Large, Multilingual, Broadcast News Corpora For Cooperative Research in Topic Detection And Tracking: The TDT-2 and TDT-3 Corpus Efforts

Christopher Cieri, David Graff, Mark Liberman, Nii Martey and Stephanie Strassel

Linguistic Data Consortium, University of Pennsylvania
Philadelphia, Pennsylvania, USA
{ccieri, graff, myl, nmartey, strassel}@ldc.upenn.edu

Abstract

This paper describes the creation and content two corpora, TDT-2 and TDT-3, created for the DARPA sponsored Topic Detection and Tracking project. The research goal in the TDT program is to create the core technology of a news understanding system that can process multilingual news content categorizing individual stories according to the topic(s) they describe. The research tasks include segmentation of the news streams into individual stories, detection of new topics, identification of the first story to discuss any topic, tracking of all stories on selected topics and detection of links among stories discussing the same topics. The corpora contain English and Chinese broadcast television and radio, newswires, and text from web sites devoted to news. For each source there are texts or text intermediaries; for the broadcast stories the audio is also available. Each broadcast is also segment to show start and end times of all news stories. LDC staff have defined news topics in the corpora and annotated each story to indicate its relevance to each topic. The end products are massive, richly annotated corpora available to support research and development in information retrieval, topic detection and tracking, information extraction message understanding directly or after additional annotation. This paper will describe the corpora created for TDT including sources, collection processes, formats, topic selection and definition, annotation, distribution and project management for large corpora.

1. Introduction

This paper describes the TDT-2 and TDT-3 corpora that were created to support the DARPA program in Topic Detection and Tracking and are now available for general use. The research goal in the TDT program is to create the core technology of a news understanding system that can process multilingual news content categorizing individual stories according to the topic(s) they discuss

The DARPA-sponsored research program in Topic Detection and Tracking (TDT) began with a pilot study in 1997. In 1998, the program expanded to include new research sites and enlist the U.S. National Institute of Standards and Technology (NIST) to perform technology evaluation and Linguistic Data Consortium (LDC) to create the corpus. TDT has now finished its third phase. Charles Wayne's (2000) paper, also presented at this LREC, will discuss the overall goals and results of the project in greater detail. The interested reader is also encouraged to visit www.nist.gov/speech/tdt3/tdt3.htm for a discussion of the evaluation metrics and final results.

The program's orientation toward the linguistic content of reported news shapes the corpus. For the TDT research tasks of:

·segmentation - divide news stream into individual stories

·topic detection - identify new topics in the news

·topic tracking - identify all stories that discus a selected topic

·first story detection - identify the first story to discuss a selected topic

·story link detection - identify all pairs of stories that have any topic in common

sites may rely only on the raw content provided; formatting, paragraph breaks and story headers as may appear in newswire are not available to the research sites during the evaluation. The corpora accommodate these tasks but also future projects by providing multiple versions of the raw data files. This will be discussed further below.

2. Data

The TDT corpora are multi-modal. TDT-2 contains daily samplings from January through June, 1998 of two television broadcasts, three radio programs, three newswires and one web site devoted to news. They are: ABC World News Tonight, CNN Headline News, Public Radio International's The World, Voice of America English & Mandarin news radio, newswires from the Associated Press, New York Times and Xinhua services and the web pages of the Singapore based news agency Zaobao. For the broadcast sources, LDC recorded the entire half-hour or hour broadcast; for the newswire sources, LDC sampled approximately 80 stories per day. TDT-2 contains over 600 hours of audio yielding 53,620 English and 18,721 Chinese stories.

TDT-3 corpus adds two English sources: NBC Nightly News and MSNBC The News with Brian Williams and extends the collection from October through December of 1998. There are 475 hours of English and 121 hours of Mandarin audio in TDT-3 yielding 31276 English and 12,341 Chinese stories.

Anticipating future projects, LDC collected Voice of America Spanish radio broadcasts and newswire from El Norte's news service during the same period and has continued the collection through 1999 adding CCTV Mandarin television broadcasts and Spanish broadcasts from ECO and Univision. The complete collection starts with six sources in January 1998 and continues, adding sources, so that there are 16 sources from August through December 1999. In the process of creating the TDT corpora, LDC collected over two years of broadcast and newswire data, more than half of which is held in reserve for future use.

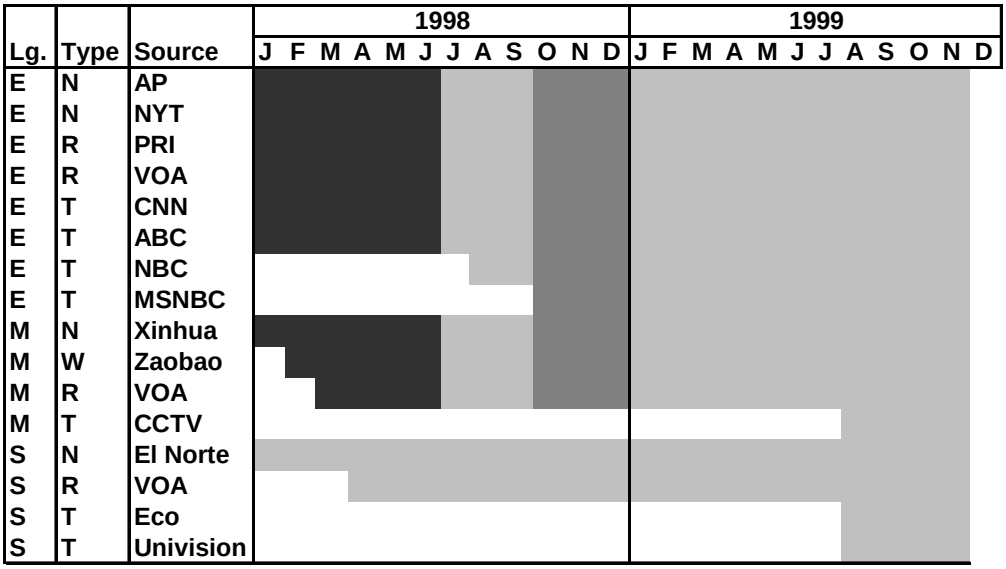

Figure 1: Data Sources in TDT. Use: ■ = used in TDT-2, ■ = used in TDT-3, ■ = not yet used.
Language: E=English, M=Mandarin, S=Spanish Type: N=Newswire, R=Radio, T=Television, W=web site

3. TDT Topics

Event and *topic* are important concept in TDT annotation. A TDT *event* is **a specific thing that happens at a specific time and place along with its necessary prerequisites and consequences**. For example, in the case of the China Airlines Crash, the crash of the plane and the resulting injuries and fatalities are considered part of the same event. A TDT *topic* is then a collection of **related events and activities**. To render more consistent the judgments about what constitutes "related", annotators use a set of *rules of interpretation*. These rules state, for each type of event: crimes, natural disasters, scientific discoveries, scandals, etc, what other events may be considered related. TDT-2 topics include: the Clinton-Lewinsky scandal, the Winter Olympics in Nagano, the 1998 elections in the Phillipines, the Karla Faye tucker trial and the Pope's visit to Cuba. The reader is encouraged to visit www.ldc.upenn.edu/Projects/TDT for complete topic lists.

LDC senior annotators define topics by selecting stories at random such that each month's collection for each source has an equal a priori probability of contributing a topic. Reading the randomly selecting story, LDC staffers attempt to identify the stories seminal event. Some news reporting such as sports scores, exchange rates and discussions of trends may lack a discernible seminal event. In these cases the story is rejected. Because the percentage of such material varies by source, in the end some sources contribute fewer actual topics. When a seminal event is identified, LDC staff create a topic explication that provides the "what", "where" and "how" of the seminal event. The appropriate rule of interpretation is then consulted to determine which other types of events and activities may be considered related for that event type. The *topic explication* codifies this information for future annotators. An example topic explication follows. During the evaluation of the tracking and first story detection tasks, sites learn the "definition" of a topic not from the topic explication but from (typically four) on-topic "training" stories. The delivered corpora contain both the topic explications and the training stories.

Leonid Meteor Shower 中文

Seminal Event

WHAT: Earth passes through comet Tempel-Tuttle's trail of debris, creating a spectacular light show in Earth's atmosphere.

WHERE: The meteor shower is visible in Europe and Asia.

WHEN: Late October through November 1998; the shower peaks on November 16-17.

Topic Explication

The 1998 Leonid Meteor Shower was particularly strong, and scientists from around the world gathered to watch the display. Although the shower was expected to peak over Eastern Asia, the best viewing actually occurred in Europe. On topic: Stories covering scientists' forecasts for the shower; reports on its observation; concerns over the possibility of the meteors damaging artificial satellites (which proved to be unfounded).

Rule of Interpretation

Rule 7: Science and Discovery News

Figure 2: A TDT-3 topic explication showing seminal event, topic explication and relevant rule of interpretation.

4. Annotation

We define annotation as any process of adding judgment to all, a component of, or a subset of a corpus. The input to any annotation effort is itself a corpus of either written or spoken linguistic performance with or without prior annotation. In the case of spoken data, a transcript is already a kind of annotation, encoding subtle human judgments about what was uttered. Under this

broad definition of annotation it is not only possible but indeed typical that the input to an annotation process be an already annotated (transcribed or otherwise) corpus. It is also typical that annotations layer one upon the other to build an increasingly rich resource. In the case of a named entity annotation based upon a transcript of the audio of a news broadcast, the audio is the raw data; the transcript is the first level of annotation and the named entity tagging, based upon the transcript, is the second layer.

4.1. Transcription & Text Normalization

LDC collects the source material as newswire text and broadcast audio. Research sites are permitted to work directly from the broadcast signal; however most work from text intermediaries. The newswires arrive at LDC as electronic text with some form of markup either ANPA (American Newswire Publishers Association) standard markup or some proprietary scheme. LDC normalizes the newswires converting their various types of markup into a standard SGML markup. For the broadcast sources, LDC acquires some form of text intermediary, either:

· closed captioning of a television program
· commercially available transcripts
· transcripts produced specifically for the project

Ultimately these texts will have story boundaries inserted and corrected. We call the result **reference text**. To ensure that the research systems focus on linguistic content, LDC removes any formatting information and metadata from the reference text and tokenizes it one word per line. We call this **tokenized text**. In 1998, sites were permitted to work with the tokenized text or with **ASR text**, the output of automatic speech engines. In 1999, sites were constrained to use only the ASR output. NIST and Dragon Systems produced the ASR text for the project. To accommodate future use, LDC publishes the audio and all forms of the text intermediaries. This has allowed, for example, the use of the TDT audio data in subsequent TREC Spoken Document Retrieval evaluations.

Because TDT focuses on real world data, no attempt was made to produce high quality transcripts of the kind used for speech recognition projects (Hub4 and Hub5 for example). However, the raw audio data remains available and organizations may transcribe or annotate a portion to suit their needs. In fact, LDC re-transcribed a few hours of broadcast from the unused August 1998 data so that NIST could use it in speech recognition technologies. LDC will release those careful transcripts and the corresponding audio files in 2000.

4.2. Segmentation

Once the text intermediaries are ready, LDC annotators provide ground-truth for the segmentation task by listening to the audio to while reading the transcripts to determine story boundaries; the newswire comes with story boundaries although there are occasional anomalies. Each story boundary is marked in the reference text and, in the case of audio transcripts, includes a time stamp to provide the time offset of that boundary from the beginning of the audio file. Because story boundaries are removed from the tokenized text, a separate set of boundary tables defines the reference segmentation in terms of word offsets from the beginning of the text file.

4.3. Topic-Story and First Story Annotation

In Topic-Story and First Story annotation, the notion of topic is explicit. In Topic-Story annotation, staff read a list of typically 20 topics that have been defined in advanced and refined after topic research. Working with a day's newswire or broadcast sampling, annotators then read each source story it discusses any of the topics in the list. LDC has designed a custom interface that presents the stories and topics to the annotator and collects their judgments in a relevance table. The relevance tables are delivered to research participants or not depending upon the phase of research. The interface also assigns work to annotators so that double-blind experiments to determine annotation consistency are possible. Indeed since mid 1998, these experiments have been a regular part of topic-story annotation. This type of annotation is done exhaustively so that for a corpus of 54,000 stories and 100 topics, the number of decisions encoded is well over 5 million and represents more than 6,000 person-hours of effort. Topic-Story annotations support Topic Detection and Topic Tracking.

First-Story Annotation supports research in Topic Detection and First Story Detection. Here, the annotators' task is to select a story at random from the corpus, determine its seminal event, define a topic and locate the first story in the corpus to discuss that topic. Customized search engines and substantial knowledge of the corpus (the same annotation team did first-story annotation after topic-story annotation) helped annotators in this task. Typically, annotators use a combination of relevance ranked searches and increasingly narrow date restrictions to locate the first on-topic story.

4.4. Story-Story Annotation

Concerns over the time and difficulty involved in explicitly defining an event and topic, lead the TDT participants to add Story-Story Linking in 1999. Here the annotator's task is to read a seed story selected at random from the corpus, compare it to another story and judge whether the two stories discuss the same topic. Although the concept of event and topic and the rules of interpretation are the same as in the other types of annotation, no specific topic is predefined. This type of annotation is probably closest to the real world use of search engines where users have a rough idea of what they want but have not yet defined the bounds of their search.

5. Corpora

LDC distributes the speech and text portions of the two TDT corpora separately. The speech corpora contain the complete 16 bit, 16KHz digital audio files for each broadcast unit in NIST SPHERE format. Researchers can acquire all of the audio or just the broadcasts from a specific source. The complete TDT-2 audio in compressed format fills more than 70 CDs. The TDT-3 audio to be released in 2000 will be of similar size. The TDT text corpora are divided into sets of related files. Each set is stored in a directory whose name matches the filename extension of each file in the set.

The file names are composed of five parts separate by underscores. The first part is the date of transmission in YYYYMMDD format. That is, the year is written first in four digits followed by the month and day written in two digits each. This format preserves date ordering when the

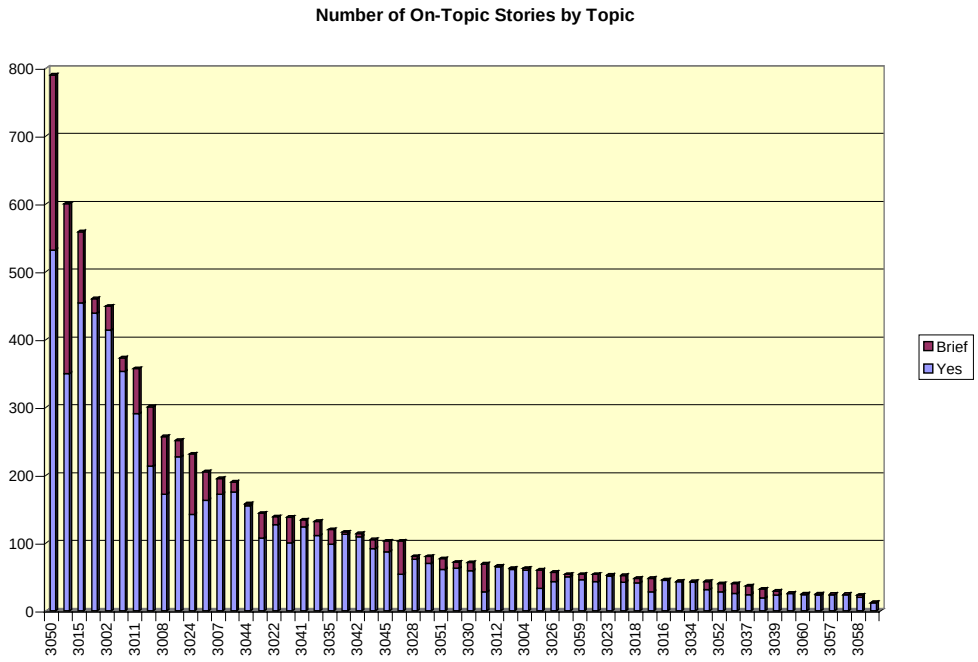

Figure 3: The number of on-topic stories for each topic defined in TDT-3. Note that the percentage of brief mentions varies by topic and that the topics are sorted by total number of on-topic stories (yes+brief).

file names are ASCII sorted and avoids confusion with dates before and after the millenium date change (19991231-20000101). The second and third parts of the file name are the start and end times of the broadcast in 24-hour format (ie. 1:30PM is written 1330). The last two components are abbreviations for the broadcast source and program name.

TDT data files are divided into sets as follows.

The SGM set contains SGML encoded reference text that retains the metadata and preserves the formatting of the original source.

The TKN set contains text from which all metadata and formatting has been removed and in which all words have been placed on their own line. TKN files begin with a <DOCSET> tag that also indicates:

- type of text
- source file ID
- date of collection
- source name
- source language (ie. English or Mandarin)
- content language (ie. Native or English)

The value of the content language field is either "Native" indicating no translation of the source or the language into which the content has been translated. IN TDT-2 and TDT-3 documents collected in Chinese are also available as their English translations. After the <DOCSET> tag, each record in the file begins with a <W> tag providing a record ID and a single word.

<DOCSET type = CAPTION
fileid = 19981001_0130_0200_CNN_HDL
collect_date = 19981001_0130 collect_src = CNN
src_lang = ENGLISH content_lang = NATIVE>
<W recid=1> Congress
<W recid=2> and
<W recid=3> the
<W recid=4> president
<W recid=5> are
<W recid=6> celebrating
<W recid=7> the
<W recid=8> budget
<W recid=9> surplus
<W recid=10> and
<W recid=11> fighting
<W recid=12> over
<W recid=13> whether
<W recid=14> it
<W recid=15> should
<W recid=16> go
<W recid=17> in
<W recid=18> your
<W recid=19> pocket.

The TKN file contains the most accurate version of the pure text content; there are no story boundaries or metadata in the TKN files.

The TKB_BND files contain story boundaries. There is one similarly named TKN_BND file for each TKN file. The TKN_BND file begins with a <BOUNDSET> tag that is, other than its name, homomorphic to the <DOCSET> tag in the TKN file. After the <BOUNDSET> tag, each line begins with a <BOUNDARY> tag that provides the:

- document number
- document type (NEWS, MISCELLANEOUS or UNTRANSCRIBED)
- beginning and ending times of the story expressed in seconds with two decimals of precision
- the beginning and ending words of the story expressed as record in the TKN file

In TDT-3, there are 1957 SGM files and the same number of TKN and TKN_BND files. Examples follow.

<BOUNDSET type=CAPTION
fileid=19981001_0130_0200_CNN_HDL
collect_date=19981001_0130 collect_src=CNN
src_lang=ENGLISH content_lang=NATIVE>

<BOUNDARY docno=CNN19981001.0130.0000
doctype=MISCELLANEOUS Bsec=0.00 Esec=18.93
Brecid=1 Erecid=62>

<BOUNDARY docno=CNN19981001.0130.0018
doctype=NEWS Bsec=18.93 Esec=77.15 Brecid=63
Erecid=216>

For each broadcast file, there is, in theory, an AS file containing the output of an automatic speech recognition engine having run over the source audio. AS0 files contain the output of the first ASR engine used in the TDT program, Dragon System's, with Mandarin audio as input. AS1 files contain ASR text produced by NIST using a recognizer developed by BBN and taking English audio as input.

The 122 AS0 and 731 AS1 files begin with a <DOCSET> tag and then continue with a <W> tag in each record. The <W> tags in the AS0 and AS1 files, indicate the beginning time of the hypothesized word and its duration. The *Clust* field provides an emic clustering of the words. In some cases the clustering will correspond to different speakers but this is not guaranteed. Music overlays or changes in channel characteristics, for example, may also cause a chance in the cluster assigned to a word. The last field, *Conf*, provides the system's confidence in its performance. The recognizer NIST used to create the AS1 files does not output clusters or confidence ratings so the appropriate fields have been filled with "NA" values in those files. The AS files also contain <X> tags which mark sounds the recognizers could not interpret. These could be periods of silence, musical interludes, background noises or simply unrecognized words. An example follows.

<DOCSET type=ASRTEXT
fileid=19981001_0800_0900_VOA_MAN
collect_date=19981001_0800
collect_src=VOA
src_lang=MANDARIN content_lang=NATIVE
proc_remarks="Dragon Mandarin ASR">
<W recid=1 Bsec=0.01 Dur=0.36 Clust=0 Conf=0.75> xx
<W recid=2 Bsec=0.38 Dur=0.59 Clust=0 Conf=0.81> yy
<X Bsec=0.97 Dur=1.40 Conf=NA>

There is a boundary file corresponding to each AS file. These are called AS0_BND and AS1_BND and their form is identical to that of the TKN_BND file.

To support translingual research, LDC has provided a best-of-breed translation of each Chinese text. Using Systran's Chinese-English software, LDC has created MTAS0, MTAS0_BND, MTTKN and MTTKN_BND files that respectively contain the translations of the content in the AS0, AS0_BND, TKN and TKN_BND. Records in the MTAS0 and MTTKN files contain an additional "tr" field that indicates whether the Systran software translated the word or not. Note that this is not an evaluation of translation quality but a simple indication of whether the word is Chinese or English (ie. whether character set used to encode the word is ASCII or one of the Chinese character encodings). There are 122 MTAS0 (and MTAS0_BND) files and 519 MTTKN (and MTTKN_BND) files.

The corpus-info directory contains file with basic statistics on the corpora including how many of each file type exist, how many on-topic stories there are per topic, and how many stories were judged to be on-topic for more than one topic.

The TOPICS file set contains the topic table for each type of annotation. The topic relevance table contains a record for each instance in which a story was judged to discuss a topic. If a story discusses more than one topic, there are multiple records for that story.

Each record in the topic relevance table lists the topic ID, the story id and id of the file containing it, and the level of relevance. “Yes” means the story discusses the topic; “Brief” means that the story is primarily about something else but mentions the topic briefly.

<ONTOPIC topicid=3001 level=YES
docno=APW19981031.0167
fileid=19981031_0409_0607_APW_ENG
comments=NO>
<ONTOPIC topicid=3001 level=BRIEF
docno=APW19981101.0193
fileid=19981101_0506_0631_APW_ENG
comments=NO>

The story link table has a record for each story pair compared. The fields in individual records give the document id of the seed story and the story to which it was compared and a label of "Yes", "No" or "Brief"

<LINK seed_docno=ABC19981027.1830.1520
comp_docno=ABC19981030.1830.0047 label=N>
<LINK seed_docno=ABC19981027.1830.1520
comp_docno=CNN19981024.1130.0562 label=Y>

6. References

ACE, 2000, Automatic Content Extraction [www.nist.gov/speech/tests/ace].

Bird, Steven, David Day, John Garofalo, John Henderson, Christophe Laprun and Mark Liberman, 2000, ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

Bird, Steven and Mark Liberman, 1999, Linguistic Annotation Page, [www.ldc.upenn.edu/annotation]

Doddington, G. (1999). The 1999 Topic Detection and Tracking (TDT) Task Definition and Evaluation Plan. Available at http://www.nist.gov/TDT.

Graff, David and Steven Bird, 2000, Many Uses, Many Annotations for Large Speech Corpora: Switchboard and TDT as Case Studies, In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

LDC, 2000, Linguistic Data Consortium Homepage [http://www.ldc.upenn.edu]

Strassel, Stephanie, Dave Graff, Nii Martey and Christopher Cieri, 2000, Quality Control in Large

Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora. In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.

TalkBank, 2000, NSF TalkBank Program [www.talkbank.org]

TIDES, 2000, DARPA Program in Translingual Information Detection Extraction and Summarization [www.arpa.mil/ito/research/tides]

VOA, 2000, Voice of America page, [www.voa.gov]

Wayne, Charles, 1998, Topic Detection & Tracking: A Case Study in Corpus Creation & Evaluation Methodologies. in Proceedings of Language Resources and Evaluation Conference, Granada, Spain, May 1998.

Wayne, Charles, 2000, Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation In Proceedings of the Second International Language Resources and Evaluation Conference, Athens, Greece, May 2000.