Comparing Effectiveness in TDT and IR

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

Keywords: evaluation (general), filtering/routing, event detection and tracking

Abstract

Many of the research tasks in Topic Detection and Tracking have counterparts in Information Retrieval research. However, the two research communities evaluate their tasks differently, which makes it very difficult to determine the extent to which they can help each other. In this study, we compare the performance of the TDT tracking task to the IR filtering task, and show that they have nearly identical effectiveness. We also show a method for using tracking to predict error rates for the TDT First Story Detection (FSD) task. We then show that FSD performance is what tracking predicts. More importantly, we show that with current approaches, FSD performance has probably reached the limits of effectiveness.

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 three new tasks: segmentation of news shows into discrete stories, detection of the onset of previously unseen news topics, and tracking of news topics given a few sample stories. Most approaches to these 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. Such choices were not a surprise because the TDT tasks are similar in many aspects to IR tasks:

- Segmentation [Ponte and Croft, 1997] is very similar to IR efforts that break documents into smaller pieces. The most obvious similarity is to TextTiling [Hearst and Plaunt, 1993], but there are also similarities to best-passage identification [Mittendorf and Schäuble, 1994].

- Detection shares some of the ideas of document clustering [Willett, 1998], particularly those of unsupervised clustering. The major differences are that "news topics" are not the same as the "subject groups" that have more typically been used in IR, that no explicit clusters need to be formed, and that detection is necessarily an on-line task, where a story is completely processed before the next can be considered.[1]

- Tracking is an analog of the Information Filtering task in TREC [Voorhees and Harman, 1999]. Each starts with a query and/or some sample relevant documents, and then monitors a stream of arriving documents for ones that match the query.

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

[1] TDT actually includes two variations on detection. One requires explicit clustering and the other does not. In this study, we focus on the latter; it is called first story detection below.

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 not shown to date [DARPA, 1999] whether the performance is what would be expected of IR approaches. This study addresses that omission.

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 use reductions between and amongst TDT and IR tasks to show two results. First, we will show that tracking performance is exactly what IR filtering performance would predict. Second, 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.

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 both areas. We next discuss the tracking effectiveness that is predicted by information filtering technology in Section 4. In Section 5 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 6.

2 Tracking, filtering, and first story detection

We will use two TDT tasks—tracking and first story detection—and one IR task—filtering—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 Filtering

The IR task of filtering starts with a set of queries. A system monitors a stream of arriving documents. Each document is compared to every query and, if it is sufficiently relevant, the document is "retrieved" for that query.

In this study, we use the TREC-8 filtering task [Voorhees and Harman, 2000] as a basis for our work. That task used TREC queries 351–400 and was run against approximately 200,000 *Financial Times* documents from 1992–1994. The documents were presented in date order, though because we did not do adaptive filtering for this study, the ordering turns out to be unimportant.

In order to make filtering more similar to tracking, we modified the task as follows:

1. The original filtering query was discarded. Instead, a "query" was created using typical relevance feedback approaches, taking $N_t = 1$ or $N_t = 4$ relevant stories and choosing the best terms. This means that filtering starts the same way it does in tracking (described below).
2. The scores were all normalized *after the fact* to fall into the range [0,1]. This provides an approximation of the tracking requirement that scores be comparable across topics. Although this *post hoc* normalization would not be appropriate for a filtering evaluation, it is acceptable for the comparison that we do in Section 4.

2.2 Tracking

The TDT tracking task is fundamentally similar to filtering. Each begins with a representation of a topic and then monitors a stream of arriving documents. 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).

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., 1999]. 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.

2.3 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 an 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 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.

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

[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.

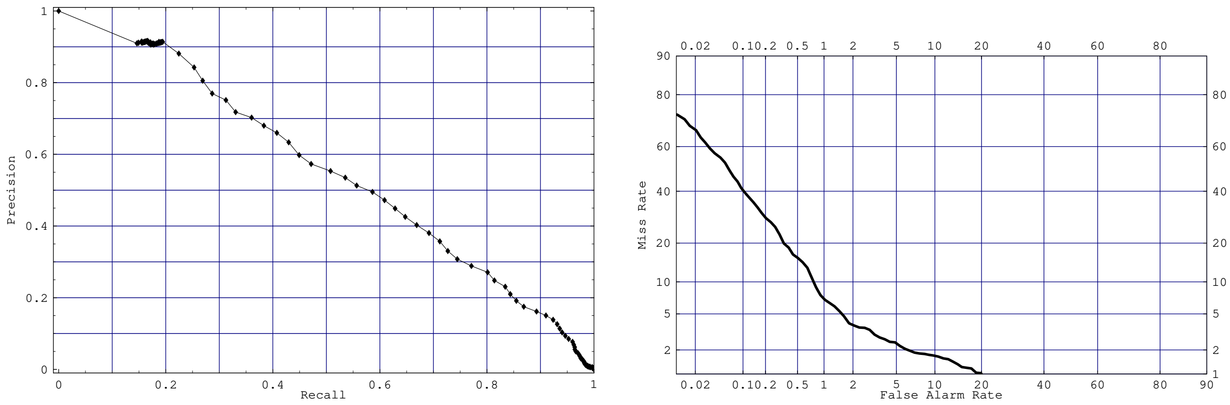

Figure 1: A recall/precision graph and a DET curve. The DET curve was calculated from a TDT-2 tracking evaluation, with $N_t = 4$. The recall/precision graph is the result of converting the DET curve to recall and precision. Note that the resulting graph is *not* an interpolated 11-point graph.

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	$\frac{A}{A+B}$	Proportion of relevant material that is retrieved
Precision	$\frac{A}{A+C}$	Proportion of retrieved material that is relevant
Miss	$\frac{B}{A+B}$	Proportion of relevant material that is not retrieved; this is the same as 1–Recall
False alarm	$\frac{C}{C+D}$	Proportion of non-relevant material that is retrieved; this is also called fallout
Richness	$\frac{A+B}{A+B+C+D}$	Proportion of the collection that is relevant; this is also called generality[Salton and McGill, 1983]

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. Figure 1 shows an example of such a graph that represents average performance on the TDT-2 tracking task. Note that the desirable part of the graph at the right and the top, where recall and precision are respectively perfect.

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.

The right of Figure 1 shows a DET curve that corresponds to the recall/precision graph on the left. 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. The results in this study will suggest that the distributions are at least close to normal.

3.2 Precision or False alarms

Ignoring issues of Gaussian scales, what are the arguments for using a DET curve rather a recall/precision graph? The primary argument in favor of using false alarm over precision is that the former removes topic richness from the evaluation, thereby focusing on the effectiveness of the system and not the quality of the corpus. It can be shown that precision and false alarm are related by the prior probability that a document will be relevant (richness) as follows:

$$precision = \left(1 + \left(\frac{false\ alarm}{1 - miss}\right) \cdot \left(\frac{1 - richness}{richness}\right)\right)^{-1}$$

This relationship means that for a system with constant error rates, increasing richness will increase precision. In Section 4 we will see precisely that effect when filtering and tracking are compared across two corpora.

The argument in favor of precision rather than false alarms has two aspects. First, precision is a measure that seems to be more intuitive to people. It indicates that if some number of documents are retrieved, $P\%$ of them are likely to be relevant. False alarm rate, on the other hand, indicates that if a non-relevant document appears, there is an $F\%$ chance that it will be retrieved. Whether that means being swamped by non-relevant material depends upon the prior probability that a document is non-relevant.

A second argument against the use of false alarms is that most information organization technology—including those discussed in this study—have such a high error rate that only extremely low false alarm rates are interesting. In Figure 1, for example, the DET curve does not even enter the picture until a 50% recall rate, far beyond what most people are interested in. The axes can be stretched to include a lower false alarm rate, but since those rates are at the tails of the distribution, the statistics are often not robust. The DET curve appears to be more useful for technology that has a low error rate.

For this study, we are comparing system technologies and the evaluation corpora vary widely—even tracking and first story detection, although they start with the same corpus, use different portions of it for their runs. For that reason, despite the problems just mentioned, the primary measure used below will be the DET curve. We also show recall/precision graphs to illustrate the differences between the two.

4 Expected TDT Performance

As mentioned above, one purpose of the TDT Pilot Study was to determine the effectiveness of IR techniques on the TDT tasks. Researchers in the Pilot Study chose to answer that question by applying IR methods to the tasks and showing empirically how well they performed. In this section, we will endeavor to show the relationship between the TDT Tracking task and a parallel IR task, Filtering.

We will start by comparing the two tasks on a DET curve. The tracking task is easily evaluated there since that is its intended evaluation measure. Figure 2 shows the performance of tracking at two N_t values in dark lines. Since better performance is closer to the origin, the plot shows that 4 sample stories yields a better topic representation than does only one. The figure also shows filtering performance at the same N_t values. In this case, the different between N_t values is substantially more pronounced: it appears to be much more difficult to build an accurate "query" out of a single document in filtering than it does in tracking.

One thing that the graph shows is that tracking performance at $N_t = 4$ is near the performance that filtering achieves with similar starting information. Although the tasks were run on completely different corpora, and had different definitions, tracking performance is approximately what filtering performance predicts. We hypothesize that the wildly different performance of the tasks for $N_t = 1$ is because news topics are more focused (e.g., "Oklahoma City bombing") than TREC filtering queries (e.g., "drug legalization benefits"). As a result, a single story is a good representative of a news topic, but it might take several documents to isolate the information pertinent to a hidden query.

Because IR systems are more traditionally evaluated using recall/precision graphs, we have also converted the performance of both tasks into that format.[3] Figure 3 shows somewhat different information but yields roughly the same conclusion. The filtering task at $N_t = 1$ results in an impressively poor plot, but the other three tasks are quite good and not all that different.

[3] These recall/precision numbers were generated directly from the DET curve, by converting miss and false alarm numbers into recall and precision values. The curve is *not* an 11-point interpolated graph.

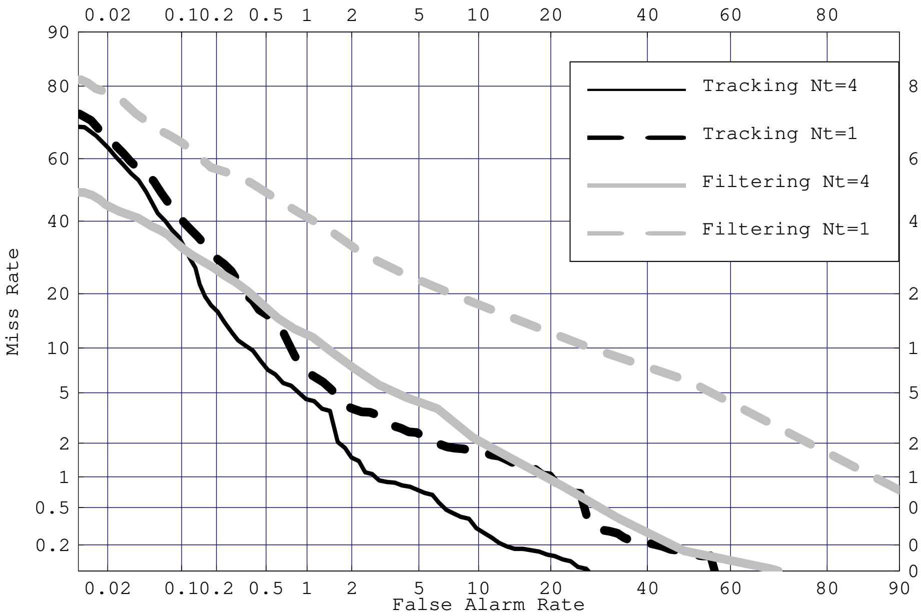

Figure 2: DET plot of two filtering and two tracking runs, each with the “query” generated from $N_t = 1$ or 4 stories.

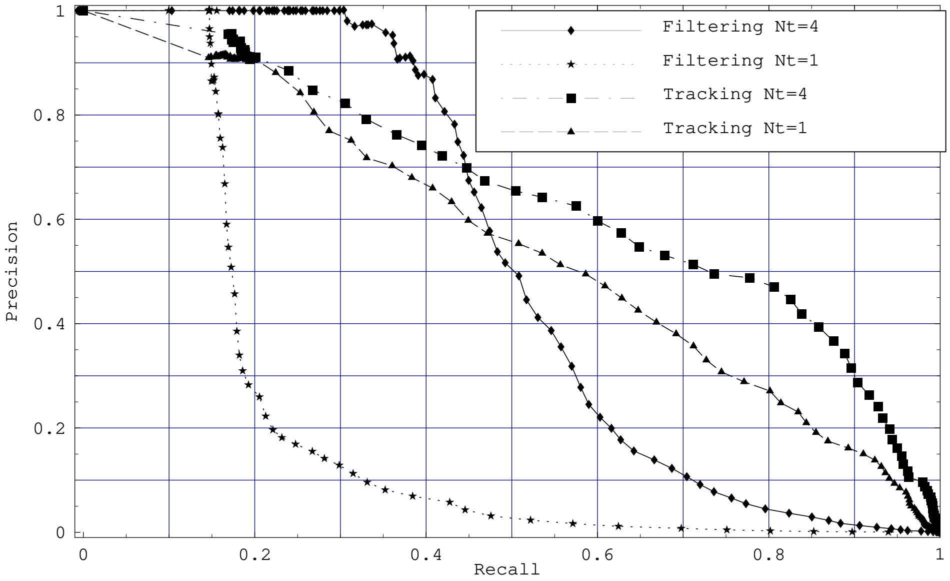

Figure 3: Recall/precision graph of performance for a modified version of the TREC-8 filtering task.

The sets of graphs in Figures 2 and 3 provide an opportunity to highlight some differences in the two evaluation graphs that were mentioned in Section 3.2. In many IR settings, the high-precision end of a graph is the most important—say, recall below 30 percent, the left-hand side of Figure 3. The corresponding portion of the DET curve is a miss rate above 70% in Figure 2. The DET curve barely gives notice to that portion of the curve—the filtering task at $N_t = 4$ falls to negligible false alarm rates at a 50% miss rate.

But there is a larger problem with Figure 3: it purports to compare performance of systems on two different collections. The IR community generally accepts that as invalid, and this curve clearly illustrates why. The DET curve of Figure 2 clearly shows that tracking and filtering with $N_t = 4$ have the same false alarm rate at a miss rate of approximately 35%—i.e., the systems have identical rates of error there. That correspond to a recall of 65%, which in Figure 3 shows precision values that are different by approximately 40%. The problem is that the two corpora have different richness—different densities of relevant documents. A DET curve compares system effectiveness; a recall/precision graph compares system effectiveness on a particular corpus (richness).

In this section, we have illustrated some differences between DET and recall/precision graphs, hoping to show the merit of comparing across tasks. In particular, we showed that the error rates that are typical of a high-quality IR filtering system imply tracking performance that is identical that resulting from the TDT research.

5 Bounds on First Story Detection

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 is to apply tracking technology as described above. 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. 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?

5.1 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 n topics to date. What is the probability of a miss or false alarm 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 n 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, 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,

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

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:

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

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 assume that all topics track with the same error rate[4] this results in:

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

Similarly, the upper bound can be shown to be:

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

5.2 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.[5] If we assume that all topics have an equal number of relevant stories, then there are 23.9

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

[5] 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., 1999]

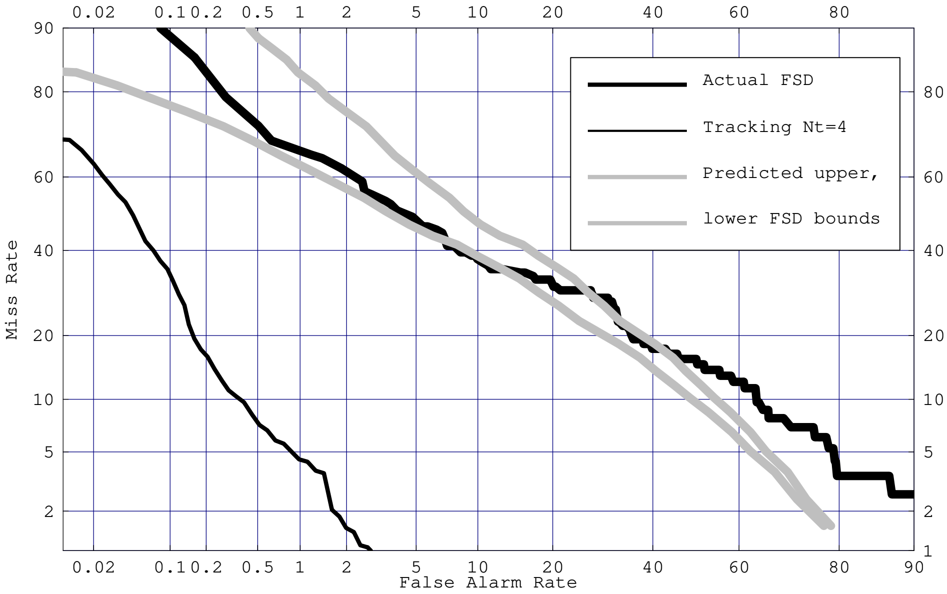

Figure 4: The lower-left graph is a tracking DET curve for $N_t = 4$. The upper part of the graph shows the lower- and upper-bound predicted performance for tracking-based FSD error rates in grey, as well as the actual system performance of an FSD system in black.

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 that large, so this approximation for N is reasonable.)

Figure 4 shows the system performance of a tracking system run on the TDT-2 evaluation set (see Section 2.2). This performance is comparable to the best systems in the TDT-2 evaluation workshop [DARPA, 1999]. The figure also shows the resulting upper- and lower-bound performance figures for FSD that result from the conversion described above. The dark line running approximately between them is the actual performance of an FSD system. Note that the FSD error rates fall nicely within the performance that is predicted by tracking. This result suggests that our FSD system is working about as well as we could expect.

Figure 5 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 curve, however, because it is a better match to the FSD-by-tracking approach described earlier. The predicted effectiveness of FSD is still close enough to the bounds to believe that it is what tracking predicts.

5.3 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 4 (the lower-left curve). A system that misses less than 10% of the first stories while generating only 0.5% false alarms is acceptable for many applications. (Certainly we'd prefer a system that is even

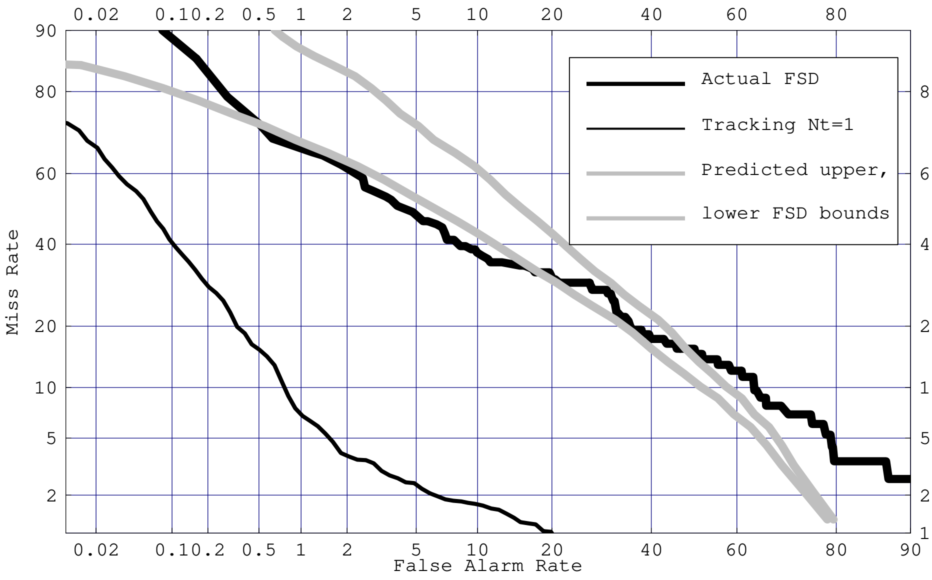

Figure 5: The lower-left graph is a tracking DET curve for $N_t = 1$. The upper part of the graph shows the lower- and upper-bound predicted performance for tracking-based FSD error rates in grey, as well as the actual system performance of an FSD system in black.

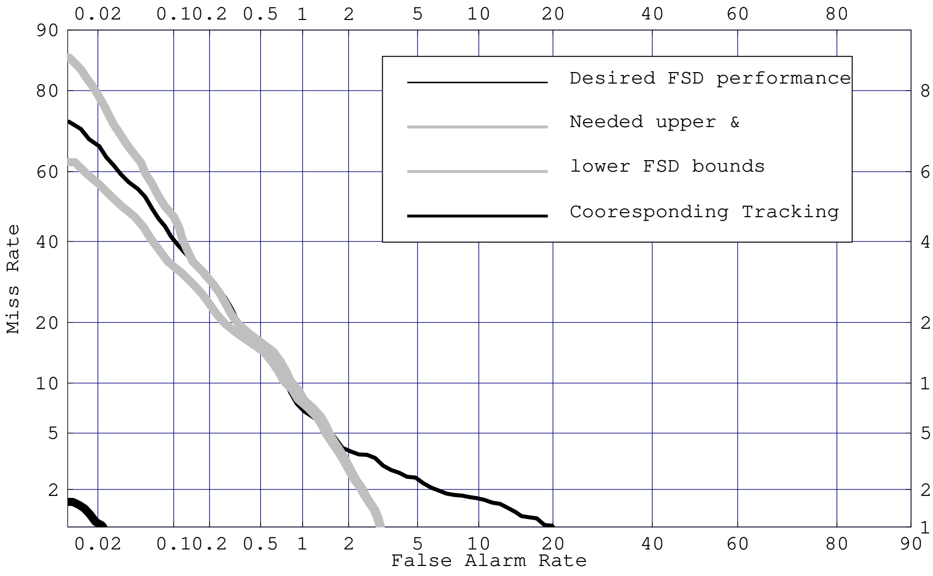

Figure 6: Shows desired FSD performance in black surrounded by reasonable confidence intervals. The extreme lower-left curve is the corresponding tracking performance.

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

Figure 6 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 4 and 5. Further, as shown in Section 4, that level of effectiveness is comparable to that achieved by many years of filtering research at TREC. There is little reason to believe that tracking 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 shown that there is little reason to believe that tracking-based FSD effectiveness can be raised to the point that the technology is widely useful.

6 Conclusions

In this study, we have compared and contrasted two graphs that are used for comparing effectiveness tradeoffs. The first graph shows how recall and precision are inversely related. The second graph, the DET curve, presents the tradeoffs between miss and false alarm. Because the DET curves are not dependent on topic richness, they are preferable for comparing system performance across *different* corpora.

We used DET curves to compare effectiveness on the IR filtering task and the TDT tracking task. We showed that the results of tracking are almost precisely what we would expect given the power of a state-of-the-art filtering system. This sort of comparison may prove more important as more information organization tasks are created: it may not be feasible to construct evaluation corpora for all tasks.

We also reduced the TDT first story detection problem to the TDT tracking problem. We showed that when FSD is based upon tracking, then current FSD performance is what we would expect. 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. 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., 1999]. The results was not very effective, but it 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 no 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.

References

[Allan et al., 1998a] 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 et al., 1999] Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R., and Caputo, D. (1999). Topic-based novelty detection: 1999 summer workshop at clsp, final report. Available at http://www.clsp.jhu.edu/ws99/tdt.

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

[Cieri et al., 1999] 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] DARPA (1997). Proceedings of the TDT workshop. University of Maryland, College Park, MD (unpublished).

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

[Garey and Johnson, 1979] 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.

[Hearst and Plaunt, 1993] Hearst, M. A. and Plaunt, C. (1993). Subtopic structuring for full-length document access. In *Proceedings of ACM SIGIR*, pages 59–68.

[Martin et al., 1997] 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.

[Mittendorf and Schäuble, 1994] Mittendorf, E. and Schäuble, P. (1994). Document and passage retrieval based on hidden markov models. In *Proceedings of ACM SIGIR*, pages 318–328.

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

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

[Papka et al., 1999] 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.

[Ponte and Croft, 1997] Ponte, J. and Croft, W. (1997). Text segmentation by topic. In *Proceedings of the First European Conference on Research an d Advanced Technology for Digital Libraries*, pages 113–125.

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

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

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

[Voorhees and Harman, 1999] 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 and Harman, 2000] 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.

[Willett, 1998] Willett, P. (1998). Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5):577–597.

[Yang et al., 1998] 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.