NIST'S 1998 TOPIC DETECTION AND TRACKING EVALUATION (TDT2)

Jon Fiscus, George Doddington, John Garofolo, Alvin Martin
National Institute of Standards and Technology
100 Bureau Drive, Stop 8940, Gaithersburg, MD 20899-8940 USA
Jonathan.Fiscus@nist.gov
http://www.itl.nist.gov/div894/894.01/tdt98/tdt98.htm

ABSTRACT

This paper presents a summary of the 1998 Topic Detection and Tracking (TDT) tasks and the results of the 1998 TDT evaluation. The purpose of TDT is to develop technologies for retrieval and automatic organization of Broadcast News and Newswire stories and to evaluate the performance of those technologies. The TDT project builds on and extends the technologies of Automatic Speech Recognition and Document Retrieval with three tasks: 1) Story Segmentation, 2) Topic Detection and 3) Topic Tracking. Each of the tasks simulates a hypothetical operational system that requires incoming data to be processed time synchronously. The 1998 TDT evaluation (TDT2) continues the work of the TDT pilot study conducted in 1997 (TDT1) and is the first open evaluation of TDT tasks.

1. INTRODUCTION

The purpose of the TDT effort is to advance the state of the art in Topic Detection and Tracking. TDT processing addresses multiple sources of information, including both newswire (text) and broadcast news (speech). The information flowing from each source is modeled as a sequence of stories. These stories provide information on many topics. The general technical challenge is to identify and follow the topics being discussed in these stories. Three specific tasks were evaluated:

- Story segmentation
- Topic tracking
- Topic detection

The purpose of the TDT evaluation is to benchmark the performance of TDT technologies and to evaluate the effect of various factors that may affect the performance of TDT technologies. The factors that were evaluated include:

- Automatic speech recognition errors
- Automatic story segmentation errors
- Decision deferral period
- Number of training stories

1.1 Topics

The definition of "topic" is a fundamental issue and of the greatest importance. It is also a very difficult problem, one which has not been fully resolved and for which no perfect solution exists. However, for the purposes of the TDT research effort, a topic is defined to be "***a seminal event or activity, along with all directly related events and activities***" [3]. Stories will be considered to be "on topic" whenever the story is *directly* connected to the associated event.

1.2 The TDT2 Corpus

The TDT2 corpus consists of data collected during the first half of 1998 and taken from 6 sources, including 2 newswires, 2 radio programs and 2 television programs, namely:

- Newswire: Associated Press WorldStream
  New York Times News service
- Radio: Voice of America World News
  Public Radio International The World
- Television: CNN Headline News
  ABC World News Tonight

There are a total of 57 thousand stories in the corpus, including 630 hours of audio. For newswire sources each story is clearly delimited by the newswire format. For radio and TV sources, however, no segmentation is given. Instead, the audio sources were segmented into stories by hand so that each "story" discusses a single topic [1]. The style of story segmentation used in Closed Captioning was used as a guide.

The audio sources were provided in three forms [1]:

- The sampled data audio signal.
- A manual transcription of the speech.
- An automatic transcription of the speech (ASR), produced by an automatic speech recognizer with a word error rate of 30-35 percent).

Each story unit was classified and tagged as NEWS, MISCELLANEOUS, or UNTRANSCRIBED. Only the stories marked as NEWS were used in the evaluations.

100 target topics were defined for the corpus, using a random sampling procedure to cover the 6-month collection period uniformly [1]. Each topic was defined in terms of a three-part identification (what/where/when) along with an explicit description and summary of the topic. Each story in the corpus was labeled according to whether it discussed a topic, for all of the 100 target topics. The labels were either YES if the story was "on-topic", BRIEF if the topic was mentioned only briefly, or NO if the topic was not discussed at all.

The TDT2 corpus was divided into three parts to provide researchers with training data, development test data and evaluation test. The first 2 months served as training data, the second 2 months as development test data, and the last 2 months as evaluation test data.

The Linguistic Data Consortium collected and annotated the TDT2 corpus. Detailed information about the TDT2 corpus may be obtained at ' http://www.ldc.upenn.edu/TDT/'.

1.3 The Segmentation Task

Story segmentation is the task of segmenting the stream of data from a source into topically cohesive stories. Since text (newswire) sources are supplied in segmented form, this task applies only to the audio subset of the corpus (radio and TV). Segmentation of audio signals may be performed using the audio signal itself or the provided manual/automatic textual transcriptions of the audio signal.

Story segmentation performance depends on the form of the source and on the maximum time allowed before segmentation decisions must be output. Evaluation is therefore conditioned on these factors and is performed for three source conditions and three deferral periods (required conditions are in boldface):

Audio source condition:

- Manual transcription
- **Automatic transcription**
- Sampled data signal

Decision deferral period:

- Transcription (words) 100 1000 **10,000**
- sampled data (seconds) 30 300 3,000

1.4 The Tracking Task

Topic tracking is the task of associating incoming stories with topics that are known to the system. A topic is "known" by its association with stories that discuss the topic. Thus each target topic is defined by one or more stories that discuss it. To support this task, a set of training stories is identified for each topic to be tracked. The system may train on the target topic by using all of the stories in the corpus, up through the most recent training story. The tracking task is then to correctly classify all subsequent stories as to whether they discuss the target topic.

Topic tracking performance depends on the form of the source and on the number of training stories for the topic. It also depends on whether story boundaries are provided to the system. Evaluation is therefore conditioned on these factors and is performed for three source conditions and three numbers of training stories (required conditions are boldface):

Source condition:

- Newswire text and the *manual transcription* of the audio sources
- **Newswire text and the *automatic transcription* of the audio sources**
- Newswire text and the *sampled data signal* representing the audio sources

Number of training stories:

1 2 **4**

Story boundary condition:

Given Not given

1.5 The Detection Task

Topic detection is the task of detecting and tracking topics not previously known to the system. It is characterized by a lack of knowledge of the topic to be detected. Therefore the system must embody an understanding of what a topic is, and this understanding must be *independent of topic specifics*. In the topic detection task, the system must detect new topics as the incoming stories are processed and then associate input stories with those topics. Thus this process identifies a set of topics, as defined by their association with the stories that discuss them.

Topic detection performance depends on the form of the source and on the maximum delay allowed before topic detection decisions must be output. It also depends on whether story boundaries are provided to the system. Evaluation is therefore conditioned on these factors and is performed for three source conditions and three deferral periods (required conditions are boldface):

Source condition:

- Newswire text and the *manual transcription* of the audio sources
- **Newswire text and the *automatic transcription* of the audio sources**
- Newswire text and the *sampled data signal* representing the audio sources

Decision deferral period (in terms of # of source files[1]):

1 **10** 100

Story boundary condition:

Given Not given

2. TDT2 TOPIC CHARACTERISTICS

Figure 1 is a bubble chart that depicts the number of on-topic stories as a function of topic and time. Note the considerable variability in the distribution of on-topic stories between topics and also in the density and time duration of topics for the three data sets.

A comparison of the number of on-topic stories across the three data sets indicated that the training set was statistically different from the development and evaluation test sets (90%, 95% and 99.5% confidence level for the Kolmogorov-Smirnov, Median, and Mann-Whitney tests respectively). However, the development set did not exhibit significant differences from the evaluation set.

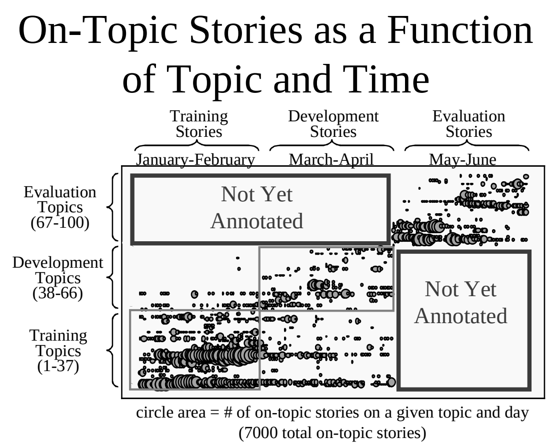

Figure 1. TDT-2 Corpus Topic Distribution

3. EVALUATION

All of the TDT tasks are cast as detection tasks. Detection performance is characterized in terms of the probability of miss and false alarm errors (P_{Miss} and P_{FA}). These error probabilities are then combined into a single detection cost, C_{Det}, by assigning costs to miss and false alarm errors:

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{NOT.target}$$

where

- C_{Miss} and C_{FA} are the costs of a Miss and a False Alarm, respectively,
- P_{Miss} and P_{FA} are the conditional probabilities of a Miss and a False Alarm, respectively, and
- P_{target} and $P_{NON.target}$ are the *a priori* target probabilities ($P_{target} = 1 - P_{NON.target}$).

For TDT2 evaluation, the cost of miss and false alarm were set equal to each other ($C_{Miss} = C_{FA} = 1$) and the *a priori* target probabilities were set to values appropriate for each task. Details of error probability computation are given in the evaluation plan [3].

[1] The source data is divided into files. Each file is a chronologically ordered collection of data from a single source – either a half hour or one hour in the case of broadcast news and an average of about 20 stories for newswire.

For the evaluation of story segmentation, a short (50 word) evaluation interval was scanned through the input source. The correctness of the segmentation was judged at each position of this interval: a false alarm was declared if the segmentation algorithm placed a boundary in the interval while no reference boundary existed in the interval, and a miss was declared if the segmentation didn't place a boundary in the interval while a reference boundary did exist in the interval. For cost-based evaluation of segmentation, 0.3 was assigned as the *a priori* probability of a story boundary existing in an evaluation interval. This assignment was based on the statistics of the training corpus.

For the evaluation of topic tracking, each topic was evaluated separately. Results were then combined for all topics either by pooling all trials ("*story weighted"*) or by weighting the trials so that each topic contributed equally to the result ("*topic weighted"*). For cost-based evaluation of topic tracking, 0.02 was assigned as the *a priori* probability of a story discussing a target topic. This assignment was based on the statistics of the training corpus.

For topic detection, evaluation was limited to the evaluation topics that had been defined and tagged during corpus development. Evaluation was performed by postprocessing the detection system output. First, each of the evaluation reference topics was mapped to the system-defined output topic with the lowest detection cost. Then, detection errors were tabulated for this system topic with respect to the corresponding reference topic. Results were then combined for all reference topics in the same way as for topic tracking. For cost-based evaluation of topic detection,a probability 0.02 was assigned as the *a priori* probability of a story discussing a target topic. This assignment was based on the statistics of the training corpus.

NIST developed a software suite for TDT2 evaluation, TDT2eval Version 0.6. This suite was used to produce the results presented in this section. It is also available for general use and may be accessed from NIST's TDT98 web site, 'http://www.nist.gov/speech/tdt98/tdt98.htm'.

3.1 Participants

Eleven research sites participated in NIST's 1998 TDT2 evaluation; 5 corporate and 6 academic. The groups were: GTE Internetworking's BBN Technologies (BBN), Columbia University (CIDR), Carnegie Mellon University (CMU), Dragon Systems (Dragon), General Electric (GE), IBM's T.J. Watson Laboratories (IBM), SRI International (SRI), University of Iowa (UIowa), University of Massachusetts (UMass), University of Maryland (UMd) and University of Pennsylvania (Upenn). Table 1 indicates the task(s) in which the sites participated. Task participation was voluntary with the proviso that the required condition be processed.

Site IDs	TDT2 Evaluation Tasks		
	Segmentation	**Tracking**	**Detection**
BBN		X	X
CIDR			X
CMU	X	X	X
Dragon	X	X	X
GE		X	
IBM	X		X
SRI	X		
UIowa	X	X*	X*
UMass		X	X
UMd		X*	
UPenn		X	X

Table 1. 1998 TDT Evaluation Task Site Participation
* Submitted after the December 21, 1998 deadline

3.2 Story Segmentation Results

Five research sites participated in the story segmentation evaluation: CMU, Dragon, IBM, SRI and UIowa. Of the nine segmentation evaluation conditions, all segmentation task participants were required to run their primary system on the common condition, namely RecognizedASR source texts(ASR) using a 10000-word decision deferral period. Sites were encouraged to submit results in the remaining eight evaluation conditions and to submit contrastive systems.

Figure 2 is a bar chart showing the segmentation costs achieved by the participants for three source conditions, namely the required ASR transcription and two manual transcriptions.[2] The lowest segmentation cost on ASR text was 0.14, achieved by CMU. The lowest segmentation cost for manual transcriptions was 0.11, achieved by Dragon on both versions.

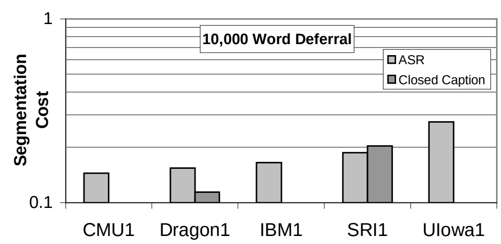

Figure 2. 1998 TDT-2 Primary Tracking Systems

As expected, Dragon's segmentation performance improved when the manual transcripts were used. Note also that SRI achieved improved performance by extracting prosodic information from the speech waveform and incorporating this with the ASR text.

The evaluation plan defines three "Decision Deferral Periods". These periods define the amount of future material a segmentation system can use before making a decision. Figure 3 contains the segmentation costs achieved by three sites that submitted system outputs for different decision deferral periods. The extended decision deferral periods were helpful for the SRI system, but not for the CMU or UIowa systems. In fact, the CMU system, which had the lowest segmentation cost, used substantially fewer than 100 words to make decisions.

[2] In terms of missing each boundary by a fixed number of words, segmentation costs of 0.1 and 0.2 are roughly equivalent to 11 and 25 words respectively.

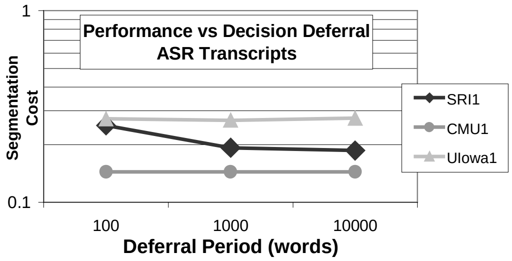

Figure 3. Effect of deferral period on ASR segmentation

3.3 Topic Tracking Results

Eight research sites participated in the topic tracking evaluation: BBN, CMU, Dragon, GE, UIowa, UMass, UMd, UPenn. As in the segmentation task, all tracking task participants ran a primary system on the required evaluation, which was to track topics from both Newswire and ASR sources, using four training stories per topic, and with reference story boundaries given for the ASR text sources.

Figure 4 is a bar chart showing topic tracking costs for the primary systems. The figure indicates a range of performance for the required condition between 0.0056 and 0.0445, with BBN achieving the lowest cost. BBN's tracking cost of 0.0056 corresponds to missing 14% of the on-topic stories and falsely detecting 0.2% of the off-topic stories. In addition to the required condition, five of the participants submitted system outputs for the source condition which used manual transcriptions. In most cases, tracking performance improved slightly when using manual transcriptions rather than ASR.

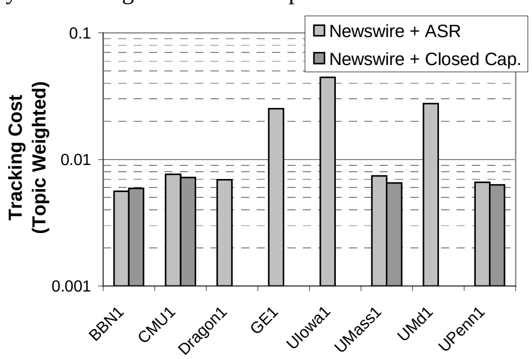

Figure 4. 1998 TDT Primary Tracking Systems

The tracking cost function evaluates a systems' ability to make YES/NO decisions. Tracking systems are also required to output a score for each decision. By varying the decision threshold *ex post facto*, a Detection Error Tradeoff (DET) curve [2] may be produced. Figure 5 is a composite DET curve generated for all the primary systems. Note that the DET curve for the UPenn1 system is better (i.e., has lower tracking cost) than that for the BBN1 system in some regions. In the cost evaluation, however, BBN's tracking cost was lower than UPenn's, because BBN's decision threshold produced a more optimal tradeoff between miss and false alarm errors.

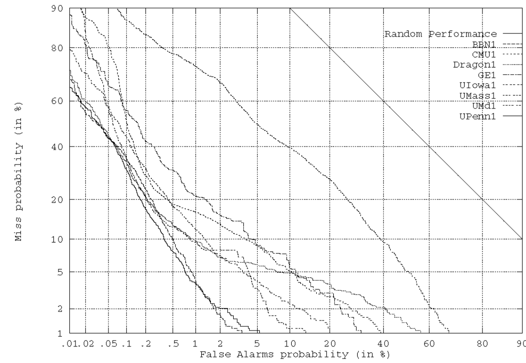

Figure 5. 1998 TDT Tracking System DET Curves

The tracking evaluation supported two interesting contrastive evaluations: variations in the number of training stories, and variations in story segmentation for broadcast news text sources. Figure 6 shows the results of three systems, Dragon1, UMass2, and UPenn1, where the varied parameter is the number of training stories. In all cases, performance was considerably better when systems were presented with four training stories rather that one, with an average of 38% relative improvement in performance.

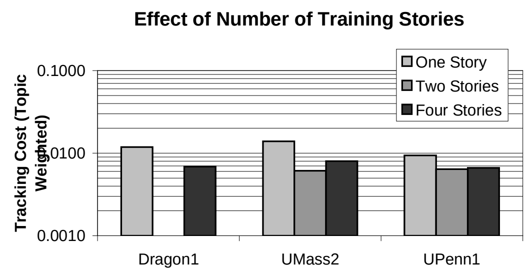

Figure 6. Effect of topic training performance on tracking

The second contrastive tracking evaluation replaces the given reference story boundaries in the ASR texts with the output of an automatic story segmentation algorithm. This contrast represents a fully automated topic tracking system from newswire and broadcast news audio sources. As expected, Figure 7 shows a marked degradation for the newswire and ASR text source condition when the story boundaries change from the given reference to automatic. There is appreciably more degradation here than the difference between Newswire+ASR (with given story boundaries) and Newswire+Closed Captioning.

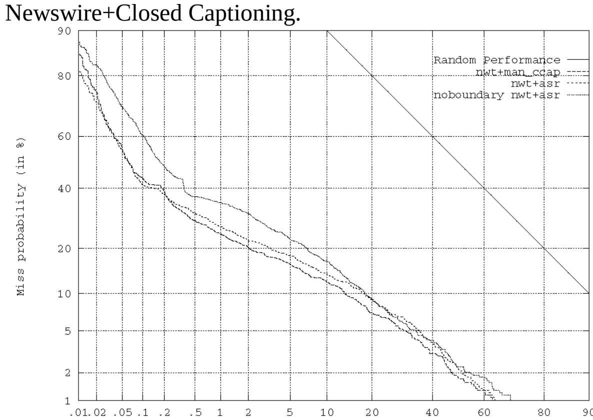

Figure 7. Effect of Automatic Segmentation on Tracking

3.4 Topic Detection Results

Eight research sites participated in the topic detection evaluation: BBN, Columbia University, CMU, Dragon

Systems, IBM, UIowa, UMass, and Upenn. The required evaluation condition was to detect topics in the newswire+ASR source transcripts, deferring decisions for up to ten source files, and using given reference story boundaries.

Figure 8 is a bar chart which shows the topic detection results for each site's primary system for the required condition. The figure indicates a range of performance for the required condition between 0.0042 and 0.0095, with IBM achieving the lowest cost. IBM's detection cost of 0.0042 corresponds to missing 20% of the documents and falsely including 0.07% of the documents. Five of the participants also submitted outputs for the source condition which used manual transcriptions. In most cases, detection performance improved slightly for the manual transcriptions. The improvements were similar to those seen for the tracking task.

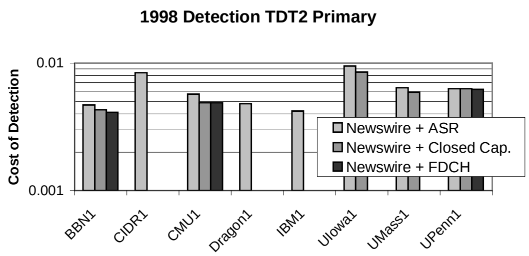

Figure 8. 1998 TDT Primary Detection Systems

As with the tracking evaluation, the detection evaluation supported two contrastive evaluations, one varying the decision deferral period and the second varying the source of ASR story boundaries. The bar chart Figure 9 shows a small improvement with extended decision deferral periods (an average of 7% relative improvement).

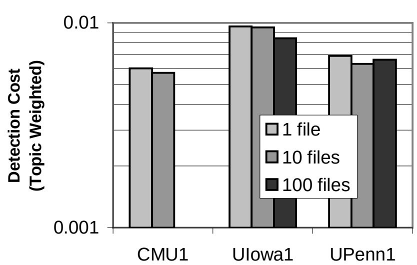

Figure 9. Effect of Decision Deferral on Detection

Figure 10 illustrates the effect of automatic ASR story boundaries on the performance of the CMU1 detection system. The detection costs have been computed by dividing the corpus into two sets, 1) broadcast news "audio source" transcripts, and 2) newswire "text sources", *after* mapping the reference topics to the system-defined topics. There is a 16% relative increase in detection cost when the system is run on Newswire+Closed Caption transcripts as opposed to Newswire+ASR transcripts (with given story boundaries). After division into the audio and text source subsets, the relative increase in detection costs are 27% and 12% respectively. Thus a majority of the additional errors occur in the audio source subset. When the system output from the Newswire+ASR tests (with automatically determined story boundaries) is divided into newswire and audio components, the relative increase in detection cost on the audio subset jumps dramatically by 97%. So, as with tracking, detection performance appears to be quite adversely affected by automatic story segmentation.

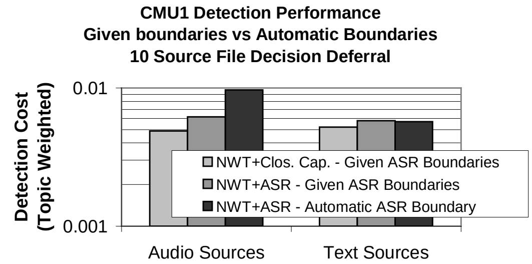

Figure 10. Effect of Automatic Segmentation on Detection

4. CONCLUSIONS

The first TDT Benchmark test was successfully completed and involved eleven research sites. The errors introduced by ASR errors appear to affect tracking and detection similarly. Automatic segmentation of ASR text degrades tracking and detection more than ASR errors alone. Decision deferral periods appear to be useful for detection, more so than for segmentation.

DISCLAIMER

The views expressed in this paper are those of the authors. The test results are for local, system-developer-implemented tests. NIST's role was one that involved working with the community to define the evaluation task definitions, develop and implement scoring software, and score and tabulate the results. The views of the authors and these results are not to be construed or represented as endorsements of any systems or as official findings on the part of NIST or the U. S. Government.

REFERENCES

1. Cieri, C., et al. "The TDT-2 Text and Speech Corpus" Proceedings of the DARPA Broadcast News Workshop, 28Feb-3Mar 1999.

2. Martin, A. et al. "The DET Curve in Assessment of Detection Task Performance", EuroSpeech 1997 Proceedings Volume #4, pp. 1895-1898

3. "1998 TDT-2 Evaluation Specification Version 3.7" http://www.nist.gov/speech/tdt98/tdt98.htm