Chapter 1

Topic Detection and Tracking Evaluation Overview

Jonathan G. Fiscus and George R. Doddington
Information Access Division,
Information Technology Laboratory,
National Institute of Standards and Technology
Gaithersburg, Md 20899

Abstract The objective of the Topic Detection and Tracking (TDT) program is to develop technologies that search, organize and structure multilingual, news oriented textual materials from a variety of broadcast news media. This research program uses controlled laboratory simulations of hypothetical systems to test the efficacy of potential technologies, to gauge research progress, and to provide a forum for the exchange of research information. This chapter introduces TDT's evaluation methodology including: the Linguistic Data Consortium's TDT corpora, evaluation metrics used in TDT and the five TDT research tasks: Topic Tracking, Link Detection, Topic Detection, First Story Detection, and Story Segmentation.

1. INTRODUCTION

The objective of the Topic Detection and Tracking (TDT) program is to develop technologies that search, organize and structure multilingual, news oriented textual materials from a variety of broadcast news media. This research program uses controlled laboratory simulations of hypothetical systems to test the efficacy of potential technologies, to gauge research progress, and to provide a forum for the exchange of research information.

The TDT program began in 1997 with a pilot study involving a small set of researchers who identified potential technologies for automatically organizing news texts. The community continued to meet and to design the primary components of the project, the experimental research tasks, and the data resources needed for research.

Perhaps the most important concept in TDT is that operational systems of the future will process data continuously as it is collected. Most previous research on text retrieval and information organization has been focused on static, retrospective text archives [1]. In contrast, TDT technologies operate on data collected in real time and from a variety of sources and potentially in a variety of languages.

The second concept fundamental to TDT is the notion of an event, or in TDT parlance a *topic*. In TDT, a topic is defined to be "a seminal event or activity along with all directly related events and activities." Since TDT focuses on processing news data, a natural way to organize news articles is by the reported events.

During the pilot study and intervening years, the community selected and defined five research tasks that simulate deployable TDT systems. The tasks were named, *Topic Tracking*, *Link Detection*, *Topic Detection*, *First Story Detection* and *Story Segmentation.*

The National Institute of Standards and Technology (NIST) has administered three open evaluations of the TDT tasks since 1998 [2,3,9]. The NIST TDT website [4] contains information about the evaluations as well as numerous papers and presentations given at the TDT workshops that NIST held after each evaluation.

The remainder of this chapter discusses details of the TDT program's evaluation infrastructure. There are four more sections in this chapter. First, TDT terminology is discussed; this includes the definition of a story and a topic. Second, the data used for research, the TDT corpora, are introduced. The third section is a brief introduction to detection task evaluations, the evaluation formalism used in TDT. The fourth section contains explanations of each of the evaluation tasks.

2. TDT DEFINITIONS: STORIES, EVENTS, AND TOPICS

In the course of preparing corpora for the TDT program, the Linguistic Data Consortium (LDC) [5,6] transcribed hundreds of hours audio recordings collected from TV and radio news broadcasts. Since the unit of retrieval for the TDT program is stories, the LDC annotated the broadcasts with story boundaries. To aid the LDC's annotation of story boundaries, the community agreed that a story is "a topically cohesive segment of news that includes two or more declarative independent clauses about a single event." While the definition doesn't address stories that discuss multiple events, which happens frequently in the TDT corpora, the definition enabled the LDC to tag the data with story boundaries with adequate reliability.

The definition of topic has changed over the course of the program. In the TDT pilot study, the notion of a topic was limited to be an "event", meaning something that happens at some specific time and place along with all necessary preconditions and unavoidable consequences. Later, in the second year, the definition of a topic was broadened to include, in addition to the triggering event, other events and activities that are directly related to it. This definition has persisted for the ensuing years. Formally, the TDT definition of a topic is "a seminal event or activity, along with all directly related events and activities." A story is considered "on topic" when it discusses events and activities that are directly connected to that topic's seminal event. Therefore, for example, a story on the search for survivors of an airplane crash, or on the funeral of the crash victims, will be considered a story on the crash event. Obviously there must be limits to this inclusiveness. (For example, stories on FAA repair directives that derive from a crash investigation would not be considered stories on the crash event.) Since definition of a topic's extension to related events is not readily agreed upon, the LDC has created topic annotation guidelines to improve agreement and consistency of topic labelling. [5]

3. TDT CORPORA

The LDC provided three corpora to support TDT research [5]: the TDT Pilot corpus, the TDT2 corpus and the TDT3 corpus. These corpora are collections of news, including both text and speech, from a number of sources in both English and Mandarin.

The TDT Pilot corpus contains 26K news stories from the Reuters newswire service and transcripts of CNN broadcasts. The corpus spans the period from July 1, 1994 to June 30, 1995. TDT researchers annotated the corpus with 25 events, (using the initial definition of events).

The TDT2 corpus spans the first six months of 1998 and contains 74K news stories from six English and three Mandarin newswire and broadcast news sources. Newswire data are rendered using the original electronic text, with the addition of consistent SGML tagging to minimize formatting differences among various sources. Radio and television material is rendered as digital audio, as human-generated transcripts, and as mechanically-generated transcripts produced by an Automatic Speech Recognition (ASR) system. In addition to these forms, the Mandarin data has been translated to English using the SYSTRAN translation system. The TDT2 corpus is annotated for 100 topics in English, 20 of which have also been annotated in Mandarin. The LDC also annotated 100 topics is support of the 1999 Johns Hopkins Summer Workshop [4].

The TDT3 corpus spans Oct-Dec 1998 contains 45K news stories from eight English newswire and broadcast news sources, and three Mandarin newswire and broadcast news sources, all of which are organized identically to the TDT2 corpus. There are 240 topics annotated in the TDT3 corpus: 120 topics were judged against the whole corpus (including both English and Mandarin), and 120 have been partially annotated against the English portion of the corpus.

Each story in the TDT2 and TDT3 corpora is tagged according to whether it discusses the defined topics. These story-topic tags are assigned a value of *YES,* if the story discusses the target topic, or *BRIEF* if that discussion comprises less than 10% of the story. Otherwise, the (default) tag value is *NO*.

There were two styles of complete annotation used for topic tagging. For the first style of annotation, the annotators were given a list of 20 topics to annotate at a time. The annotator would read a story and judge whether or not the story discussed any of their 20 topics. While this process was thought to be the best way to annotate TDT data, it was labor intensive. During 2000, the second and more efficient technique called *search-guided* annotation reduced the labor by using a search engine to limit the number of stories an annotator had to read. This protocol gave each annotator a single topic to work on at a time and a relevance-ranked list if stories which he/she read until they reached a point of diminishing return. Early investigations suggest that the latter technique produces better consistency presumably due to a reduced cognitive load.

4. EVALUATION METHODOLOGY

TDT is a technology research and development program. At the core of the program is the "technology evaluation cycle" employed by DARPA-sponsored R&D programs in the speech field for many years, Figure 1. The cycle essentially has five phases: task definition, system design, system building, system testing, and system refinement. After the refinement, developers re-evaluate their systems in order to assess how the refinements have affected performance. Periodically, (every year for the TDT program,) there is a community-wide technology evaluation that culminates in a meeting to discuss recent research and progress, and possibly modify the task definition.

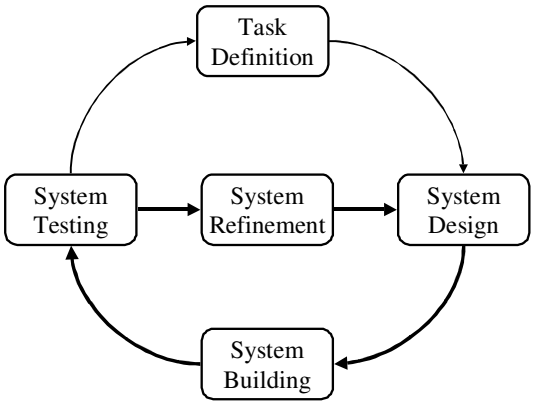

Figure 1. DARPA Evaluation Cycle

This evaluation cycle requires a considerable amount of infrastructure. Not only must the community agree on the evaluation tasks, but they must also agree on the corpora (for training, development and evaluation sets), evaluation metrics, data formats, and system I/O requirements. Consensus on these issues are reached through periodic meetings and codified in the TDT task specification [4]. As the program changes, the task specification codifies in detail the expectations of systems operation. The remainder of this section discusses key components of the task specification.

4.1 TDT Tasks Evaluated as Detection Tasks

All of the TDT tasks are cast as detection tasks. That is, a system is presented with input data and a hypothesis about the data, and the system's task is to decide whether the hypothesis about this data is true. This is called a trial. If the hypothesis is true, the trial is called a target; if not, the trial is called a non-target trial.

A target story can be correctly detected as a target, or a story can be missed in which case the error is called a *missed detection*. A non-target trial can be correctly determined to be a non-target, or it can be falsely detected in which case the error is called a *false alarm.* Table 1 summarizes the contingency matrix of detection system responses.

		Reference Annotation	
		Target	Non-Target
System Response	YES (a Target)	Correct	*False Alarm*
	NO (Not a Target)	*Missed Detection*	Correct

Table 1. Contingency table of detection system responses

Along with the actual decisions, a detection system emits a score. The detection decision is based on this score, which indicates how strongly the evidence suggests that the trial is a target trial. While systems are at liberty to construct their own decision score space, the scores must be comparable across topics and corpus, i.e., a score of 1.0 "means" the same for two different topics, across languages or across sources. Indeed this is a considerable challenge for TDT systems.

There are two techniques for representing performance based on missed detections and false alarms; the detection cost function (C_{Det}) and the decision error tradeoff curve (DET) curve [7]. The former is a "single number" performance measure that estimates system performance at a particular operational point using the actual decisions (YES/NO), and the latter is a visualization of the tradeoff between missed detections and false alarms using the distribution of decision scores.

Since TDT evaluations use many topics, the global assessment of system performance is accomplished by averaging both the detection cost function and DET curves across topics. In TDT, we call these topic-weighted performance metrics. The major advantage to using the averages is that confidence intervals are trivially established for performance measurements as well as outliers are easily identified. Alternatively, global performance could be assessed using a trial-weighted detection cost function and DET curve. In TDT, this is called a story-weighted measure since the trials are typically decisions based on stories. The disadvantage of a story-weighted measure is that topics with disproportionately large numbers of trials can swamp smaller topics.

The remainder of this section discusses the detection cost function and the DET curve.

4.2 Normalized Detection Cost Function

Detection system performance is characterized in terms of the probabilities of missed detection and false alarm errors (P_{Miss} and P_{Fa}). These error probabilities are linearly combined into a single detection cost, C_{Det}, by assigning costs to missed detection and false alarm errors and specifying an *a priori* probability of a target.

The cost model provides a convenient framework for evaluating systems that exhibit a performance trade off between P_{Miss} and P_{Fa}. Intuitively, when a user employs a searching or filtering technology, they're doing so to reduce their workload, e.g., you want to find all the stories that discuss an event while not reading millions of stories. For the user, there's a fixed cost for reading a story, and an increased cost for reading a non-target story, since the time spent was wasted. Thus, the detection cost function uses C_{Miss} and C_{Fa} as

estimates of these costs. Linearly combining P_{Miss} and P_{Fa} using the assigned costs would be sufficient if the richness of targets and non-targets were identical. However, in TDT, and most other filtering applications, the difference is several orders of magnitude. Therefore, a term is needed to compensate for the difference in target richness, hence P_{Target}. The resulting formula for C_{Det} is

$$C_{Det} = (C_{Miss} * P_{Miss} * P_{Target} + C_{Fa} * P_{Fa} * (1 - P_{Target}))$$

$$P_{Miss} = \#Missed\ Detections / \#Targets$$
$$P_{Fa} = \#False\ Alarms / \#Non\text{-}Targets$$

Where

- C_{Miss} and C_{Fa} are the costs of a missed detection and a false alarm respectively, and are pre-specified for the application,
- P_{Miss} and P_{Fa} are the probabilities of a missed detection and a false alarm respectively and are determined by the evaluation results, and
- P_{Target} is the *a priori* probability of finding a target as specified by the application.

For each TDT task, the evaluation specification states C_{Miss} and C_{Fa}. Their values are set using previous experience with detection systems development. For most TDT evaluation tasks, they are set to 10 and 1 respectively. Note that these constants are arbitrarily chosen, and their value is less important than their consistent use. P_{Target} is based on corpus statistics and is a measure of the richness of on-topic stories in the training data. Again, any reasonable choice will suffice as long as the value is used consistently.

While C_{Det} is a convenient measure to assess performance, its dynamic range makes it difficult to interpret, e.g., good performance results in detection costs on the order of 0.001. Therefore, in TDT we use a *Normalized Detection Cost*, or $(C_{Det})_{Norm}$. The goal of normalization is to ground the performance to a more meaningful range. This is accomplished by expanding the dynamic range in the "good performance" range of the scale. To do so, we divide C_{Det} by the minimum expected cost achieved by either answering YES to all decisions or answering NO to all decisions. The resulting normalized cost still has a minimum of zero, but now a cost of 1.0 means a system is doing no better than consistently guessing YES or NO. The derivation of the normalized detection cost formula is as follows:

$$(C_{Det})_{Norm} = C_{Det} / MIN((C_{Miss} * 1.0 * P_{Target} + C_{Fa} * 0.0 * (1 - P_{Target})), (C_{Miss} * 0.0 * P_{Target} + C_{Fa} * 1.0 * (1 - P_{Target})))$$

$$(C_{Det})_{Norm} = C_{Det} / MIN(C_{Miss} * P_{Target}, C_{Fa} * (1 - P_{Target}))$$

4.3 Detection Error Tradeoff Curves

Detection Error Tradeoff (DET) curves are visualizations of the tradeoff between of missed detection (P_{Miss}) rate and the false alarm (P_{Fa}) rate. The curves are constructed by sweeping a threshold through the system's space of decision scores. At each point in the score space, P_{Miss} and P_{Fa} are estimated and plotted as a connected line.

Figure 2 is a DET curve from the 1999 tracking evaluation. The Y-axis is the probability of missed detection and the X-axis is the probability of false alarms. Since missed detections and false alarms are types of errors, improvements in performance will be shown by lines moving closer to the lower left hand corner. Note that the normal deviant scale (expressed as percentages) is used on both axes. The normal deviant scale has advantages over linear scales. It expands the "high performance" region, and resulting straight lines indicate normality of the underlying error distributions of P_{Miss} and P_{Fa}.

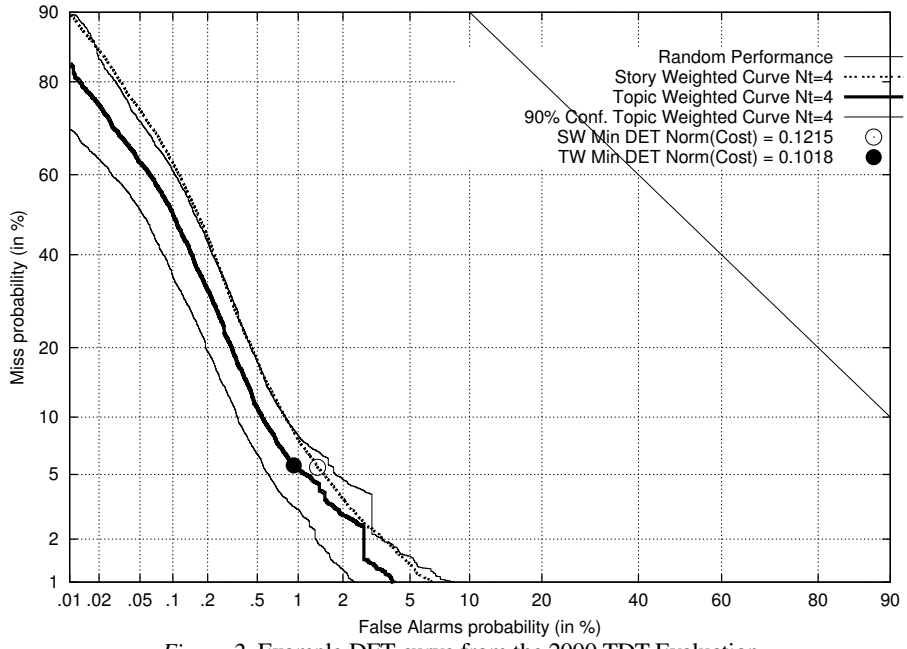

Figure 2. Example DET curve from the 2000 TDT Evaluation

The method described above generates a story-weighted DET curve. Story-weighted DET curves suffer from the same vulnerabilities as story-weighted measures discussed earlier, so TDT uses a topic-weighted DET curve to match the topic-weighted $(C_{Det})_{Norm}$. Topic weighted DET curves are made as follows: sort the stories in order of decision scores separately for each topic. Again, step through the score space, but rather than calculate global P_{Miss} and P_{Fa}, compute the average of P_{Miss} and P_{Fa} across topics. Since means are estimated, variances can also be computed which allows computation of confidence region.

Figure 2 contains both a story-weighted and topic-weighted DET curves. Also presented are the P_{Miss} and P_{Fa}, points corresponding to the minimum in the detection cost function for each curve type. Note the disparity in the story- versus topic-weighted curves. Either technique would be an appropriate means of assigning performance; the benefit to using topic-weighted DET curves is the ability to calculate 90% confidence DET curves and topic-weighted curves have lower variances that story-weighted curves.

5. TASK DEFINITIONS

There are five evaluation tasks in the TDT program. The tasks can vary in focus and size from hypothetical applications to enabling technologies. In brief, the goal of each of the tasks is:

- Topic Tracking – detect stories that discuss a target topic,
- Link Detection – detect whether a pair of stories discuss the same topic,
- Topic Detection – detect clusters of stories that discuss the same topic,
- First Story Detection – detect the first story that discusses a topic, and
- Story Segmentation – detect story boundaries

5.1 Topic Tracking

TDT topic tracking systems detect stories that discuss a previously known topic. A topic is "known" by its association with stories that discuss it. Tracking systems are given these sets of on-topic stories and a portion of the evaluation corpus to train models on. The systems are tested by their ability to find on-topic stories within the remainder of the corpus.

Developers must adhere to three key system design issues.

First, tracking systems must train and test on each topic independently. Systems cannot make use of any other topic's definition, which would

presumably make the task easier. As a by-product of topic independence, the training epochs, the portion of the evaluation corpus used for training the systems models, differ from topic to topic. Since the number of evaluated stories differ from topic to topic, the topic-weighted detection cost function is the preferred system performance metric. Independence of topic has a major advantage. Since the evaluation protocol creates orthogonal topics, stories that discuss multiple topics are evaluated separately for each topic and thus are handled gracefully.

The second system design issue is decision score normalization across topics. Decision scores should "mean" the same thing across topics, so for instance a decision score of 15.0 for one story and one topic indicates the same amount of evidence supporting an on-topic decision for another story and another topic. Mathematically, not only do the means of the underlying target/non-target decision score distributions have to match, but also the variances. Note that this task would be much simpler if systems were allowed to make use of other evaluation topics for score normalization; however, formulating the task as such makes the systems deal with issues of evidence reliability to some extent.

The third system design issue requires tracking systems to be multilingual. Systems must track topics in all languages within the corpus regardless of all training/test language pairs. No doubt, this is a daunting task and requires considerable infrastructure. To make this task more accessible to small researchers, the evaluation corpus includes English translations for the Mandarin texts.

Tracking systems are evaluated using the topic-weighted normalized cost function and the topic-weighted DET Curve, both of which were described in section 4.

There are many experimental conditions identified in the evaluation plan, each enabling developers and NIST the opportunity to decompose system performance on factors that are thought to affect system performance. The TDT 2000 evaluation plan calls out the following conditions: the number of training stories, the number of negative example training stories, the language of the training stories, the form of the broadcast news data, and reference vs. automatic story boundaries.

5.2 Link Detection

TDT link detection systems detect when a pair of stories discuss the same topic (i.e., the stories are "linked" by a common topic). These systems answer the YES/NO question: "do these two stories discuss the same topic?" and output a decision score that the answer is YES. The actual decisions and decision scores are used to calculate $(C_{Det})_{Norm}$ and DET curves respectively as described in section 4.

This task can be thought of as a core capability from which topic tracking and topic detection systems can be built. The link detection task is related to topic tracking with one training story, but rather than track the stories through time, the link detection task sub-samples the story space to be more efficient. Otherwise, a system would need to evaluate *N*(N-1)/2* story pairs.

There are advantages to the link detection paradigm. As defined, the task does not require annotator effort to define topics as in topic tracking or topic detection. Performance can be evaluated using human judgements on random story pairs as to whether or not they discuss the same topic without a formal statement of topic. Since the topic space does not need to be organized into orthogonal clusters of stories, handling of stories on multiple topics is a non-issue.

Another advantage to link detection is the ability to separate performance of monolingual and cross-lingual story pairs. Since system judgements on each story pair are made independently of each other, assessing performance based on any division is simply a matter of sub-sampling the story pairs.

The task is more flexible than the tracking task because there are provisions for systems to take advantage of deferral periods, (a specified amount of future data that can be processed before making decisions on the current story).

There are relatively few evaluation conditions defined by the evaluation plan. For the TDT 2000 evaluation, those conditions were the form of the broadcast news data and the deferral period.

5.3 Topic Detection

The topic detection task evaluates technologies that detect novel, previously unknown, topics. As in the tracking task, topics are defined by associating together stories that discuss the topic. However, topic detection systems are not given *a priori* knowledge of the topic. Therefore, systems must embody an understanding of what constitutes a topic, and this understanding must be independent of topic specifics. The task is multilingual and therefore systems must build clusters that span languages.

The systems detect clusters of stories that discuss the same topic. The concept of clustering is easily applied to news stories, but the assessment of performance is difficult because stories frequently discuss multiple topics. This phenomenon not only means the topic clusters are dependent on previously processed stories, but also that decomposition of performance into casual subsets is misleading.

The evaluation protocol must deal with the issue of topic independence. The multi-topic stories are declared unscorable even though the systems perform clustering on all test stories. Thus, multiple topic stories may influence a system, but they do not contribute to the error measures.

Performance assessment for topic detection uses the detection cost model but with two variations: P_{Miss} and P_{Fa} are calculated after mapping system-defined topics to reference topic clusters, and DET clouds are used rather than DET curves. P_{Miss} and P_{Fa} are calculated for each reference topic cluster using the system-defined cluster that has the lowest detection cost. This reference to system-defined cluster mapping permits system clusters to "map" to any number reference clusters. This mapping is the least cost mapping; therefore, the reported topic detection scores are the minimum score[1]. Second, DET curves are not used since decision scores are not meaningful in the context of detection systems. Instead, detection performance assessment makes use of DET clouds, i.e., a point for each topic's P_{Miss} and P_{Fa}, are plotted on a DET-scaled graph, see Figure 3. The DET cloud also includes the system's topic-weighted average P_{Miss} and P_{Fa}.

As in the other evaluation tasks, stories marked as BRIEF are declared unscoreable and as such are left out of calculations of P_{Miss} and P_{Fa} for the topic.

There are a number of evaluation conditions defined by the evaluation plan. For the TDT 2000 evaluation, those conditions were the source language (English, Mandarin and both English and Mandarin), the form of the broadcast news transcripts, reference vs. automatic boundaries, and the decision deferral period.

[1] A globally optimised mapping that would enforce a 1:1 mapping would yield higher measured detection costs. While such an algorithm is straight forward, it is computationally expensive and it could degenerate to a very long search.

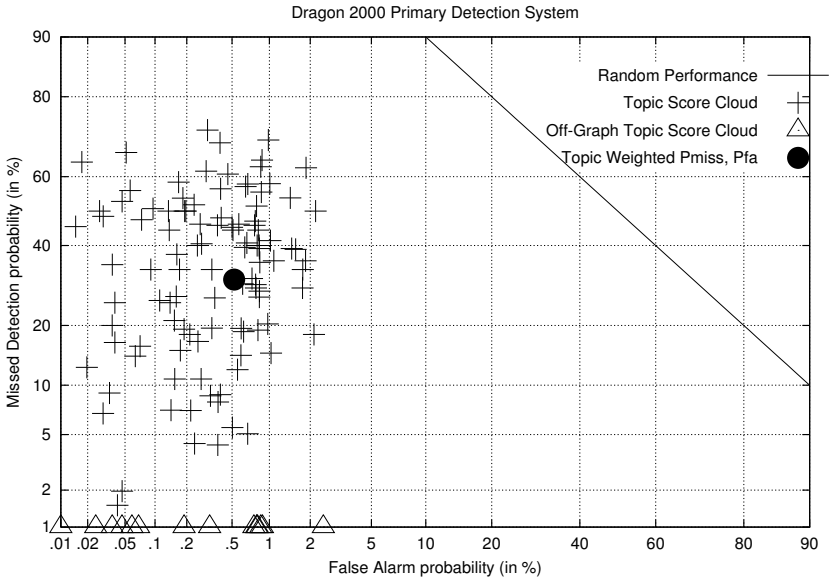

Figure 3. Example DET Cloud from Dragon's 2000 Primary Topic Detection System

5.4 First Story Detection

The first story detection task (FSD) evaluates technologies that detect the first story to discuss topic. This special case of the topic detection task focuses on the specific aspect of topic detection associated with novel information detection, i.e., knowing when to start a new cluster. The task parameters are essentially the same as topic detection. The real difference is in what the system outputs.

FSD systems output an actual decision, either YES or NO, in response to the question: "does this story discuss a new topic?" and a decision score that the answer is YES. While there are relatively few first stories in a corpus, performance assessment for this task uses DET curves in addition to normalized first story detection cost using the same protocol as defined in the evaluation methodology section.

Like topic detection, the FSD evaluation assumes that first stories always discuss a single topic. The TDT annotations of topics disprove this assumption, so the evaluation ignores first stories that are ambiguous, i.e., stories known to discuss a previously seen topic.

Unlike other tasks, stories labelled as BRIEF mentions of a topic are considered as potential non-first stories. However, they are not used as first story candidates.

For the TDT 2000 evaluation, FSD was strictly an English task. The restriction was a pragmatic decision made by the community to streamline the evaluation. The task has the additional evaluation conditions involving the form of the broadcast news transcripts, reference vs. automatic story segmentation, and decision deferral periods.

5.5 Story Segmentation

The story segmentation task evaluates technologies that detect story changes. The systems segment streams of automatically transcribed text into TDT-style stories. In TDT, a story is a "topically cohesive segment of news that includes two or more declarative independent clauses about a single event." The notion of story explicitly excludes commercials from being stories, and therefore systems are not evaluated on boundaries between consecutive commercials.

In TDT, story segmentation is seen to be an enabling technology since all retrieval is story based. This implies that all automatically transcribed speech data will need to be segmented by stories. As previously discussed, TDT is multilingual; the segmentation task is not an exception. Rather than requiring segmenters to work on English translations of Mandarin texts, segmentation systems work on native orthographies.

Performance assessment of segmentation systems makes use of the detection cost model, but the derivation of the missed detection and false alarm probabilities is quite different compared to the other TDT tasks. System performance is judged by determining how well computed story boundaries agree with reference boundaries. This agreement will be judged with an evaluation interval, nominally 15 seconds, that is swept through the input data. The technique is a derivation of the method proposed by Beeferman, et al. [8] The evaluation interval is chosen to be long enough to include all computed boundaries that might reasonably be associated with a true reference boundary, but short enough to exclude unreasonable associations and multiple reference boundaries (i.e., whole stories).

Evaluation is performed by sweeping the evaluation interval through the input source stream and judging the correctness of the segmentation at each position of the interval:

1. If there is both a computed boundary and a reference boundary within the interval, then segmentation is judged correct.
2. Likewise, if there is neither a computed nor a reference boundary within the interval, then segmentation is judged correct.
3. However, a missed detection is declared if there is no computed boundary within an interval that contains a reference boundary,

4. Moreover, a false alarm is declared when a computed boundary exists within an interval that doesn't contain a reference boundary.

The evaluation conditions for the segmentation tasks are the language of the material, the form for the broadcast news data and the decision deferral period, measured in seconds. Note that the task ignores newswire texts since newswire services routinely include story segmentations.

6. SUMMARY

In this chapter, the Topic Detection and Tracking evaluation methodology was introduced. The TDT evaluation methodology is codified by the TDT corpora and the TDT evaluation specification document. The evaluation specification covers three major topics; structure of the TDT corpora, the TDT evaluation metrics, and the TDT research tasks: Topic Tracking, Link Detection, Topic Detection, First Story Detection, and Story Segmentation.

REFERENCES

[1] Voorhees, E, Harman, D., "Overview of the Eighth Text REtrieval Conference (TREC-8)", NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC 8)

[2] Fiscus, J., Doddington, G., Garofolo, J., and Martin, A., "NIST's 1998 Topic Detection and Tracking Evaluation (TDT)", Fifth European Conf. On Speech Comm. and Tech., Vol. 4, pp. 247-250.

[3] Fiscus, J., and Doddington, G., Results of the 1999 Topic Detection and Tracking Evaluation in Mandarin and English, 6th International Conference on Spoken Language Processing, October 2000, Beijing China, SS(06)-05, paper 320.

[4] TDT Homepage at the National Institute of Standards and Technology, http://www.nist.gov/TDT

[5] Cieri, C., Graff, D., Libermann, M., Martey, N., Strassel, S., "Large, Multilingual, Broadcast News Corpora for Cooperative Research in Topic Detection and Tracking: The TDT-2 and TDT-3 Corpus Efforts", Second International Conference on Language Resources and Evaluation, 31 May - 2 June, 2000, pp. 925-930.

[6] Linguistic Data Consortium TDT Corpora Homepage, http://www.ldc.upenn.edu/Projects/TDT

[7] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., "The DET Curve in Assessment of Detection Task Performance", Eurospeech 1997, Proceedings Volume #4 Pages 1895-1898

[8] Beeferman, D., Berger, A., and Lafferty, J., Text Segmentation Using Exponential Models, In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, pp. 35-46. 1997

[9] TDT 2000 Evaluation Website, (Includes presentations and papers) http://www.nist.gov/TDT/tdt2000