THE TDT-3 TEXT AND SPEECH CORPUS

David Graff, Chris Cieri, Stephanie Strassel, Nii Martey

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104

ABSTRACT

The TDT-3 Text and Speech Corpus expands on previous phases of Topic Detection and Tracking data collections, by increasing the number of news sources being sampled, by including Mandarin Chinese as well as English news data, and by introducing new forms of topic annotation. In order to satisfy the specific data and annotation requirements of the TDT-3 Evaluation Plan[1], the LDC refined and supplemented the methods that had been used in TDT-2 corpus development[2]. There were significant changes and improvements in the process of selecting and defining target topics, in the procedures for quality assurance applied to both data content and annotations, and in the organization of the delivered corpus. In addition, the LDC created or acquired a range of resources to support research in cross-language information retrieval. These included the addition of a Mandarin Chinese component to the TDT-2 Text and Speech Corpus, the collection of a large body of Chinese-English parallel text, and adaptation of Chinese-to-English and English-to-Chinese glossing lexicons. All the resources that we have developed for use by the participants in the TDT-3 Evaluation are being added to the LDC's catalog of corpora for general availability.

1. INTRODUCTION

Phase 3 of the research program on Topic Detection and Tracking (TDT-3) builds directly on the foundation established in earlier phases of TDT: a dense, continuous sampling of news sources over an extended period of time provides an effective test bed for the development of automatic information retrieval (IR). TDT-3 extends the scope of the previous phase in four ways: (1) it includes more English news sources (NBC and MSNBC were added), (2) it covers a longer period of time (three months instead of two), (3) there are more topics to detect and track (60 instead of 30), and (4) there are two distinct languages in the sample (three sources of Mandarin Chinese news were added). The following sections describe the corpus content, the methods applied in annotating the corpus, supporting materials that were made available to aid the development of cross-language IR, and our post-evaluation efforts to refine the manual topic annotations by adjudicating the various TDT system outputs that were submitted to NIST for benchmark scoring.

2. SUMMARY OF TDT-3 DATA CONTENT AND ANNOTATIONS

All the data for TDT-3 were collected from broadcast and newswire sources between October 1 and December 31, 1998. The data sources included four newswire services (Xinhua and Zaobao for Mandarin Chinese, Associated Press and New York Times for English), four television networks (ABC, CNN, NBC and MSNBC, all in English), and two radio broadcasters (Public Radio International in English, Voice of America in English and Mandarin). Put another way, there were 8 news sources in English (2 newswire, 2 radio, 4 television), and 3 in Mandarin (2 newswire and 1 radio). All but two of these sources (NBC and MSNBC) are also present in the TDT-2 corpus, spanning January 1 through June 30, 1998 (but note that the three Mandarin sources were added to that corpus just before we began production of TDT-3; cf. Section 4.1 below).

Data sources were sampled daily over the three-month collection period (except for PRI and MSNBC, whose news programs were broadcast only on weekdays), and some were sampled multiple times per day. Table 1 shows the sampling regimen for each source, the total number of samples successfully collected, and the total yield of news stories. Note that the numbers given for sample size and samples per week represent intended targets. For newswire, the actual number of stories per sample varies between 15 and 30, and a few samples were lost due to problems with modem connections. For broadcasts, there were very few failures in the sampling, typically due to unannounced schedule changes.

Source	Sample size	Samples per week	Total Samples	Total Stories
ABC	30 min.	7	76	1012
CNN	30 min.	12	349	9003
MSNBC	60 min.	5	51	683
NBC	30 min.	7	87	846
PRI	60 min.	5	65	1575
VOA_E	60 min.	12	103	3948
VOA_M	60 min.	12	121	3371
Audio sub-total		60	852	20438
APW	20 stories	28	360	7338
NYT	20 stories	28	347	6871
XIN	20 stories	21	217	5153
ZBN	20 stories	14	180	3817
Text sub-total		91	104	23179
Grand Total		151	1956	43617

Table 1: Sampling and content of TDT-3 data collection

All television sources were broadcast with closed captioning, which was converted to text and captured electronically with the audio, to become the "reference" transcription for these sources. Professional transcription services (Federal Document Clearing House for English, Philadelphia Chinese News Service for Mandarin) were enlisted to produce rapid, lexically correct transcriptions for all radio samples.

All the audio samples and their associated transcripts were manually reviewed using a specialized interface to identify all news story

boundaries; time-stamped tags were inserted into the transcripts to mark the beginning of each story and each non-news segment. The non-news segments included introductory or "preview" announcements at the beginning of a broadcast, commercial breaks, musical interludes, and any other extended portion of the audio that did not contain a news report, such as casual banter among newscasters.

Automatic speech recognition (ASR) output was produced for all audio samples; Dragon Systems provided the ASR output for the Mandarin data, and NIST used the BBN Byblos system to produce ASR output for English.

Both broadcast reference transcripts and newswire samples were transformed into an SGML format modeled closely on TIPSTER corpora. For broadcast data, the SGML files retain all transcript content and the story boundaries (rendered using "<DOC>" tags) as marked by annotators; for newswires, only usable news stories are retained in the SGML files. In all cases, each "<DOC>" unit is assigned a unique "<DOCNO>" index value and a "<DATE_TIME>" tag identifying when the story was transmitted.

Once the SGML files were all in place, 60 target topics were selected in such a way as to assure a minimum of four on-topic stories in both languages for each topic. All news stories in the collection were manually judged for relevance to each of the 60 topics, and all stories that were found to be fully or briefly relevant to one or more target topics are presented in a "topic table" file.

In order to support the First Story Detection and Story Link Detection tasks that were defined for the TDT-3 Evaluation, an additional 120 stories were selected randomly from the English data, such that they did not relate to any of the 60 main topics. For each of these 120 "seed" stories, special annotation was done to locate the first story in the collection to describe the event discussed in the seed story. Also, each seed story was combined with 180 other stories from the English data to produce over 21,000 story pairs; each pairing of stories was presented to annotators who would decide, for each pairing, whether the two stories discussed the same topic. The corpus includes "first-story" and "story-link" tables that provide the results of these annotations.

The published version of the corpus contains the following components:

1. The English and Mandarin audio data, in SPHERE compressed format, on 68 CD-ROMs (13 for Mandarin, 55 for English).
2. The TIPSTER-style SGML data, as described above, containing the reference text for all samples; in the Mandarin data, the reference text is not segmented into words, though both newswire and transcripts include some ASCII alphanumeric strings in the stream GB-encoded characters.
3. A tokenized version of the reference text for all samples; for English data, the tokenization assigns a sequential index number to each space-separated string in the reference text; for Mandarin, it assigns a sequential index to each 2-byte GB character (including punctuation) and to each string of contiguous ASCII characters (excluding whitespace).
4. The ASR output for all audio samples, using the same format as the tokenized reference data, but with the following differences:
   - the Mandarin ASR system provides word segmentation in its output (the indexed tokens are variable-length strings), and produces only GB-encoded words (no ASCII alphanumerics);
   - both English and Mandarin ASR systems provide the start time and duration for each word, and these are provided along with the sequential index number;
   - the Dragon Systems ASR also provides "speaker cluster" and "confidence" data, which are included for each word;
   - the reference text tokenization retains all the bracketing and punctuation of the original texts, whereas the ASR systems produce no punctuation at all;
   - while the tokenized reference text is all stored together in a single directory, the ASR outputs are separated according to the system that produced them – the Mandarin ASR data (from the Dragon system) is separated from the English ASR data (from the BBN system).
5. The output of Systran Chinese-to-English machine translation (MT) operating on the tokenized Mandarin reference texts.
6. The output of Systran MT operating on the Mandarin ASR text.[1]
7. Boundary table files for each set of token-stream files listed in 3 - 6 above; the token-stream files do not preserve story boundary information, so boundary tables are provided to relate the story units (by means of their "DOCNO" indexes) to the ranges of token indexes that constitute the stories.

3. THE TDT-3 ANNOTATION TASKS

The majority of effort in producing the TDT-3 corpus was devoted to the same tasks that were applied in TDT-2: transcription of radio programs, determination and time-stamping of story boundaries in all audio data, and exhaustive labeling of all 43,600 news stories with respect to the 60 target topics.

The main topic labeling was done in three passes; in each pass, annotators were presented with one story at a time while a listing of 20 topics was kept visible on the display screen next to the story. The annotator, who had previously studied the descriptions of those 20 topics, would select a check box next to one or more topics if they were discussed in the story; the topic could be marked as "YES" if the story discussed it primarily or at length, or it could be marked as "BRIEF" if the story included a short mention of the topic. The annotator then hit a button to move on to the next story, and the results for the current story were entered into a database table.

3.1. Selection and Definition of Main Topics

A central requirement for the corpus development effort was to assure that for each topic there would be at least four on-topic stories both in the English data and in the Mandarin data. This involved a closely coordinated effort by senior annotators to determine appropriate search strategies. About 500 potential topics were considered, derived from seed stories that had been selected from all sources over

[1] The version of Systran that we used depended crucially on the presence of punctuation in the Chinese text provided as input; since the ASR output had no punctuation, we chose to insert a GB "period" character whenever there was a gap of 0.5 sec or more between the time-stamps of adjacent words. Also, for both reference and ASR text, there were small portions of Mandarin in many files that Systran was unable to translate; these remained as GB-encoded strings in the MT output files.

the full span of the 3-month corpus, in order to arrive at the 60 topics that met the requirement.

A specialized interface was developed to support this task, which involved using a version of the UPenn TDT search engine on both the Mandarin and English collections, and providing methods to select a seed story in one language, use that as a query to find related stories in the same language, and perform a suitable keyword search for related stories in the other language. Any related story that was found in the other language could then be used in turn as a query to find additional hits on the same topic. If at least four stories were found in each language for a given event-based topic, the user provided a topic title and a brief description; later, the found stories and other available resources were used to conduct research on the topic and create a detailed topic definition with background facts.

3.2. Procedures for Quality Control

A number of lessons learned from the TDT-2 corpus development were put to use in the creation of TDT-3. Also, we were in a better position to apply quality control measures because the entire text corpus was in place and fairly stable before annotation began. Among the efforts applied comprehensively to TDT-3 data are the following:

- All story segmentation was second-passed by a separate annotator, and after segmentation, all boundaries were checked algorithmically to look for cases of incorrect time stamps and missing end-of-story marks.
- About 5% of audio files went through a second independent segmentation, to measure inter-annotator consistency on this task.
- Work assignments for topic labeling were made automatically, and included a double-blind assignment of 8% of sample files to two different annotators to measure inter-annotator consistency.
- During topic labeling, where annotators had the option to reject stories due to errors in formatting or segmentation, questions and examples of rejected stories were announced and discussed frequently; all rejection judgments were reviewed by senior personnel to determine appropriate remedies for the affected stories: to either correct the segmentation (for audio sources), fix errors in file format, or (in newswire sources) eliminate an unusable story from the corpus.
- All stories judged to be on-topic were reviewed by senior annotators to assure the accuracy and consistency of the on-topic marks for all topics.
- For each of the 60 topics, an extra effort was made to look for potential misses in that portion of the corpus that included and/or led up to the first four on-topic stories.

During the final stages of preparing the corpus for delivery, an extensive set of formatting and consistency checks was executed over the entire corpus to verify full compliance with the data format specifications, and assure the coherence of cross-references between various files. Up to this point, all annotation had been based on extensive relational database tables and on the TIPSTER-style SGML data (and, for audio sources, the prior forms of text data that were transformed into SGML). At the final stage, with the annotation essentially done, much of the final quality control was imposed by a set of scripts that generated the token stream files and associated boundary tables from the SGML, ASR and MT output files, and distilled the exhaustive topic label database records into the published form of the main topic-relevance table.

In addition, a separate set of scripts was developed, independently of the corpus-creation tools, to verify that the formatting specifications and cross-reference relations were fully satisfied.

3.3. New Annotation Procedures

Two new annotation tasks were introduced for the English component of TDT3: first-story detection (FSD) and story-story links (SSL). These were applied to the 60 main topics that had been defined for full annotation, and also to an additional 120 topics selected from the English data.

For the 120 additional topics, seed stories were selected at random, and manually reviewed to ensure that (a) none of these seeds were related to the 60 main topics and (b) each of the seed stories did in fact discuss an event-based topic comparable in nature to the main topics (excluding, for example, "human interest" stories, formulaic stock market summaries, broad commentary, etc).

The FSD annotation used the same methods (i.e. the same user interface) as the main topic-selection task. In fact, the full annotation of the main 60 topics satisfied the requirements for FSD on those topics. The additional 120 seed stories were used as queries against the portion of the English corpus that preceded each seed; the annotators repeated a cycle of reviewing the returns from search engine, tagging earlier on-topic stories, and using those stories to supplement the query, until they could find no earlier mention of the topic.

The SSL annotation task involved creating a list of 120 "comparison" stories for each of 180 seed stories. (For each of the 60 main topics, one of the initial four on-topic English stories discovered during the topic selection phase was selected as sole seed story for that topic.) The selection of comparison stories involved two different methods of sampling from the English collection:

- For each seed story, 60 of the compare stories were selected from a relevance-ranked list that was output by the UPenn search engine when the seed was used as a query against the English corpus.
- The other 60 stories were selected randomly from the corpus (excluding stories already present in the ranked list), using a special sampling method provided by Jon Fiscus at NIST: the sampling was weighted according to the temporal distance between the selected story and the seed, in such a way that the overall distribution of the random selections would be similar to the relevance-ranked selections in terms of their chronological spread around each seed story.

The combination of 180 seed stories with 120 compare stories per seed yielded 21,600 story pairs. For annotation, users were presented with a display showing the seed story on one side and one compare story at a time on the other side. There was no display or creation of a topic title or description – that is, no explicit naming or specification of a topic. The user simply read both stories and decided whether they "discussed the same topic." The compare stories were presented in random order while the seed story remained on display throughout the session.

4. SUPPORTING RESOURCES

4.1. Supplements and Changes to TDT-2

In order to provide suitable training and development test data for the TDT-3 test set, the LDC produced a Mandarin supplement to the TDT-2 corpus, spanning January 1 through June 30, 1998, and using the same three Mandarin sources (VOA, Xinhua, Zaobao). The sampling was not as consistent over this period as it was in TDT-3: only Xinhua was present consistently over the entire six months; VOA and Zaobao began in mid-February, and an unexpected schedule change in the VOA Mandarin broadcasts went undetected from early April through the end of June, causing the yield of usable news content to be lower than intended (a total of about 57 hours over the 4-1/2 months of collection). Table 2 shows the total number of samples and stories in TDT-2 Mandarin.

Source	Total Samples	Total Stories
VOA_M	177	2265
XIN	484	11286
ZBN	250	5170
Total	911	18721

Table 2: Sampling and content of TDT-2 Mandarin supplement

As with TDT-3, Dragon Systems provided ASR output for the VOA audio data, and Systran was used to generate automatic English translations of the reference and ASR text.

For topic annotation, we reviewed the Mandarin data to select 20 topics from among the original 100 TDT-2 topics (which had been selected from English data in 1998); the essential criteria for determining these topics were that there be at least 4 on-topic stories in both the English and Mandarin TDT-2 collections. Once these 20 topics were designated, the established methods for topic annotation were applied over the Mandarin data.

The original corpus structure that had been established for TDT-2 when it was created in 1998 proved to be inadequate for the quantity and diversity of files that resulted when the English and Mandarin collections were combined. After an initial release of the combined corpus in June using that original corpus structure, we consulted with NIST and some of the TDT researchers to work out an improved design that would readily support further expansion beyond TDT-3. Both TDT-2 and TDT-3 releases are now organized according to this new structure.

4.2. Parallel Text

In 1998 the LDC began a project to seek out parallel text on web sites around the world. One of the major resources discovered during this effort was the Government of the Hong Kong Special Administrative Region (HKSAR), which has produced sentence-aligned parallel text in English and Mandarin for the HKASR legal code. This collection alone contains about 6.3 million words of English, and the corresponding amount of Chinese (about 11.5 million characters). More recently, we have been collecting parallel-text news data from HKSAR, and the Honk Kong "Hansards" (parliamentary proceedings), on a continuing basis; to date, these two sources have each yielded an additional 9.3 million English words and the corresponding Chinese, and they are accumulating at a rate of about 0.8 million words per month. These sources all use Big-5 character encoding for the Mandarin material.

4.3. Bilingual Word lists

In addition to parallel text, the LDC sought out sources for bilingual word lists – both English-to-Chinese and Chinese-to-English glosses – and we supplemented the available resources with some development work in-house. Since our intention was to provide as much as possible as quickly as possible, there was very little refinement of the head word inventory or the glosses. Each bilingual word list contained over 110,000 head words with, on average, 1.5 to 2 glosses each.

5. ADJUDICATION OF MAIN TOPIC ANNOTATIONS

The TDT-3 annotation project differed from TDT-2 in one important aspect of quality control. For the TDT-2 project in 1998, we had attempted to do a thorough check of "recall" accuracy for the standard topic annotation; this involved going over the data, one topic at a time, using a version of the UPenn search engine to locate stories that had been missed during the main annotation effort. This was a time-consuming and difficult task, which would become impractical for the expanded scope of TDT-3. After the TDT-2 evaluation, NIST provided LDC with the system outputs on the topic tracking task, and we reviewed all cases where a system had reported an on-topic story that we had not labeled as such. As a result of this review, we uncovered 246 on-topic stories that had not been caught by the earlier recall check.

After consultation with NIST, it was agreed that we would not attempt to do a thorough recall check on the TDT-3 data. Instead, NIST would again provide LDC with the various TDT system outputs, and we would review the cases where those outputs differed from the original LDC annotations.

Because the scale of the TDT-3 test was larger than TDT-2 in every respect (more systems, more stories, more topics), we were not able to review all discrepancies. Altogether, there were 4,991 topic-story relations that were labeled on-topic by LDC but were missed by one or more systems, and 108,274 relations that one or more systems chose as on-topic but were not labeled as such by LDC. Adjudication requires reviewing each disputed topic-story relation individually, and a complete review of over 113,000 cases would involve an effort nearly as large as the original topic annotation (43,617 stories, each read three times).

We decided to review only those cases where a majority of systems (4 or more of the 7 that submitted tracking results) made the same "error" relative to the LDC annotation. This reduced the number of cases by an order of magnitude. But even with this reduction, there was one topic (#3037) that produced an excessive number of apparent false alarms (1,202 stories hit by four or more systems, almost three times more than the next most difficult topic); interestingly, 4 systems reported no misses on this topic. To make the review of this topic manageable, we further reduced its adjudication set to include only those stories whose ranking by a majority of systems were within the overall range of the known on-topic stories. This reduced the number of false alarm reviews for topic #3037 by about half. In all, 10448 topic-story relations were reviewed. The overall changes to the topic relevance judgments are summarized in Table 3, broken down according to the amount of consensus among systems.

If we were to treat the original LDC annotations as a "system" and score it in the TDT-3 tracking task, the bottom-line score (normalized topic-weighted Ctrack) would be 0.0938, which turns out to be comparable to the best TDT-3 system scores for this task (in fact slightly worse than a couple of the BBN submissions). Of course, the bottom line scores are not that similar in nature. The probability of false alarms in the original LDC annotations was virtually zero, whereas false alarms contributed heavily to the system tracking scores, despite being valued at one-tenth the cost of misses.

In any event, this comparison of exhaustive manual annotation against automatic topic tracking makes a strong case for the suitability of using existing IR technology to reduce the cost and improve the quality of topic annotation on large corpora.

6. CONCLUSION

The creation of the TDT-3 corpus has benefited in significant ways from the experience gained in previous phases of the project. The internal consistency of the data and annotations marks a significant improvement over the original version of TDT-2, and the overall quality of the corpus has met a higher standard.

The novel demands of the TDT-3 Evaluation Plan have led us to develop a more robust, capable and economical infrastructure for the collection and annotation of large multi-modal corpora. In responding to the special needs of researchers in this project, we have discovered and made available a wider range of important linguistic resources, including parallel text and bilingual dictionaries. We expect that the range of potential uses for TDT-3 and its related resources will quickly expand far beyond the scope of the project that brought it about.

References

1. Doddington, George, "TDT3 Evaluation Specification Version 2.7," *http://www.itl.nist.gov/iaui/894.01/tdt3/tdt3.htm*
2. Cieri, C., Graff, D., Liberman, M., Martey, N., Strassel, S., "The TDT-2 Text and Speech Corpus," 1999 DARPA Broadcast News Workshop, *http://www.ldc.upenn.edu/Papers/DARPA_BN1999/DARPA_BN1999.html*

# Systems agreeing	F.A.'s reviewed	No > Yes	No > Brief	Misses reviewed	Yes > No	Yes > Brief
4	5547	130	107	572	3	17
5	2702	143	101	330	1	9
6	1018	149	102	206	3	10
7	7	7	0	66	3	9
Total	9274	429	310	1174	10	45

Table 3: Summary of topic label changes as a result of adjudication