Multiple Annotations of Reusable Data Resources: Corpora for Topic Detection and Tracking

Christopher Cieri

Linguistic Data Consortium, 3615 Market Street, Philadelphia. PA 19104, USA

Abstract

Responding to demands for very large, easily accessible, reusable news corpora to support research in the topic detection and tracking paradigm, the Linguistic Data Consortium created the TDT corpora. In addition to supporting research in the Topic Detection and Tracking program, the TDT corpora were collected and annotated with an eye toward reuse and re-annotation. Their value is confirmed in the number of projects that have benefited from part of all of the TDT corpora for new uses. The paragraphs that follow will describe the raw data and annotations in the TDT corpora and summarize their use in multiple common-task research programs.

Keywords: text corpora, information retrieval, text classification, topic detection and tracking, corpus annotation

1. Introduction

Any statistical analysis of text data requires appropriate infrastructure including annotated text corpora. Given the cost of corpus creation and annotation, demand for large, reusable corpora is on the rise. The Linguistic Data Consortium has created two such corpora over the past two years to support research and technology development in the new research paradigm of Topic Detection and Tracking. Because these corpora were developed to encourage reuse, other research programs have benefited from the availability of this data. This paper will describe the corpora created for TDT in terms of both raw materials and annotations and how they further the goals of the current project but also the larger goals of open access, standardization of formats, reusability and multiple annotation.

The DARPA-sponsored research program in Topic Detection and Tracking (TDT) began in 1997 with a small pilot study. In 1998, the program expanded to include new research sites and to exploit the expertise of the U.S. National Institute of Standards and Technology (NIST) in the evaluation of results and the experience of the Linguistic Data Consortium (LDC) in corpus creation. As of the time of writing, TDT is finishing its third phase with evaluation results due in early 2000 (see: http://www.nist.gov/speech/tdt3/tdt3.htm). The goal of research in the Topic Detection and Tracking program is to create core technology as part of a news understanding system. This system should ultimately be capable of processing streams of news content across multiple languages and media. Processing includes dividing the news streams into individual stories and categorizing them according to the topic(s) they describe. The languages can be as diverse as English and Chinese. The media include broadcast television and radio, newswires, WWW sites, newsgroups, e-mail lists or some future innovation or combination. The research tasks defined under the TDT paradigm are:

- 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

Research sites are constrained to use only the linguistic content of the raw material to accomplish these tasks. Formatting information such as the paragraph breaks and headers that may appear in newswire are not available to the research sites. Although the TDT Corpora have been created to accommodate these tasks, they have also been designed with future projects in mind. The delivered corpora contain both the reference version of each story as well as a tokenized version with all formatting and extra-linguistic material removed.

2. Collection

The TDT corpora are multiply annotated collections of broadcast news. The TDT Pilot Study Corpus was based upon a small number of stories and sources. The TDT-2 English corpus contains daily samplings from two television, two radio and two newswire sources, specifically: ABC *World News Tonight*, CNN *Headline News*, Public Radio International *The World*, Voice of America English news radio and newswires from the Associated Press and New York Times services. The broadcast sources were sampled on a daily basis over a six-month period. For the newswire sources, LDC sampled approximately 80 stories per day over the same period, January through June of 1998.

			1998											1999										
Lg.	Type	Source	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
E	N	AP	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
E	N	NYT	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
E	R	PRI	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
E	R	VOA	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
E	T	CNN	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
E	T	ABC	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
E	T	NBC							■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
E	T	MSNBC										■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
M	N	Xinhua	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
M	W	Zaobao		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
M	R	VOA			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
M	T	CCTV																				■	■	■	■	■
S	N	El Norte	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
S	R	VOA			■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
S	T	Eco																				■	■	■	■	■
S	T	Univision																				■	■	■	■	■

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

The TDT-2 Mandarin collection includes daily samples of Voice of America Mandarin radio broadcasts and Xinhua News Service's newswire plus news stories downloaded from the WWW pages of a Singapore based news agency Zaobao. Over the 180 days of collection, LDC collected over 54,000 stories and 634 hours of recorded audio in English. The 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. TDT-3 English includes more than 35,000 stories. The Mandarin data from TDT2 and TDT3 together totals 30,000 stories. In anticipation of future projects, the LDC collected Voice of America Spanish radio broadcasts and newswire from El Norte's news service during the same period. In preparation for the next generation of TDT research, LDC 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. Figure 1 summarizes the collection schedule.

3. Annotation

Annotation, is herein defined as any process of adding interpretation to all, a component of or a subset of a corpus. This is a broader definition of annotation than, for example, the one employed by the Corpus Encoding Standard. The input to any annotation effort is itself a corpus. Initially, this input corpus may be simply a collection of raw data. Where the object of study is written linguistic performance, the written text constitutes raw data. However, where the object of study is spoken linguistic performance, the raw data is the audio data. 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 (eg. transcribed) corpus. Thus is it 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.

3.1. Transcription & Text Normalization

While TDT newswire is collected as encoded text, many TDT sources are television or radio broadcasts. TDT research sites are permitted to work directly from the broadcast signal; however, sites typically work from text intermediaries. To accommodate future use, both the audio and text intermediaries are available. This allowed TREC participants in the Spoken Document Retrieval track to select TDT-2 audio for the 1999 evaluation. Because TDT project managers have a strong interest in systems that can handle the errors present in real world, no attempt is made to produce high quality transcripts of the kind used for speech recognition projects (Hub4 and Hub5 for example). Text intermediaries come from a number of sources. Where closed-captioning or commercial transcripts are available, LDC acquires those to serve as text intermediaries. Otherwise, LDC transcribes or contracts the transcription of the broadcast television and radio sources. In TDT-2, sites were given the option to use these manually transcribed stories; the TDT-2 corpus contains both the manually and automatically transcribed texts of the ABC broadcasts. In TDT-3, researchers are permitted to work only with the errorful output of automatic speech recognition system running over the broadcast audio. Dragon Systems and NIST provide the reference ASR text for the project that LDC tokenizes (one word per line) before delivering to the sites. In the case of newswire text LDC first removes any headers, time stamps or other extra-linguistic data. Figure 2 compares reference and tokenized data in the case of a newswire story. The TDT corpora contain both the reference texts and the tokenized versions with all formatting removed.

Although the project's bias toward real world data prevents LDC from transcribing the broadcast audio to the same specifications used in speech recognition research, the raw audio data remains available and organizations may transcribe or annotate a portion to suit their needs. In fact, NIST selected stories from the August 1998 data that LDC re-transcribed to Hub-4 specifications for use in the 1999 evaluation.

<DOC> <DOCNO>APW19980104.0002</DOCNO> <DOCTYPE>NEWS STORY</DOCTYPE> <DATE_TIME>01/04/1998 00:02:00</DATE_TIME> <HEADER>w2488 &Cx1f; wstm-u i &Cx13; &Cx11; BC-Cambodia-PolPot 01-04 0570</HEADER> <BODY> <SLUG>BC-Cambodia-Pol Pot</SLUG> <HEADLINE>Pol Pot has fled Cambodia, Thai minister claims</HEADLINE> &UR; By ROBIN McDOWELL &QC; &UR; Associated Press Writer &QC; <TEXT> PHNOM PENH, Cambodia (AP) _ The mystery surrounding Pol Pot deepened Sunday after Thailand's foreign minister claimed that the Khmer Rouge leader had fled Cambodia. Earlier, Chinese diplomats here denied allegations he had been granted asylum in China. They could not be reached for comment Sunday. One of modern history's most secretive figures, Pol Pot was last seen by independent observers last October at the Khmer Rouge base of Anlong Veng in northern Cambodia. Breaking an 18-year silence he denied to a Western reporter that he had orchestrated the killings of as many as 2 million of his countrymen in the mid-1970s. (material deleted) </TEXT> (PROFILE (WS SL:BC-Cambodia-Pol Pot; CT:i; (material deleted) (LANG:ENGLISH;))) </BODY> <TRAILER>AP-NY-01-04-98 0002EST</TRAILER> </DOC>	<DOCSET type=NEWSWIRE fileid=19980104_0002_0418_APW_E NG collect_date=19980104_0002 collect_src=APW src_lang=ENGLISH content_lang=NATIVE> <W recid=1> The <W recid=2> mystery <W recid=3> surrounding <W recid=4> Pol <W recid=5> Pot <W recid=6> deepened <W recid=7> Sunday <W recid=8> after <W recid=9> Thailand's <W recid=10> foreign <W recid=11> minister <W recid=12> claimed <W recid=13> that <W recid=14> the <W recid=15> Khmer <W recid=16> Rouge <W recid=17> leader <W recid=18> had <W recid=19> fled <W recid=20> Cambodia. <W recid=21> Earlier, <W recid=22> Chinese <W recid=23> diplomats <W recid=24> here <W recid=25> denied <W recid=26> allegations <W recid=27> he <W recid=28> had <W recid=29> been <W recid=30> granted <W recid=31> asylum <W recid=32> in <W recid=33> China.

Figure 2: A TDT2 story in both SGML-encoded reference form and in tokenized format. Note that paragraph information, headlines, sluglines, datelines and the like have all been removed.

3.2. Segmentation

To support the segmentation task, LDC annotators provide reference segmentation for each broadcast story in the corpus by reading over the transcripts while listening to the audio to determine the story boundaries. Story boundaries are already marked in newswires. 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. Figure 3 shows the boundary table records for two TDT-2 stories a portion of the text that the records segment.

Boundary Table	<BOUNDARY docno=APW19980104.0002 doctype=NEWS Brecid=1 Erecid=533> <BOUNDARY docno=APW19980104.0012 doctype=NEWS Brecid=534 Erecid=724>		
Tokenized Text	<W recid=515> Pressed <W recid=516> against <W recid=517> the <W recid=518> Thai <W recid=519> frontier, <W recid=520> the <W recid=521> royalists <W recid=522> are <W recid=523> continuing <W recid=524> to <W recid=525> hold <W recid=526> out <W recid=527> against <W recid=528> a <W recid=529> far <W recid=530> superior.	<W recid=531> Hun <W recid=532> Sen <W recid=533> army. <W recid=534> Seven <W recid=535> skiers <W recid=536> were <W recid=537> killed <W recid=538> and <W recid=539> at <W recid=540> least <W recid=541> one <W recid=542> person <W recid=543> was <W recid=544> missing <W recid=545> after <W recid=546> avalanches	<W recid=547> hit <W recid=548> two <W recid=549> separate <W recid=550> ski <W recid=551> parties <W recid=552> in <W recid=553> the <W recid=554> Selkirk <W recid=555> Mountains <W recid=556> in <W recid=557> southeast <W recid=558> British <W recid=559> Columbia, <W recid=560> police <W recid=561> said <W recid=562> Saturday.

Figure 3: The two boundary table records above impose segmentation on tokenized text shown below them. Story APW19980104.0002 ends at word 533. Story APW19980104.0012 begins at word 534.

3.3. Topic Explicit Annotation

The notions of event and topic are crucial to TDT annotation. A TDT *event* is defined as a specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences. 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 defined as collection of related events and activities. To increase the consistency of judgements about what constitutes "related", annotators refer to a set of rules of interpretation. These rules state, for each type of event: crimes, natural disasters, scientific discoveries and scandals, etc, what other events should 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, the China Airlines Crash and the Pope's visit to Cuba (for the complete list see: http://www.ldc.upenn.edu/Projects/TDT2). During the technology evaluation, topics are "defined" by presenting sites with examples of (typically four) on-topic stories. Within TDT, there are two types of annotation that explicitly define topics. They are Topic-Story annotation and First Story annotation.

3.3.1. Topic-Story Annotation

In Topic-Story annotation, annotators read stories and decide whether they discuss any of the topics defined. During any session, an annotator will typically work with 20 predefined topics and all of the stories collected from a single source for a single month. LDC has designed a custom interface that presents the stories and topics to an annotator who reads the story and indicates whether it discusses any of the topics. All of the judgements are stored in a relevance table that is ultimately delivered to research participants. 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 5,400,000 and represents more than 6,000 person-hours of effort. The

result of Topic-Story annotation support the research goals of Topic Detection (find topics in the news) and Topic Tracking (find all stories related to a given topic). Topic Story annotations are stored in a separate topic relevance table so they can be incorporated or not as appropriate. Figure 4 provides an example of a record from the topic relevance table and the story to which it refers.

Relevance Table	<ONTOPIC topicid=20001 level=YES docno=ABC19980110.1830.1008 fileid=19980110_1830_1900_ABC_WNT comments=NO>
SGML Text of Story	<DOC> <DOCNO> ABC19980110.1830.1008 </DOCNO> <DOCTYPE> NEWS STORY </DOCTYPE> <DATE_TIME> 01/10/1998 18:46:48.41 </DATE_TIME> <BODY> <HEADLINE> CONSUMER ELECTRONICS SHOW </HEADLINE> Byline:JACK SMITH, AARON BROWN High:MAKING TECHNOLOGY WORK FOR YOU Spec:COMPUTERS / TECHNOLOGY / ELECTRONICS / CONSUMERS <TEXT> <TURN> <ANNOTATION> spkr:AARON_BROWN </ANNOTATION> President Clinton's point man on the financial crisis in Asia is heading towards Indonesia tonight. <ANNOTATION> (voice-over) </ANNOTATION> Deputy Treasury Secretary Lawrence Summers will pressure government and business leaders there to put in place the belt-tightening measures required by the I.M.F. in... <ANNOTATION> (on camera) </ANNOTATION> ...exchange for billions to bail out failing Indonesian banks and businesses. </TEXT> </BODY> <END_TIME> 01/10/1998 18:47:09.84 </END_TIME> </DOC>

Figure 4: The topic relevance table indicates which stories (as defined by the boundary table) discuss TDT topics. In this case, story number ABC19980110.1830.1008 discusses topic 20001, the Asian Economic Crisis.

3.3.2. First Story Annotation

A close cousin to Topic-Story Annotation, First-Story Annotation supports research in Topic Detection and First Story Detection. In this case, annotators read a story selected at random from the corpus and locate the first story in the corpus to discuss this topic. Annotators may use their substantial knowledge of world events to narrow the search and are aided by the use of customized search engines. Typically, annotators use a combination of relevance ranked searches and increasingly narrow date restrictions to locate the first on-topic story. As proof of concept, LDC annotators used search engines to locate the first story for each of the topics that were exhaustively annotated under Topic-Story Annotation. Figure 5 shows the First Story entries for topic 20001, the Asian Economic Crisis, as well as a sample of the first English story in the corpus to discuss this topic.

3.4. Story-Story Linking

A concern over the time and difficulty involved in explicitly defining an event and topic, lead the TDT participants to add Story-Story Linking in phase three. In this case, annotators read a seed story selected at random from the corpus and compare it to another of other stories also selected at random. The goal is to determine whether the stories under comparison discuss the same topic. The concept of event and topic and the rules of interpretation are the same as in the other types of annotation. However, 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. The results of the Story-Story Linking annotation will be available when the evaluation results are released sometime after this paper is printed.

First Mandarin Story: <ONTOPIC topicid=20001 level=YES docno=XIN19980101.0018 fileid=19980101_0016_1116_XIN_MAN comments=NO>
First English Story: <ONTOPIC topicid=20001 level=YES docno=APW19980104.0286 fileid=19980104_0720_0851_APW_ENG comments=NO>
... SEOUL, South Korea (AP) _ International financier George Soros met with President-elect Kim Dae-jung Sunday and expressed interest in increasing investment in South Korea, Kim's aides said. Soros' move was welcome news for South Korea as it tries to woo foreign investment and regain foreigners' confidence in its economy after needing to arrange rescue loans from the International Monetary Fund in early December. ...

Figure 5: The first story table indicates the first story to discuss a selected topic. Note above the records for the first English and Mandarin stories in the corpus to discuss topic 2001, the Asian Economic Crisis. The third row gives a sample of the first English story.

4. Evaluation and Quality Control

Within the DARPA sponsored TDT project evaluation is a crucial component. Research is evaluated formally at least once each year with dry-run evaluations in between. The TDT corpora are subject to equivalent scrutiny. Annotators' judgements are checked for quality in four different processes. During *Precision QC*, all of the stories judged to discuss a specific topic are viewed as a group by a senior annotator who looks for false alarms. During *Recall QC*, external processes such as search engines are used to locate stories that may perhaps discuss a specific topic. Senior annotators review those candidates to locate instances where LDC annotators may have missed an on-topic story. In addition to these ex post facto processes, between 5% and 10% of all stories are independently annotated by multiple annotators in a double-blind process. Discrepancies are counted to yield measures of inter-annotator consistency and then resolved by a senior annotator. Finally, the system output from the research sites is compared with the human judgements and adjudicated by LDC senior annotators. Where the systems catch mistakes made by the human annotators, the corrections are fed back into the corpus before the final scoring is concluded.

5. Other Uses

The TDT corpora have a number of features that render them susceptible to re-annotation and re-use. The raw material, broadcast news and newswire from well-known providers, is intrinsically interesting. The combination of both broad (2 years) and deep (hundreds of stories per day) coverage makes the corpus useful for projects that hope to observe linguistic correlates of trends in news reporting over time. The languages represented are English, Mandarin and Spanish. The media include television, radio, newswire and WWW news sites.

The corpus contains both original audio and text intermediaries in both reference form and in tokenized form with all formatting and extra-linguistic material removed. The text intermediaries include manually and automatically created transcripts. The text files and tables are parseable SGML with DTDs provided. The audio files are broadcast-length with story boundaries encoded in tables. All annotation is encoded in separate tables so as not to clutter the reference text. Finally, all material is available through LDC.

Because the TDT corpora were designed for multiple use, they have served a number of functions. To date, the entire TDT-2 corpus has been used in the TREC Spoken Document Retrieval evaluation. TDT-2 will also serve as the base for a new DARPA-sponsored program in Automatic Content Extraction whose tasks are being defined at the time of writing but will likely include the extraction of named entities from errorful data. Smaller pieces of the TDT data have been carefully re-transcribed to support evaluation in automatic speech recognition as part of the DARPA Hub-4 program. Finally, the on-topic stories from TDT-2 were re-annotated and used in the NSF-sponsored summer workshop in speech and language processing which is held each summer at John's Hopkins University. For this work, the on-topic stories were collected in groups by topic, chronologically ordered and tokenized into individual sentences. LDC staff then annotated each sentence to indicate 1) whether that specific sentence discussed the selected topic and 2) whether that sentence presented any new information.

The TDT annotation tools reference a common set of data tables in a standard format. After three years of evolution, annotation practice has stabilized opening the possibility of a unified system to handle the various kinds of annotation desired. A number of projects have grappled with the problem of systems for multiple annotation. One such system is the Gate system developed at University of Sheffield. Another is the Talkbank project underway at CMU and the LDC. It is our hope that future annotations to corpora like TDT will benefit from standard practices and common tools.

6. Conclusions

The TDT corpora were design to support multiple annotation and reuse. The size and variety of the corpora make them interesting to numerous research projects in speech and language processing. The use of standard file formats and stand-off markup encourage use on multiple platforms and allow re-annotation of either the raw data or the annotations themselves. The intensive reuse of the TDT corpora for other projects including Hub-4, ACE and TREC SDR and its re-annotation for the John's Hopkins summer workshop in speech and language processing confirm that such corpora provide real benefit to multiple research communities.

References

Bird, Steven and Mark Liberman (1999) A formal framework for linguistic annotation, Technical Report MS-CIS-99-01, Computer and Information Science, University of Pennsylvania. (ftp://ftp.cis.upenn.edu/pub/sb/papers/cis-9901/cis-9901.pdf)

Doddington, George (1999) The 1999 Topic Detection and Tracking (TDT3) Task Definition and Evaluation Plan Version 2.7 (http://www.nist.gov/speech/tdt3/doc/tdt3.eval.plan.99.v2.7.ps)

Gaizauskas, Rob, et al., GATE User Guide, ILASH, Univ. of Sheffield, UK (http://www.dcs.shef.ac.uk/research/groups/nlp/gate/system_docs/user_guide/main/main.html)

Ide, Nancy and Greg Priest-Dorman (1999) Corpus Encoding Standard - Document CES 1. Version 1.5. (http://www.cs.vassar.edu/CES/CES1-1.html#ToCDef)