[1]This system was originally developed for the 1999 summer workshop at Johns Hopkins University's Center for Language and Speech Processing.[Allan et al., 1999b]

non-first stories—they can cause possible false alarms. The remaining stories (approximately 19,000) *must* be processed by the system, but are not evaluated for correct scoring.

Our FSD system is the same as the tracking system. However, rather than comparing incoming stories to a centroid, they are compared to *every* story that had appeared in the past. If the new story exceeds a threshold with any one of the stories, it is considered old, else it is considered new. This approach is similar to the model used by all FSD systems fielded at TDT.

Note that all TDT tasks assume that stories are on a single topic. During evaluation, stories that are on multiple topics are not judged (since there are multiple valid answers: it is likely to be both a first story and a non-first story). This assumption is not unreasonable in practice since only a 2–3% of the stories discuss multiple topics.

3. EVALUATION MEASURES

IR and TDT system evaluations both depend upon a notion of "relevance." In IR a document is or is not relevant to a query; in TDT a story is or is not on a topic.[2] Systems generate scores for every document with respect to the query or topic: the intent is for higher scoring documents to have greater likelihood of being relevant. Most IR tasks simply present the resulting list of ranked documents to the user. Some IR tasks—and all TDT tasks—require that a threshold be chosen such that only documents with a score above the threshold are selected. TDT has the extra requirement that scores are expected to be comparable across topics.

The effectiveness of a system can be evaluated at any particular threshold value by use of the familiar 2×2 contingency table:

	Retrieved	Not retrieved
Relevant	A	B
Not relevant	C	D

A, for example, represents the number of documents that were both retrieved *and* relevant to the query or topic, $A+C$ is the total number of documents retrieved, $A+B$ is the size of the relevant set, and so on. The following commonly used measures from IR and TDT can be expressed in terms of those numbers: Recall, $\frac{A}{A+B}$, is the roportion of relevant material that is retrieved. Precision, $\frac{A}{A+C}$, is the proportion of retrieved material that is relevant. Miss, $\frac{B}{A+B}$, is the proportion of relevant material that is not retrieved; this is the same as 1–Recall. False alarm, $\frac{C}{C+D}$, is the proportion of non-relevant material that is retrieved; this is also called fallout. Finally, richness, $\frac{A+B}{A+B+C+D}$, is the proportion of the collection that is relevant; this is also called generality[Salton and McGill, 1983]

[2]In fact, in TDT a story can also be "briefly" on a topic, meaning that there is a short mention of the topic in a story that is primarily off-topic.

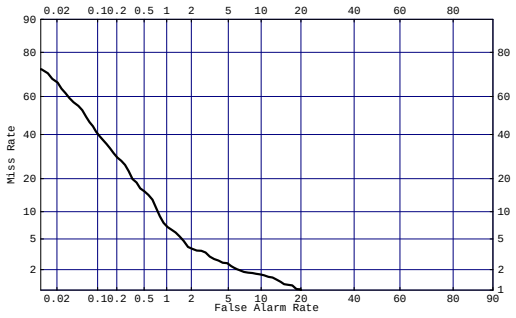

Figure 1: A sample DET curve calculated from a TDT-2 tracking evaluation, with $N_t = 4$.

Several other measures have also been proposed, including measures such as normalized recall and precision, F, and E that combine the measures above [van Rijsbergen, 1979, Salton and McGill, 1983]. We do not consider those measures in this study, even though some of them have appeared in TDT research reports [Allan et al., 1998b, Yang et al., 1998].

3.1 Tradeoffs between measures

The most common measures within the IR community are recall and precision. It has been well established empirically that the two measures are inversely related. The popular recall/precision graph shows how the two are inversely related.

The TDT research community has chosen a signal detection model of evaluation, using miss and false alarm as the preferred measures. Because those are error measures, the goal is to minimize them both. The tradeoff between them is shown on a Detection Error Tradeoff (DET) curve [Martin et al., 1997], a variation on operating characteristic curves [Swets, 1988]. The DET curve was adopted for TDT, but the ideas behind it are far from new in the IR community. The derivation of Swets' model of evaluation [van Rijsbergen, 1979] used the same approach, for example.

Figure 1 shows a sample DET curve. False alarm rate is on the x-axis and miss rate is on the y-axis, meaning that "good" systems have performance curves toward the lower-left of the graph, while the upper-right has high numbers for both error rates.

The axes of the DET curve are on a Gaussian scale—i.e., such that the normal deviate is linear (every standard deviation from the mean advances the same distance on the axes). The result of this is that if the distributions of relevant and of non-relevant document scores are normal, then the resulting DET curve will be a straight line.

It is a source of open debate whether the DET curve is preferable to a recall and precision graph, and whether miss and false alarm are somehow "better" than recall and precision as system effectiveness measures. Because this study focuses on two TDT tasks, we adopt that community's evaluation methodology, and leave the debate for other forums.

4. BOUNDS ON FSD

The goal of first story detection (FSD) is to monitor a stream of arriving news stories and to mark each of them with a score indicating the likelihood that the story is the first on its topic. For example, a successful FSD system would give a high score to the first story on an earthquake and a low score to all following stories that discuss that same earthquake. "First" is defined within the evaluation corpus, so it might actually represent a story well into the news topics broader coverage.

One possible solution to FSD (and more or less the one used by all FSD systems fielded in TDT to date) is to apply tracking technology as described in Section 2.1. Intuitively, the system marks the first story of the corpus with a very high score (it *must* be the first story on any topic in the corpus). It then begins tracking that story. If the second story tracks, it is assigned a low FSD score (since it is on an existing topic). If it does not track (is not on the same topic as the first story), it is assigned a high FSD score, and the system starts tracking that one, too. At any point, the system is tracking numerous topics—in fact, if the system makes an FSD false alarm, it will be tracking some topics in multiple ways.

It should be clear that a perfect tracking system (for $N_t = 1$) yields a perfect FSD system. However, tracking systems are far from perfect. What sort of FSD performance can we expect from a state-of-the-art tracking system?

4.1 Relating tracking and FSD

In order to derive an expected tracking-based FSD performance curve, we need to relate the error measures for both tasks. Suppose that all topics are independent, and that we have encountered $i-1$ topics to date. What is the probability of a miss or false alarm (fa) on the next story? An FSD miss will occur if the story *is* first, but some existing topic tracks it by mistake—alternatively, that it is *not* the case that *none* of the already existing $i-1$ topics accidentally tracks it. That is, the probability that we miss the first story for topic i is:

$$P_{fsd}(miss, i) = 1 - (1 - P_{track}(fa))^{i-1}$$

When averaged over all the first stories of N topics in a collection, and assuming that all topics track with the same erro rate,[3] the topic-weighted average value is:

$$\begin{aligned} P_{fsd}(miss) &= \frac{1}{N}\sum_{i=1}^{N} P_{fsd}(miss, i) \\ &= 1 - \frac{1}{N} \cdot \frac{1 - (1 - P_{track}(fa))^N}{P_{track}(fa)} \end{aligned}$$

An FSD false alarm means that the story was marked as first when it was not. That requires that the story's correct topic misses (fails to track), and that no other topic incorrectly tracks it. This value is more complicated to calculate because it depends upon the number of topics that have already been seen and how they are distributed.

We consider two possibilities to provide lower and upper bounds on the false alarm rate. Recall that a false alarm means that a non-first story was marked as first. That means that a false alarm can only occur on the second or later story in a topic. Assume there are N topics in the corpus. For the lower bound, we assume that every one of the N first stories has been seen before any of the non-first stories, so $N-1$ topics could possibly have incorrectly tracked the topic. That results in the lower bound possibility of a FSD false alarm for topic i:

$$P^{\perp}_{fsd}(fa, i) = P_{track}(miss, i) \cdot \prod_{j=1, j\neq i}^{N} (1 - P_{track}(fa, j))$$

To get the upper bound, we consider the case where every story in a topic occurs before any story in another topic. That is, all stories on topic one arrive, then all on topic two, and so on. That means that for topic i, only earlier topics can incorrectly track:

$$P^{\top}_{fsd}(fa, i) = P_{track}(miss, i) \cdot \prod_{j=1}^{i-1} (1 - P_{track}(fa, j))$$

To find topic-weighted average false alarm rates, we average each of these over all N topics in the corpus. For the lower bound, if we again assume that all topics track

[3] In fact, we make the slightly weaker assumption that the arithmetic and geometric means of the error rates are the same across topics.

with the same error rate this results in:

$$\begin{aligned} P^{\perp}_{fsd}(fa) &= \frac{1}{N}\sum_{i=1}^{N} P_{fsd}(fa,i) \\ &= \frac{1}{N}\sum_{i=1}^{N}\Big[P_{track}(miss,i) \\ &\qquad \cdot \prod_{j=1,j\neq i}^{N}(1-P_{track}(fa,j))\Big] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\Big[P_{track}(miss,i)\cdot(1-P_{track}(fa))^{N-1}\Big] \\ &= \frac{1}{N}\sum_{i=1}^{N}\big[P_{track}(miss,i)\big]\cdot(1-P_{track}(fa))^{N-1} \\ &= P_{track}(miss)\cdot(1-P_{track}(fa))^{N-1} \end{aligned}$$

Similarly, the upper bound can be shown to be:

$$P^{\top}_{fsd}(fa) \approx P_{track}(miss)\cdot\frac{1}{N}\cdot\frac{1-(1-P_{track}(fa))^{N}}{P_{track}(fa)}$$

An important assumption in the derivations above (one that makes the math possible) is that topic error rates are independent—that is, the similarity of a story to topic A has no effect on the probability of its being similar to some other topic B. Although we know that assumption is not entirely valid, we believe that dependencies between topics are rare enough that they can be ignored for our purposes. In the TDT-2 corpus (discussed below), approximately 66,000 stories were judged against 193 randomly chosen topics. 10,029 stories were judged relevant to some topic, and of those only 321 (3.2%) were relevant to multiple topics. Since stories rarely overlap with multiple topics, it is plausible that dependencies between topics do not have any substantive impact on the error rates for tracking them.

4.2 Expected FSD performance

The results above give us a way to predict lower and upper bounds on FSD error rates given tracking error rates. (We emphasize that the predictions only make sense if we assume that the FSD system uses an approach that is based upon tracking.) We will do this by generating a tracking DET curve and then transforming it into lower- and upper-bound FSD error curves: each point on the tracking curve generates two FSD-bound points. We will then show that actual FSD performance falls into that range.

An important parameter of the conversion is the value of N, the number of topics in the evaluation corpus. The TDT-2 evaluation corpus contains 21,255 stories. Of those, 2,847 are known to be relevant to one of 119 topics.[4] If we assume that all topics have an equal number of relevant stories, then there are 23.9 stories per topic, which implies about 900 topics in the evaluation corpus. (The random sampling technique that generates topics makes it unlikely that large topics have been missed, so the average size is probably smaller, and the number of topics slightly larger. It turns out that the bounds are not very sensitive to the value of N once it is above 200-300, so this approximation for N is reasonable.)

Figure 2 shows the system performance of a tracking system run on the TDT-2 evaluation set (see Section 2.1). This performance is comparable to the best systems in the TDT-2 evaluation workshop [DARPA, 1999]. The figure also shows (as a pair of black lines) the resulting upper- and lower-bound performance figures for FSD that result from the conversion described above. The actual FSD system performance is not shown (to reduce graph clutter); it runs near the center of that range. The actual performance is comparable to the best FSD systems fielded in TDT to date. We show the 90% confidence interval[5] for system performance to make it clear that the predicted performance figures are consistent with actual results. (We are showing two confidence intervals, so it is not surprising that the predicted values do not lie entirely within the actual confidence values. What is important is that they substantially overlap.) This result suggests that our FSD system is working about as well as we could expect.

Figure 3 shows a similar pair of graphs, but this time the tracking task is run with $N_t = 1$. Tracking is not as accurate with a single training story, so the tracking curve shows higher error rates. We show this set of curves, however, because it is a better match to the FSD-by-tracking approach described earlier.

4.3 Difficulty of improving FSD

The predicted and actual error rates of a tracking-based FSD system are in fact not very good: they are unacceptably high for all but a few applications, no matter what threshold on the DET curve is used. Although tracking performance is adequate for a wider range of tasks, it is not sufficient to achieve effective FSD. We will show that to realize a high-quality FSD system based on tracking, we will have to construct a nearly-perfect tracking system. There is no reason to believe that current technology can yield such a system, which suggests that FSD systems built around tracking technology cannot be meaningfully improved.

We assume that "reasonable" FSD performance is approximately equal to the tracking DET curve shown in Figure 2 (the lower-left curve). A system that misses less than 10% of the first stories while generating only 1% false alarms is acceptable for many applications. (Certainly we would prefer a system that is even better,

[4] That includes topics that were generated for the TDT-2 evaluation, as well as additional topics that were generated for a workshop on Novelty Detection in the summer of 1999.[Allan et al., 1999b]

[5] The 90% confidence interval is calculated by the official NIST-produced TDT evaluation software. It makes the simplifying assumption that for a fixed threshold, miss and false alarm values have a normal distribution across topics.

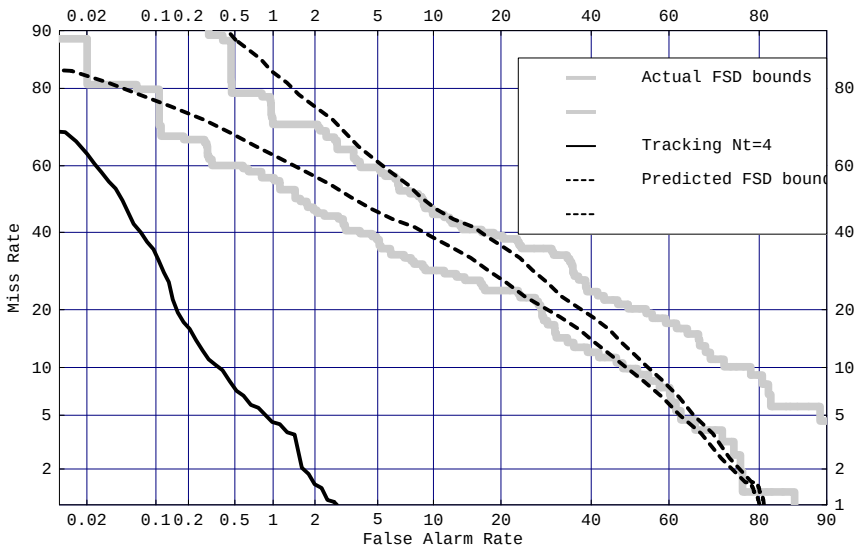

Figure 2: The lower-left solid line is a tracking DET curve for $N_t = 4$. The upper part of the graph shows the lower- and upper-bound predicted performance for tracking-based FSD error rates in black dashed lines. The gray lines show the 90% confidence interval for system performance of an *actual* FSD system.

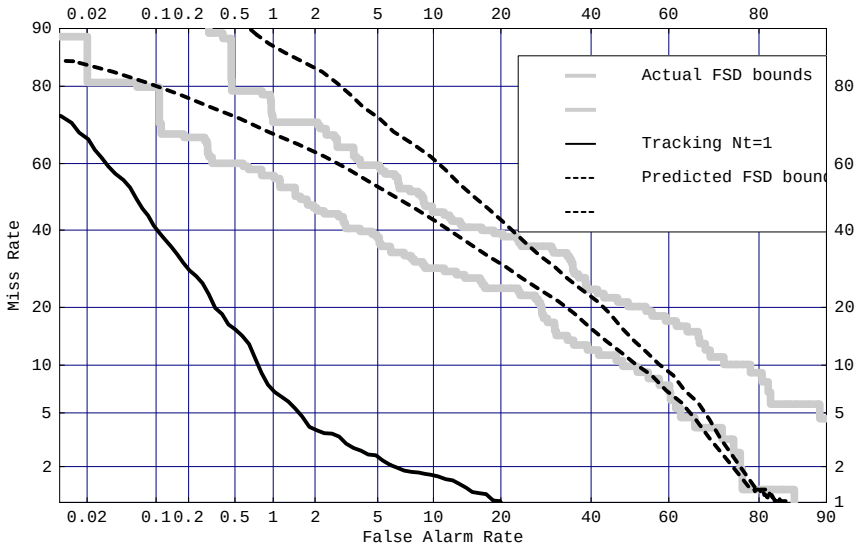

Figure 3: The lower-left solid line is a tracking DET curve for $N_t = 1$. The upper part of the graph shows the lower- and upper-bound predicted performance for tracking-based FSD error rates in black dashed lines. The 90% confidence interval for actual FSD system performance is shown in gray.

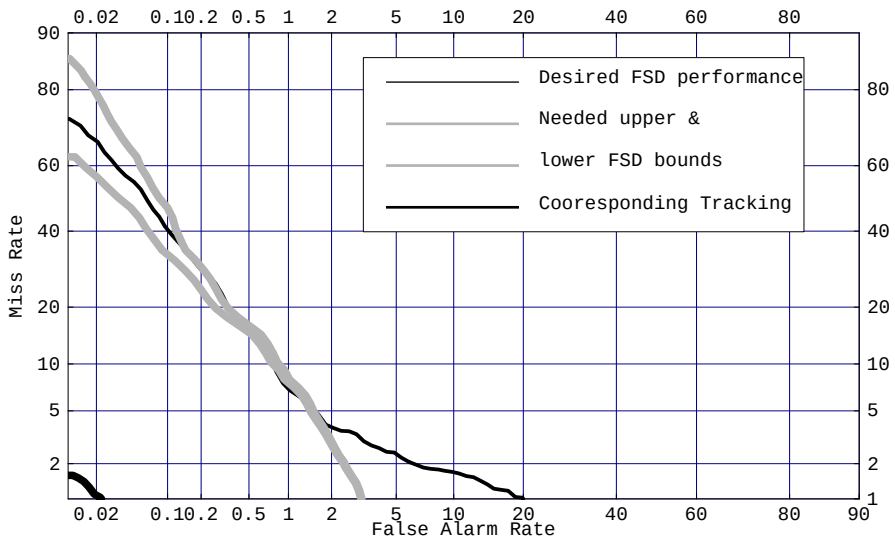

Figure 4: Shows desired FSD performance in black surrounded by reasonable confidence intervals. The extreme lower-left curve is the corresponding tracking performance.

but even this modest goal will be a tremendous challenge.)

Figure 4 shows the desired FSD curve (it is really just the tracking curve again) and lower- and upper-bounds on errors that encompass it. In order to achieve those bounds, we had to improve tracking performance for $N_t = 1$ by a factor of 20. The resulting DET curve is a small line segment in the lower left of the figure.

None of the research in TDT-1, TDT-2, and TDT-3 has resulted in a tracking DET curve that is substantially better than the ones in Figures 2 and 3. Further, as we have argued elsewhere [Allan et al., 2000], tracking effectiveness is as good as filtering effectiveness, and is comparable to that achieved by many years of filtering research at TREC. There is little reason to believe that tracking and filtering technology will ever improve 20-fold.

We have shown how to reduce the FSD problem to a tracking task. We have also shown that a given error rate in tracking results in substantially worse error rates in a corresponding FSD system. Most importantly, we have argued that there is little reason to believe that tracking-based FSD effectiveness can be raised to the point that the technology is widely useful.

5. CONCLUSIONS

In this study, we showed that it is possible to reduce the TDT first story detection problem to the TDT tracking problem. This sort of reduction is extremely important in showing how tasks within and across programs are related to each other. Knowing about such relationships may help avoid redundant research or unnecessary investigative dead-ends. We hope that by describing this technique, we will encourage other such efforts to show relationships between related information technology problems.

We showed that when FSD is based upon tracking, current FSD performance is what we would expect. Indeed, we claim that this relationship between FSD and tracking exists as long as FSD is based upon any full-text similarity comparison (i.e., any approach where we use a text similarity function between a new story and any kind of topic representation).

We also argued that tracking (and filtering) are not likely to make the sort of improvements that are necessary to achieve high-quality FSD results. We view this last result as the main contribution of this study, and a clear call to the TDT research community not to expect large improvements in processing by simple parameter tuning. We have shown that any effort to base an FSD system on a tracking approach is unlikely to succeed. This suggests that new approaches, ones that do not model tracking directly, are necessary. For example, in the summer workshop we tried modeling all previously seen topics as a large collection of names and places, without regard to the specific topics [Allan et al., 1999b]. The results were not very effective, but they illustrated some interesting points about how names and places are used in topics. We have begun similar work that looks at objects and their changing relationships to see if novelty stands out in that way.

We have not given up on tracking since it has value in its own right. We have started work that attempts to leverage the "rules of interpretation" that were used in

the creation of the TDT-2 (and TDT-3) topics. Those rules break news into 11 categories of topics (e.g., scandals, natural disasters, etc., plus an additional "miscellaneous" topic) and describe how the topic relates to the underlying event—i.e., its scope. We believe that modeling topics relative to their class of topic may result in improvements in tracking and therefore in FSD.

The problem of event-based organization of information is an interesting and important one. The TDT studies have shown empirically that generic IR approaches can be adjusted slightly to address such organization tasks, though in the case of tracking-based FSD, only dramatic improvements in tracking will help. Future improvements in both tasks are not likely to come from modifications to the generic approach, but from applying task-specific information about how news topics and events are related and defined.

Acknowledgments

The authors thank Martin Rajman, Charles Wayne, Daniel Gildea, Rose Hoberman, and David Caputo, the other five members of the Topic-based Novelty Detection summer 1999 workshop at Johns Hopkins University, who helped with the broader question of effective First Story Detection. The authors also wish to thank Daniella Malin for her help creating some of the tracking and filtering runs that were evaluated in this study.

The JHU/CLSP summer workshop, where this study was germinated, was funded by the DARPA and by the National Science Foundation under grant IIS-9820687. The first and second authors' work was also supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912, and in part by the Air Force Office of Scientific Research under grant number F49620-99-1-0138. The third author's work was also sponsored by DARPA and monitored by the Space and Naval Warfare Systems Command under Contract number N66001-97-D-8501. The opinions, views, findings, and conclusions contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

6. REFERENCES

Allan, J., Callan, J., Sanderson, M., Xu, J., and Wegmann, S. (1999a). Inquery and trec-7. In *The Seventh Text Retrieval Conference (TREC-7)*, pages 201–216. NIST Special publication 500-242.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998a). Topic detection and tracking pilot study: Final report. In *Proceedings of DARPA Broadcast News Transcription and Under standing Workshop*, pages 194–218.

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

Allan, J., Lavrenko, V., and Jin, H. (2000). Comparing effectiveness in TDT and IR. Technical Report IR-197, CIIR, Department of Computer Science, University of Massachusetts, Amherst.

Allan, J., Papka, R., and Lavrenko, V. (1998b). On-line new event detection and tracking. In *Proceedings of ACM SIGIR*, pages 37–45.

Cieri, C., Graff, D., Liberman, M., Martey, N., and Strassel, S. (1999). The TDT-2 text and speech corpus. In *Proceedings of the DARPA Broadcast News Workshop*, pages 57–60.

DARPA (1997). Proceedings of the TDT workshop. University of Maryland, College Park, MD (unpublished).

DARPA, editor (1999). *Proceedings of the DARPA Broadcast news Workshop*, Herndon, Virginia.

Garey, M. R. and Johnson, D. S. (1979). *Computer and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman and Company, New York.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and ybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proceedings of EuroSpeech'97*, pages 1895–1898.

NIST (1999). 1999 topic detection and tracking evaluation project (TDT3). http://www.nist.gov/speech/tdt3/tdt3.htm.

Papka, R. (1999). *On-line New Event Detection, Clustering, and Tracking*. PhD thesis, Department of Computer Science, University of Massachussetts.

Papka, R., Allan, J., and Lavrenko, V. (1999). UMass approaches to detection and tracking at TDT2. In *Proceedings of the DARPA Broadcast News Workshop*, pages 111–116.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240:1285–1293.

van Rijsbergen, C. (1979). *Information Retrieval*. Butterworths, London.

Voorhees, E. and Harman, D. (1999). Overview of the seventh text retrieval conference (TREC-7). In *The Seventh Text Retrieval Conference (TREC-7)*, pages 1–24. NIST Special publication 500-242.

Voorhees, E. and Harman, D. (2000). Overview of the eighth text retrieval conference (TREC-8). In *The Eighth Text Retrieval Conference (TREC-8)*. NIST Special publication. Forthcoming.

Yang, Y., Pierce, T., and Carbonell, J. (1998). A study on retrospective and on-line event detection. In *Proceedings of ACM SIGIR*, pages 28–36.