Abstract
The TREC Spoken Document Retrieval Track (SDR) and the Topic Detection and Tracking (TDT) project have annotated the same corpus with different styles of relevance judgements, using different notions of topic. We compare the behavior of a topic tracking system using relevance judgements from TDT with that of the same system using relevance from the SDR in order to investigate the influence of differences document relevance judgements on the behavior of the tracking system.

1. Introduction
A wide variety of language technologies are principally concerned with determining what a document is about. For instance, in information retrieval (IR), the user enters short phrases (queries) into the system to direct a search of a previously collected corpus for documents that are relevant to what the user had in mind. In a topic tracking system (as in the Topic Detection and Tracking project (TDT)), the user already has one (or a few) documents that describe an interesting event, and uses these seed stories to direct a search of an incoming stream of documents for documents that are on topic. Other examples abound of language technologies that determine what a document is about, e.g. classification, summarization, and filtering. The TDT community has a precise definition of a document’s topic: [1]

A topic is defined to be a seminal event or activity, along with all directly related events and activities.

Although the boundaries of a topic are arbitrary, the LDC has established a well-documented set of guidelines to establish consistency of topic labeling for TDT. Other communities have adopted other definitions of relevance, aboutness, categories. It is not the purpose of the paper to discuss philosophically the merits of the various types of aboutness. (see e.g. [2]) Instead, we are concerned with whether one of these technologies (a topic tracking system) is influenced by differences in the way topics are defined and judged, either in its overall performance, or in its internal tuning. Since the Spoken Document Retrieval Track (SDR) at The Ninth Text Retrieval Conference (TREC-8) built its corpus from the TDT-2 corpus, we have a unique opportunity to make a comparison between the SDR style of topics and judgments and the TDT style of topics and judgments on the same corpus.

Both SDR and TDT have definitions of a topic, but for the purposes of this paper a topic is any subset of documents that can be delineated with a series of binary judgments. In the SDR track, the judges formulated initial ideas of topics based on their knowledge of current events. Queries were then formulated, and topics were selected for inclusion in the evaluation by an initial check with a search engine to ensure that some relevant documents were contained in the corpus. After the evaluation, the judges pooled the results of the systems in the evaluation, and rendered binary relevance judgments on documents that were highly ranked by at least one system. The judgments are reusable, subject to the assumption that unjudged documents are irrelevant [3]. In TDT, the LDC formulated the topics by initially selecting a seed story. After some background research, judgment standards were formulated for each topic. Judges then read every story in the corpus (exhaustive judgments) and rendered binary judgments. More details of the procedure are discussed in [4]. Again, it is not the purpose of this paper to discuss the merits of these different procedures leading to binary judgments. Nevertheless, the differing details of these procedures, in particular the difference between pooled judgments and exhaustive judgments, are one possible explanation of the results to be presented.

2. Experiment
The results presented here contrast tracking experiments using two distinct sets of topics. Both experiments use the manual transcripts (typically close-captioned) of the Feb-June, 1998 English broadcast data from the TDT-2 corpus, the part of the TDT-2 corpus that was used in the SDR track. To make the experiments comparable, some selection of both the TDT and SDR topics was required: For our two sets of topics, we selected the 48 SDR topics and 51 TDT topics that had 5 or more on-topic documents within the SDR corpus. The first four (chronologically) stories on each topic were held out for use as possible seed stories, so that tracking experiments with up to four
seed stories could have identical test sets. The tracking runs presented here all have one seed story (the last one of the four held-out stories.)

![Graph](image)

**Figure 1**: DET curves for tracking on SDR topics and TDT topics

Our tracking system is closely based on our topic detection system [5]; at its core is a dual-threshold incremental clustering using a document-document similarity function. This approach has proven highly successful in the DARPA TDT evaluation [6]. For our document-document similarity function we use a symmetrized version of the Okapi formula [7] which scores documents with other documents. We allow the score of two documents \( d^1 \) and \( d^2 \) to depend upon a cluster \( c \) (generally the cluster to which the earlier of these two documents belongs) so that

\[
Ok(d^1, d^2; c) = \sum_{w \in d^1 \cap d^2} t_w^{1} t_w^{2} \left( idf(w) + 2 \lambda \frac{n_{w,c}}{n_w + n_{c}} \right) \tag{1}
\]

where \( t_w \) is the adjusted term frequency of word \( w \) in document \( i \) (“warped” according to [7] and then normalized so that \( \sum_w t_w^i = 1 \) independent of the length of \( d^i \) and \( idf(w) \) is the traditional Okapi inverse document frequency of word \( w \) (taken from training data; i.e. TDT-2.) Furthermore, \( n_w \) is the number documents (so far) that contain word \( w \), \( n_{c} \) is the number of documents (so far) in cluster \( c \) and \( n_{w,c} \) is the number of documents in the cluster which contain the word; the \( \lambda \) is an adjustable parameter that controls the “dynamic weight” of the cluster-dependent part of the word weight. Furthermore, the score of a document with a cluster is given by the mean

\[
Ok(d, c) = \frac{1}{|c|} \sum_{d \in c} Ok(d, d^i; c). \tag{2}
\]

To perform a tracking experiment, we form all of the training stories into one or more clusters. Clustering proceeds (independently for each cluster) by scoring each test story against that cluster. Two thresholded decisions are made on the basis of the score. If the score exceeds the upper threshold \( \Theta_m \), the story is merged into the cluster (thus affecting future scores through \( n_{c} \).) If the score exceeds the lower threshold \( \Theta_d \), the system outputs “YES” to indicate that the story is on-topic, but does not merge it into the cluster.

![Graph](image)

**Figure 2**: merging threshold

In Fig. (2), we vary the merging threshold \( \Theta_m \), the higher threshold which controls when a cluster’s statistics are updated by a story. We plot \( C_{n,\Theta_m} \) at the optimum decision threshold \( \Theta_d \) against merging threshold \( \Theta_m \) for the two halves (January - March, April - June) of the TDT-2 corpus. In the limiting case on the right hand side of the graph, the system never merges a story into a cluster. We observe significant performance gains when some merging occurs, but which decays rapidly if the threshold is set too low.

Choosing the \( \Theta_d \) threshold is a crucial matter; for the system is allowed only one threshold across all topics. Thus it is important that the typical range of range of scores be independent of topic. The threshold controls the operating point of the system, allowing the user to tradeoff misses with false alarms. We present the principal results in the form of a DET curve [8] in Fig. (1), in order to make comparisons across a range of thresholds and system operating points. We also display curves of constant cost \( C_{n,\Theta_m} \), a linear combination of miss and false alarm probabilities chosen to represent an expected cost of operating a tracking system and normalized so that \( C_{n,\Theta_m} = 0 \) is perfect performance \( C_{n,\Theta_m} = 1 \) is random performance.

### 3. Results

We observe that there is a substantial difference in the overall cost of performance between the two sets of top-
ics, much larger than typical performance differences between competitive tracking systems on the same set of topics. The question naturally arises whether an identifiable feature distinguishes the two sets of topics.

Although an official tracking run has only one threshold across all topics, it is useful to analyze the behavior of the tracking system by retrospectively determining a separate ideal threshold for each topic. For example, a topic that triggers many false alarms would have an ideal threshold higher than the actual threshold of the system. Ideal threshold is a measure of the background noise level that the system must overcome to detect topical documents. The concept of ideal thresholds is also useful for determining how much improvement in tracking performance can be obtained by score normalization (as opposed to improvements in the underlying document-document similarity function.) We show in Fig. (3) a histogram of the ideal thresholds for TDT topics and SDR topics. We observe that the typical ideal threshold is higher for the TDT topics than for the SDR topics. We also observe, Fig. (4) that our system’s performance was systematically better on topics with higher ideal threshold than lower ideal threshold. This systematic effect is likely to account for the system’s overall better performance on the TDT topics. We note also that this system performance difference cannot be related to topic size (number of documents judged relevant per topic.)

4. Conclusion

We think that our tracking system is responding to a fundamental difference in character between SDR topics and TDT topics. A particular consequence of this difference is that our system should be run at a higher threshold on SDR-style topics. Inspection of the SDR topics (from a TDT event-based point-of-view) suggests that the topics fall into several broad categories. Some SDR topics refer to a single event, and correspond well to TDT topics. Other SDR topics request information about multiple events (“What foreign countries has Pope John Paul II visited...?”) but due to the limited time frame of the corpus, there may be only one event in the corpus for the system to discover. Finally, some SDR topics are informational in nature (“Is tuberculosis still a major health concern in the world?”) and do not correspond to TDT-style topics.

5. Acknowledgements

This work is supported by DARPA under SPAWAR contract number N66001-99-2-8916.

6. References


