THE TDT-2 TEXT AND SPEECH CORPUS

Chris Cieri, David Graff, Mark Liberman,
Nii Martey, Stephanie Strassel

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104

ABSTRACT

This paper describes the creation and content of the TDT-2 corpus in the context of the TDT-2 research project it supports and in comparison to previous and subsequent efforts

1. INTRODUCTION

The second phase of the Topic Detection and Tracking research project (TDT-2) was larger and more ambitious in virtually every respect compared to its predecessor, the TDT Pilot project. This was especially true with regard to the size and scope of the data collection needed to support the research. This paper will summarize the relevant facts about the TDT-2 corpus, and contrast it with the TDT Pilot corpus and with the upcoming TDT-3 corpus, which is being prepared for use in 1999. We will briefly describe the procedures used at the LDC to collect and prepare the data, and to elicit, check and maintain the human judgements that comprise the most important feature of the corpus, the annotations of topic relevance. Alternative methods of data creation will be reviewed, some of which may be applied in the next phase of TDT. This will lead to a discussion of the issues that arose in selecting and defining topics, resolving and maintaining the scope of a topic, and organizing information about the topical relatedness of news stories. We will explain how these issues were addressed during the creation of the TDT-2 corpus, and present some alternative approaches that were suggested or explored during the course of the project.

2. PROJECT OVERVIEW

The TDT-2 corpus was created to support three research tasks defined in the evaluation plan for the project: segmentation, detection and tracking. We will provide a very brief description of them here. The interested reader should consult Charles Wayne's overview of the task in [5] and [6], George Doddington's description of the evaluation specification in [1] and [2] and NIST's summary of 1998 results elsewhere in this volume. In the TDT model, news from multiple sources streams through a system that *segments* the news streams into individual stories based only upon their content, *detects* stories that discuss novel events and *tracks* stories related to a target topic throughout the data stream. In support of those tasks, the TDT-2 corpus includes:

- Original reference audio which the research sites may use to generate their own text
- SGML-structured, segmented reference text used only as background information and excluded from the evaluation tasks
- Tokenized, non-segmented text
- ASR output for all audio generated by Dragon Systems
- Boundary tables indicating the position of story boundaries within tokenized files and ASR output
- NIST's index imposing a canonical chronological ordering of stories
- Relevance tables indicating which stories are related to each topic

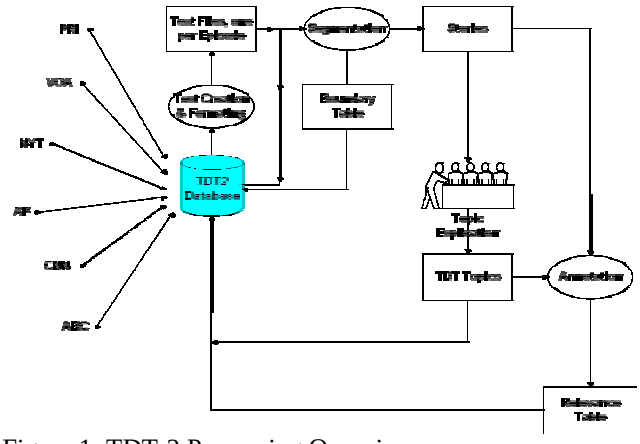

Figure 1: TDT-2 Processing Overview

The data for the TDT-2 corpus came from newswire, television and radio, all sampled on a daily basis to yield, on average, over three hundred news stories per day. Nearly half of those stories were drawn from the audio content of broadcasts, which comprised, on average, about five hours of digital recordings per day. Accumulated over

the 180 days of collection, the net result was over 54,000 stories and 634 hours of recorded audio. LDC also has a video tape archive of all the television sampling.

3. DATA ACQUISITION

The six sources collected were: Associated Press's *World Stream*, the New York Times news service, Public Radio International's *The World*, Voice of America's English news, ABC's *World News Tonight* and CNN's *Headline News*. The broadcast radio and television sources were captured from the broadcast airwaves, from cable or, in the case of VOA, from satellite receiver and the worldwide web. All of the audio sources were captured to disk as 16bit, 16KHz NIST SPHERE files. (For the first two months of collection, VOA broadcasts were only available as 16 bit, 11KHz files gathered from VOA's web site, but these proved unsuitable for the ASR technology used in the project; as a result, only the text transcriptions are included for the January and February VOA programs.) The newswire text came to LDC via 24-hour dedicated modem feeds from the services; four sets of about 20 stories each were selected each day from each service for inclusion in the corpus, yielding over 1100 newswire stories per week. The PRI and ABC programs were recorded as often as they aired: 5 hours per week for PRI, 3.5 hours per week for ABC. CNN Headline News, a 24-hour broadcast news format, was recorded for 30 minutes 3 or 4 times per day, yielding 12 hours per week. Figure 2 summarizes TDT-2 sources, their quantities and the methods used to collect them.

	Program	Weekly Volume	Audio Source	Text Source
AP	World Stream	560 stories	NA	Modem
NYT	News Service	560 stories	NA	Modem
PRI	The World	5 hours	broadcast	Transcript
VOA	English News*	12 hours	satellite, web	Transcript web
ABC	World News Tonight	3.5 hours	broadcast	Closed caption FDCH
CNN	Headline News	12 hours	broadcast	Closed caption

Figure 2: Six TDT-2 sources showing weekly volume and collection methods. *VOA modified its programming during the collection; the name of the program we recorded also changed.

4. SEGMENTATION

LDC staff produced the reference segmentation of the corpus against which the evaluation systems would be scored. (Newswire data is already divided into stories, however in some cases, due to transmission or other errors, the data would contain story fragments that needed to be concatenated into whole stories.) For the most part, audio segmentation of the broadcast news sources was a two-pass procedure. In the first 2*RealTime pass, human annotation staff listened to the broadcast audio with the audio waveform and text on display, inserted story boundaries and identified non-story segments. During a 1*RealTime second pass, annotators confirmed or adjusted existing story boundaries. The story boundaries are present in the reference text and duplicated in a boundary table. In the ASR output and tokenized text streams, story boundaries are removed and systems need to consult the boundary tables.

5. TOPIC EXPLICATION

In TDT-2, LDC defined 100 topics based upon a stratified random sample of the six news sources collected January through June of 1998. The sampling gave each month of data from each source an equal chance of being represented. Within any month/source, stories were selected at random. In some cases the randomly selected story was a list of sports scores or stock market quotes and was therefore rejected. If one could determine that the story was about an event in the news it would be used to define a topic.

TDT-2 topics are based on an assumption that news stories are about events. A TDT-2 event *is an activity that happens at a specific place and time and all of its necessary causes and unavoidable consequences*. On February 3, 1998, when a U.S. Marine jet sliced a funicular cable, the car's crash to earth and the subsequent injuries and rescue efforts were all unavoidable consequences and thus part of the event. Rules of interpretation specify the scope of related events also to be considered part of the same topic. In this example, stories about the investigation, the Marine pilot, the repercussions for his unit, the victim's families and their quest for justice were all on topic.

In TDT-2, topic definition was a collaborative process with annotators negotiating the scope of a topic among themselves, the sponsors and the research sites. The randomly selected story was often neither the best nor even a good representative of the seminal event. Annotators, therefore, researched each event elsewhere in the news. Recognizing that TDT-2 topics need to retain some fluidity in response to changes in the real world, annotation fed back into topic definition in a continuous loop so that as we encountered new stories we reevaluated and often modified the bounds of the topic. Rules of interpretation and topics in their entirety on LDC's TDT-2 web page referenced in [3].

6. TOPIC LABELING

The lion's share of building the TDT-2 corpus was devoted to topic labeling. Annotation staff worked with the daily news files reading each story and deciding how it related to the 100 TDT-2 topics. For each topic-story pair the annotator could render a decision of *yes*, *brief* or *no* meaning that story was about the topic, mentioned the topic only briefly or was not about the topic. Any mention of a topic warranted a label of at least *brief*. Stories that were primarily about something else but discussed the target topic in more than 10% of their volume were labeled *yes*. This was in keeping with the premise that news stories could be about more than one topic. For TDT-2, LDC staff made five passes over the data. In each pass, staff labeled a story with respect to 20 topics on average. A custom interface stepped annotation staff through the stories, recorded annotator's progress and logged their decision into an Oracle database. Figure 3 shows the yield of TDT-2 annotation with topics on the x-axis and number of *yes* stories per topic on the log-scaled y-axis.

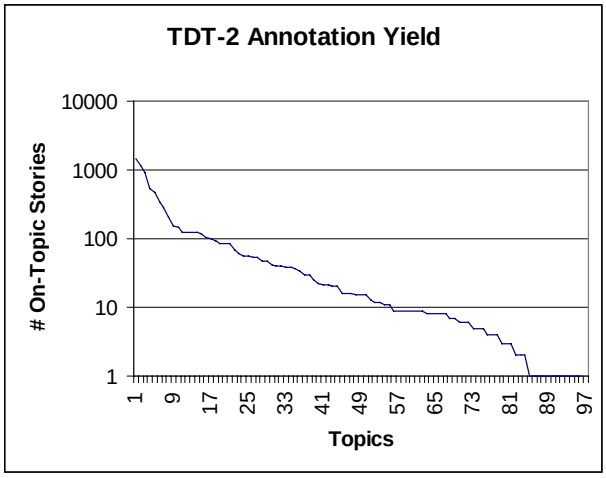

Figure 3: TDT-2 annotation yield

A very different approach to topic annotation involves labeling pairs of stories with regard to whether they discuss the same topic. This approach, referred to as "story-story linking" amounts to a manual judgement of the clustering of stories, and so does not require the initial selection and definition of target topics. Although pair-wise comparisons among tens of thousands of stories are impractical, automatic clustering techniques can eliminate a vast majority of unnecessary comparisons making this a viable annotation method. TDT-3 will use story-story linking.

7. QUALITY ASSURANCE

In TDT-2, LDC implemented three types of quality assurance: precision QC, recall QC and discrepancy QC. In precision QC, senior annotators review all stories labeled as on-topic for a particular topic to identify false alarms. In recall QC, senior annotators use a search engine to generate a relevance-ordered projection of the corpus with respect to a single topic. Queries used in recall QC can be the seminal article, a list of miscellaneous keywords, the topic explication itself or the union of all stories labeled as related to the target topic or a subset thereof. In discrepancy QC, senior annotators adjudicate any cases in which two or more annotators disagree as to the labeling of a specific story. In the later stages of TDT-2 when research sites submitted dry run and evaluation results, LDC also adjudicated all disagreements between the human annotators on the one hand and the tracking systems on the other. To support consistency studies, LDC's local copy of the database includes all judgments. However, the reference version of the corpus contains only the adjudicated results.

8. CONSISTENCY STUDIES

Since TDT-2 was something of a unique project, we also wanted to determine how consistent human annotators could be with respect to the tasks of segmentation and topic labeling. LDC assigned segmentation and topic labeling tasks to the annotation staff so that at least 5% of the work was duplicative. Initial consistency studies were somewhat suspect since staff were aware of the nature of the experiment. By the end of TDT-2, we had implemented a double-blind method of task assignment. Among TDT-2 participants consistency studies were alternately viewed as a way to measure the inherent variability of the task or a way to improve human performance. In acknowledgement of the second view, LDC annotation staff worked to improve consistency by discussing judgements via e-mail and during weekly meetings, by keeping informed on current events and by collaborating on topic research. The physical arrangement of annotation staff encouraged collaboration and consistency scores were generally good. The kappa statistic was used to measure consistency of human annotation. Where a kappa of .6 indicates marginal consistency and .7 measures good consistency, kappa scores on TDT-2 were routinely in the range of .7 to .9.

9. TDT-2 CORPUS

For purposes of system evaluation, the TDT-2 corpus is divided into three segments. Although the corpus contains the cross-product of six months of data collection labeled for each of 5 topic lists, only a subset was used during the evaluation. January and February stories labeled against topics drawn from the same period comprise the Training data. The Development-Test data includes March and April stories labeled against topics from these two months. The remaining topics, from Many and June were used with

May and June stories to create the Evaluation data. The reference corpus contains all three TDT-2 data sets plus the off-diagonal material as shown in Figure 4..

Topics from	Stories from						
		Jan	Feb	Mar	Apr	May	Jun
	Jan-Feb	Training Data					
	Mar-Apr			Development -Test Data			
	May-Jun					Evaluation Data	

Figure 4: Organization of the TDT-2 Corpus.

10. OTHER TDT DATA

As part on an ongoing DARPA research project, TDT-2 is one of a series of related corpora. Its predecessor, the TDT PILOT corpus was a sampling of two sources (one newswire and an assortment of CNN programs) on a daily basis over a year. For the 1999 TDT-3 project, LDC has collected Mandarin data from three sources: Xinhua News Agency, Voice of American Mandarin News and the Zaobao Worldwide Web site. LDC has distribution rights for the first two sources and is negotiating rights to distribute the third. LDC has been collecting the Mandarin data since the first quarter of 1998. In TDT-3, researchers will have access to all of the TDT-2 English from January to June 1998 plus the Mandarin from the same period to use as Training and Development-Test data. The TDT-3 Evaluation set will include all of the TDT-2 English sources, the three Mandarin sources and two new English sources: NBC's *Nightly News* and MSNBC's *News with Brian Williams* all collected between October and December 1998. Figure 5 summarizes the use of data sources in TDT-2 and TDT-3 projects.

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
APW	2	2	2	2	2	2	-	-	-	3	3	3
NYT	2	2	2	2	2	2	-	-	-	3	3	3
PRI	2	2	2	2	2	2	-	-	-	3	3	3
VOA Eng.	2	2	2	2	2	2	-	-	-	3	3	3
CNN	2	2	2	2	2	2	-	-	-	3	3	3
ABC	2	2	2	2	2	2	-	-	-	3	3	3
NBC								-	-	3	3	3
MSNBC										3	3	3
Xinhua	3	3	3	3	3	3	-	-	-	3	3	3
VOA Mand.			3	3	3	3	-	-	-	3	3	3
Zaobao		3	3	3	3	3	-	-	-	3	3	3

Figure 5: Data sources for TDT-2 and TDT-3 projects 2=TDT-2 data, 3=TDT-3 data. Dash indicates data collected but not yet slated for use in a specific project

REFERENCES

[1] Doddington, George, The Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan: Overview & Perspective, *Proceedings of the Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, February 1998.

[2] Doddington, George, The Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan http://www.nist.gov/speech/tdt98/doc/tdt2.eval.plan.98.v3.7.pdf

[3] Linguistic Data Consortium, Topic Detection and Tracking, http://morph.ldc.upenn.edu/TDT/

[4] National Institute for Standards and Technology, 1998 Topic Detection and Tracking Project (TDT-2), http://www.nist.gov/speech/tdt98/tdt98.htm

[5] Wayne, Charles, Topic Detection & Tracking: A Case Study in Corpus Creation & Evaluation Methodologies, *Proceedings of the First International Conference on Language Resource and Evaluation*, Granada, Spain, May 1998.

[6] Wayne, Charles, Topic Detection and Tracking (TDT): Overview & Perspective, *Proceedings of the Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, February 1998