Results of the 2003 Topic Detection and Tracking Evaluation

Jonathan G. Fiscus
National Institute of Standards and Technology
108 Bureau Drive Stop 8940
Gaithersburg, MD 20899-8940
Jonathan.fiscus@nist.gov

ABSTRACT

The National Institute of Standards and Technology (NIST) administered the sixth open evaluation of Topic Detection and Tracking (TDT) technologies in November of 2003. The TDT project supports development of technologies that automatically organize event-related news stories. The program leverages expertise in core technologies, Automatic Speech Recognition (ASR), Document Retrieval (DR), and Machine Translation (MT) to build the TDT technologies.

The participants in the 2003 TDT project built systems that organized broadcast news and newswire stories collected from Arabic, English and Mandarin Chinese sources. There were four evaluation tasks in the 2003 evaluation: new event detection, topic detection, topic tracking and link detection.

The 2004 TDT Evaluation is scheduled for Fall 2004. The evaluation will focus on the core task of monolingual TDT tasks. The community is also experimenting with a new evaluation task, Hierarchical Topic Detection (HTD). HTD aims to overcome the problems with the topic detection task – specifically topic granularity and multiple-topic stories.

1. INTRODUCTION TO TDT

The TDT project is a DARPA-sponsored evaluation-driven research program to advance the state of the art in technologies that automatically organize event-related stories from multi-lingual information streams (Wayne, 2000). The streams come from either newswire text (NWT) services or audio broadcast news services (BNews), (e.g., television, radio, or webcast).

The plethora of news information confronts users with an overwhelming amount of information in not only English, but non-English languages as well. News delivery agencies routinely deliver data in story units; stories are the building blocks from which users are able to understand the issues in the news. Technology that structures the data or limits the data to present to the user will assist the user during the conversion from raw news to higher-level descriptions of events and activities discussed in the news.

TDT defined a "topic" to be the unit that all research tasks seek to organize. In TDT, a topic is "an event or activity, along with all directly related event and activites" and an event is "a specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences." All stories are judged to be either on-topic or off-topic (Strassel, 2004), and during evaluation, systems are evaluated by how well they can replicate the human topic annotations.

In 1997, TDT began with a small pilot study to decide what technologies would be needed to build news information organization systems. As a result, the first open evaluation occurred in 1998 (Fiscus et al., 1998). NIST has conducted yearly evaluations since 1998. The NIST TDT Website http://www.nist.gov/TDT contains results of the open evaluations and continues to grow with each evaluation.

Participants are furnished with development and test corpora, via the Linguistic Data Consortium (LDC), and with an evaluation plan (Fiscus, 2004) which defines the evaluation tasks, the evaluation metrics and the system input and output requirements. The remaining sections of the introduction briefly describe each aspect of the evaluation.

1.1. TDT CORPORA AND EVALUATION TOPICS

The TDT 2004 evaluation used the LDC's TDT-4 (Strassel 2004) corpus for the evaluation that came from the Oct. 1, 2000/Jan. 31, 2001 epoch. Table 1 contains evaluation-salient statistics of the TDT-4 corpus. The 2004 evaluation marked the second use of the TDT-4 corpus. To "freshen" the corpus, 40 new topics were annotated for the evaluation, and the systems were evaluated over all 80 TDT-4 topics. All BNews sources included automatic speech recognition (ASR) transcripts and all Non-English sources included commercial English translations that the participants had the option to use depending on the evaluation conditions.

	Arabic	English	Mandarin
Newswire Sources	3	2	2
Broadcast News Sources	2	6	5
Number of stories	41728	23602	25405
Total on-topic stories	3104	1926	1303

Table 1 TDT-4 Corpus Statistic Summary

The corpus was heavily weighted towards Arabic to provide good Arabic resources for future work.

1.2. EVALUATION TASKS

The evaluation tasks have evolved since the TDT program began in 1997. Currently, there are five evaluation tasks defined for TDT: Story Segmentation, Topic Tracking, Topic Detection, New Event Detection, and Link Detection.

The TDT story segmentation task is to segment the stream of data from a source into constituent stories. This task applies to BNews, and systems work only on the native orthographies and/or audio signals. NWT text sources are not part of this task because they are delivered already segmented.

The TDT topic tracking task is to associate incoming stories with topics that are known to the system. A topic is "known" to the system by associating a topic id with one or more training stories that have been judged to discuss the topic within the training epoch[1] of the topic. The number of training stories is a parameter of the evaluation condition. We refer to the number of training stories by the variable N_t where N_t can be either 1, 2 or 4. In addition, under some evaluation conditions, the system is privy to high-scoring off-topic stories during the training epoch. The N_n evaluation parameter represents the number of off-topic stories and can have the value 0 or 2. The tracking system must then classify all subsequent stories in the evaluation epoch as to whether or not they discuss the target topic.

The TDT topic detection task is to detect, or find, all topics as the corpus is processed. Then, track the found topics for the remainder of the corpus. Topics are not "known", via training exemplars, to the system prior to processing the corpus – thus topic detection is an unsupervised training version of topic tracking.

The TDT new event detection (NED) task is to detect the first mention of an event that occurs in a sequence of chronologically ordered stories. Like topic detection, there are no topic training stories. Conceptually, the task is a glass box evaluation of a topic detection system where the task measures how well a system can detect the first story of an event and start new topic cluster. New event detection was previously called the "first story detection" task, (Fiscus 2000).

The TDT link detection task is to detect when two stories discuss the same topic, and are therefore 'linked'. Systems have no a-priori knowledge of the topic. Thus, the system must embody an understanding of what a topic is, and this understanding must be independent of topic specifics. Links are not constrained to segregate stories into a set of orthogonal topics, and there is no presumption that each story discusses one and only one topic.

1.3. EVALUATION METHODOLOGY

The evaluation model used for the TDT tasks is the detection model used by the Text Independent Speaker Recognition community (Martin et al., 1997). The model views performance as a tradeoff between two error types: missed detections and false alarms. Such systems have many operating points, so TDT evaluates system performance both by Detection Error Tradeoff (DET) curves and by the normalized detection cost function.

The Detection Error Tradeoff (DET) Curve is a graphical depiction of th tradeoff between missed detection and false alarms. The normalized detection cost (NDC) function distills performance into a single number. The cost function is a linear combination of the costs associated with missed detections and false alarms. The normalization step scales the costs to be 0.0 for perfect performance and 1.0 for the best score achievable by saying "NO" for all detections. The TDT 2003 evaluation plan (Fiscus, 2004) describes the evaluation protocols in detail.

2. TDT 2003 EVALUATION RESULTS

Four research groups participated in the evaluation (Fiscus 2004): Carnegie Mellon University (CMU), Royal Melbourne Institute of Technology (RMIT), Stottler Henke Associates, Inc. (SHAI), and University of Massachusetts (UMass) (Allan et al., 2003).

The evaluation project supported all five evaluation tasks described in section 1.2, however no group participated in the story segmentation task. The following sections discuss the system results for each task's primary[2] evaluation condition.

2.1. TOPIC TRACKING

CMU, RMIT and UMass participated in the 2004 topic tracking multi-lingual evaluation. Systems tracked topics in the three languages, Arabic, English and Mandarin, using English stories for topic training.

The topic tracking task has two primary conditions which represent two extremes of the source text quality. The high quality corpus condition includes newswire texts and human transcribed BNews with N_t=1 English training stories. The degraded corpus condition uses newswire texts and automatically transcribed BNews with N_t=4 English training stories and N_n=2 off-topic training stories. For each condition, participants had the choice to use either native orthography or English translations of non-English sources. The normalized topic tracking costs are presented in Table 2 and the DET curves are presented in Figure 1.

[1] For each topic, the evaluation corpus is divided into two epochs, the training epoch which holds the training stories, and the evaluation epoch for which systems must track the topic in.

[2] Each evaluation task has a defined "primary" evaluation condition that all task participants are required to run.

	Newswire+Bnews Human Trans. N_t =1, N_n =0		Newswire+BNews ASR Trans. N_t=4, N_n=2
	RMIT1	UMass01	CMU
NDC	0.2983	0.2135	0.2825

Table 2 2003 Primary Topic Tracking Systems

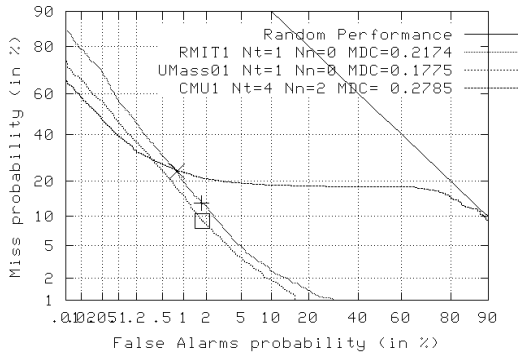

Figure 1 2003 Primary Topic Tracking System DET Curves

2.2. TOPIC DETECTION

CMU, RMIT and UMass participated in the topic detection task. The primary evaluation condition was to use the multilingual texts with either English translations of non-English sources or the native transcripts, the ASR transcripts for BNews, and a 10 file decision deferral window[3]. Table 3 contains the topic detection costs.

	Newswire+BNews Human Trans.	Newswire+BNews ASR Trans.	
	RMIT1	CMU1	UMass3
NDC	0.623	0.3035	0.3094

Table 3 2003 Topic Detection Systems

2.3. NEW EVENT DETECTION TASK

CMU, SHAI and the Univ. of Mass. participated in this evaluation. The primary evaluation condition for NED is newsire texts and BNews automatic transcripts, native English language material only. The normalized NED costs and DET curves are in Table 4 and Figure 2 respectively.

	CMU1	SHAI1	UMass1
NDC	0.5967	0.6615	0.6536

Table 4 2003 Primary NED System Scores

[3] The decision deferral window is the number of source files that can be read in advance for making a decision about a story.

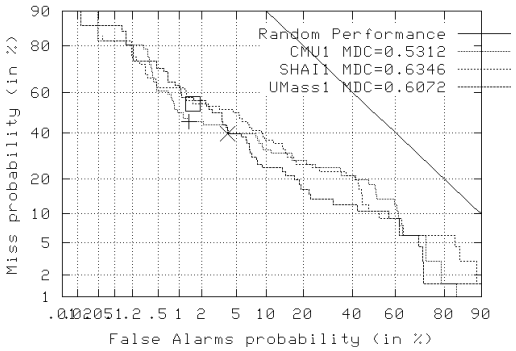

Figure 2 2003 Primary NED DET Curves

2.4. LINK DETECTION

CMU and UMass participated in the link detection task. The primary evaluation condition was to use multilingual texts with English translations of non-English sources or the native transcripts, the ASR transcripts for BNews, and a 10-file decision deferral window. The normalized link detection costs and DET curves are in Table 5 and Figure 3 respectively.

	With English Translations		Native Orthography
	CMU1	UMass1	CMU1
NDC	.2176	.1839	.2199

Table 5 2003 Primary Link Detection Scores

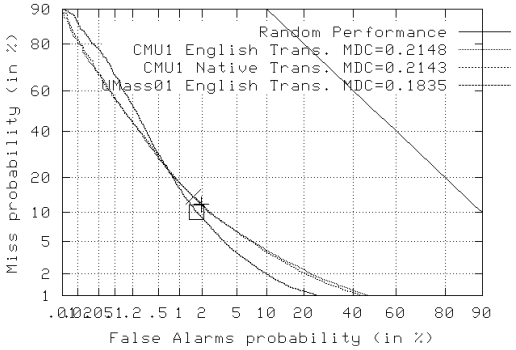

Figure 3 Primary Link Detection DET Curves

3. TDT 2004 EVALUATION PLANS

The TDT community is planning an evaluation for the fall of 2004. The thrust of the evaluation will be the core issue of TDT: being able to organize vast amounts of textual data. The existing tasks, topic tracking, topic detection, new event detection and link detection will remain TDT tasks. However, the evaluation project will change four aspects of TDT this year in order to push the technology forward. The changes to TDT this year include: (1) a variation of topic tracking, supervised topic tracking, (2) a new task, hierarchical topic detection, (3) a new corpus, TDT-5, and (4) more annotated topics for the evaluation using a new annotation scheme "time-limited, search-guided annotation".

3.1. SUPERVISED TOPIC TRACKING

Supervised topic tracking is a variation of the topic tracking task. The existing topic tracking task permitted unsupervised adaptation of the topic, however this task variation gives systems the option to query the topic annotations to simulate human-interaction. The TDT 2004 evaluation plan will fully describe the task.

3.2. HIERARCHICAL TOPIC DETECTION

The current topic detection task cannot represent varying levels of topic granularity nor evaluate stories that discuss multiple topics. A potential replacement task, hierarchical topic detection (HTD), seeks to overcome these problems.

Topic granularity is the level at which a topic is described. For instance, the "Asian Economics Crisis" is a long-running topic that pertains to many countries. Whereas, the "G-7 World Finance Meeting" topic is focused on an event – the meeting. Clearly, the breadths of the topics are different, yet the system developers tune their system to maximize performance for both topic types using the flat topic structure defined for the evaluation task.

The topic detection evaluation protocol explicitly ignores stories known to discuss multiple topics. Systems are not able to place a story in to more than one cluster so the evaluation protocol ignores them during scoring.

The proposed HTD task is to organize the texts into a directed acyclic graph (DAG) structure where nodes represent topics and leaves point to stories. In the DAG structure, nodes can be composed of subnodes, (to represent topic granularity), and stories can belong to multiple nodes, (to allow stories to discuss multiple topics).

Many issues are presently unresolved with respect to evaluating such an HTD system. The summer's research will be devoted to answering these questions.

3.3. THE TDT-5 CORPUS

The LDC will create the TDT-5 Corpus. The corpus will contain Arabic, English and Mandarin newswire components collected from a contemporaneous epoch. Unlike the TDT-2, TDT-3 and TDT-4 corpora, the corpus will not include any broadcast news but will still contain English translations. Depending on resource constraints, some foreign language material may not be used for topic annotation.

There will be many more topics annotated on the corpus. The current target is from 200-250 topics.

3.4. FAST, PARTIAL TOPIC ANNOTATION

The art of topic building and annotation has evolved during the TDT evaluation program. The topic definition strategy will not change for the TDT-5 corpus. However, the topic annotation procedure will change in order to meet the researcher's desire for 200-250 topics. Rather than using search-guided annotation (Strassel, 2004), the LDC will use a time-limited search-guided annotation procedure. The change is to place a fixed time limit on the amount of time per topic an annotator can spend. Thus, the LDC will be able to annotate more topics than the typical 40 topics used in previous years.

4. CONCLUSIONS

The 2003 Topic Detection and Tracking evaluation included researchers from CMU, RMIT, SHAI, and UMass. The lowest achieved error rates on the primary evaluation conditions are: 0.2315 for Topic Tracking by UMass, 0.3035 for Topic Detection by CMU, 0.1829 for New Event Detection by CMU, and 0.1839 for Link Detection by UMass.

The TDT 2004 evaluation in the fall of 2004, will use a new TDT corpus, and will experimentally include a new task, hierarchical topic detection.

5. DISCLAIMER

The views expressed in this paper are those of the author's. The test results are for local, system-developer-implemented tests. The views of the author's 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.

6. REFERENCES

- Allan, J., Bolivar, A., Connell, M., Cronen-Townsend, S., Feng, A., Feng, F., Kumaran, G., Larkey, L., Lavrenko, V., and Raghavan, H., "UMass TDT 2003 Research Summary" CIIR, Department of Computer Science, UMass Amherst Technical report.
- 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
- Fiscus, J., Doddington, G., "Results of the 1999 Topic Detection and Tracking Evaluation in Mandarin and English", ICSLP 2000
- Fiscus, "NIST TDT Web Site" http://www.nist.gov/TDT, 2004
- Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., "The DET Curve in Assessment of Detection Task Performance", 1997 Fifth European Conf. On Speech Comm. and Tech., Vol. 4, pp. 1895-1898
- Strassel, S. et al., 'LDC TDT-4 Corpus". http://www.ldc.upenn.edu/Projects/TDT4/, 2004
- Wayne, C., "Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation", Second International Conference on Language Resources and Evaluation, 31 May - 2 June, 2000, pp. 1487-1493.