The TDT Pilot Study Corpus Documentation

Corpus Overview

Topic Detection and Tracking (TDT) refers to automatic techniques for finding topically related material in streams of data such as newswire and broadcast news. During 1997 a pilot study was conducted to establish baselines for the state of the art. Research objectives included finding topically homogeneous sections (segmentation), detecting the occurrence of new events (detection), and tracking the reoccurrence of old or new events (tracking). This document describes the TDT Pilot Study Corpus (hereafter "corpus") which was created to support this research.

The corpus comprises a set of stories that includes both newswire (text) and broadcast news (speech). Each story is represented as a stream of text, in which the text is either taken directly from the newswire (Reuters) or is a manual transcription of the broadcast news speech (CNN). The corpus spans the period from July 1, 1994 to June 30, 1995. It contains nearly 16,000 stories, with about half taken from Reuters newswire and half from CNN broadcast news transcripts.

An integral and key part of the corpus is the annotation of the corpus in terms of the events discussed in the stories. A set of 25 target events were defined which span a variety of event types and which cover a subset of the events discussed in the corpus stories. Annotation data for these events are included in the corpus and provide a basis for training TDT systems and evaluating performance on a variety of TDT tasks.

Events might be unexpected, such as the eruption of a volcano, or expected, such as a political election. Stories that discuss unexpected events will of course follow the event, whereas stories on expected events can both precede and follow the event.

Details of the construction and format of the TDT corpus are explained in the remainder of this document. The TDT tasks and evaluation plan are detailed in the document called "The 1997 Topic Detection and Tracking (TDT) Evaluation Plan".

Source Data Selection

The data were selected from the following sources:

- Broadcast news (CNN)
- Newswire (Reuters)

A breakdown of the number of stories from each half-year is included:

Program Source	1994	1995	time-of-day
CNN Sources			
Newsnight	194	49	12:00 AM
CNN Overnight	160	44	3:00 AM
Daybreak	1056	470	5:00 AM
Newsday	567	514	12:00 PM
Newshour	453	353	3:00 PM
Early Prime	665	290	4:30 PM
The International Hour	569	176	5:00 PM
World News	641	225	6:00 PM
The World Today	899	573	10:00 PM
Total # of Stories	**1994**	**1995**	**1994+1995**
CNN	5204	2694	7898
Reuters	3520	4445	7965
CNN+Reuters	8724	7139	15863

In order to balance the corpus between speech (CNN) and text (Reuters), the Reuters collections were scanned sequentially in the time period and each story was selected with a 13.84% probability using a pseudo-random number generator. The smaller number of CNN stories for 1995 is unexplained, but is true in the source CDs and is not an artifact of the extraction programs.

Chronological Ordering

The various sources were merged so that they appear in chronological order in the corpus. The date/time information that comes with Reuters stories was used for this purpose. Unfortunately, the CNN transcriptions do not provide broadcast time, so an approximate time was assigned to the CNN stories according to the time-of-day listed in the CNN program source table above.

A sequential number was assigned to each story after the merging. That number has the form TDTnnnnnn and is placed in the TDTID tag of the resulting data.

Event Definitions

Twenty-five events were defined. Each story in the corpus was labeled according to whether the story discussed the event, or not. There were three labels used, namely "YES", "NO", and "BRIEF" (if no more than a tenth of the story discussed the event.) The event definitions comprise a "short" definition, a "long" definition, and optionally some additional discussion to clarify the event definition.

Great care was taken in labeling the stories in the corpus with event labels. The corpus was labeled twice, once by each of two independent sites. Differences were then separately reconciled, and the resulting labels are therefore regarded as "reliable".

The event definitions and labeling instructions are listed at the end of this overview.

Data Format

The final data is distributed as an SGML-tagged file of stories. The following SGML tags are used. Each tag <XXX> is paired with an ending tag labeled </XXX>.

DOC

> Surrounds each story.

DOCID

> A unique identifier for this story constructed from data included on the disks using the story. The general format is one of CNNnnn-nn.YYMMDD or REUnnn-nnnn.YYMMDD for broadcast news or newswire data respectively. YYMMDD is the date of the story, and nnn-nn is a number extracted from the meta-data for the story (the extracted number is not always exactly the indicated number of digits).

TDTID

> A second unique identifier of the stories in this collection, assigned during the building of the corpus. These IDs are of the format TDTnnnnnn and are assigned sequentially. The first story in the corpus was assigned TDT000001, the second TDT000002, and the last TDT015863.

SOURCE

> Is a textual description of the source of this story. For example "CNN Daybreak".

DATE

> Is the date in the format MM/DD/YY HH:MM. The time is omitted for broadcast news sources since it is not known.

TITLE

Is a title. For broadcast news sources, this was provided by the transcribers. For newswire, this is the headline.

HEADLINE

Not used for broadcast sources; the headline for newswire. For Reuters, this field will have the same value as the TITLE tag.

SUBJECT

For newswire, this is the slug line. For broadcast news, it is the transcriber-generated description of the type of show--e.g., "News" or "Domestic". These are not especially useful tags, but are provided in case they are helpful. This tag may repeat, with one subject per occurrence.

AUTHOR

For broadcast, not used. For newswire, the byline if present.

TOPIC

For newswire, not used. For broadcast, some JGI-provided topic descriptors, something like Library of Congress subject headings--e.g., "Actors and Actresses". If a story has multiple topics, this tag repeats. WARNING: These are ***not*** the TDT-defined event labels.

SUMMARY

For newswire, not used. For broadcast, a JGI-provided summary of the major content of the story.

TEXT

Enclosed the actual text. None of the fields above is in the text.

NOTES

For broadcast, not used. For newswire, it is a note that is often attached to the start of a story describing how it is different from an earlier version.

DATELINE

For broadcast, not used. For newswire, the dateline of the story. For example, "WASHINGTON(Reuter)".

SP

For newswire, not used. For broadcast, this is the JGI-provided name of the speaker. Occurs repeatedly.

P

Surrounds paragraphs.

B

Surrounds text that should be emboldened. Probably appears only in broadcast and is JGI-provided.

I

Surrounds text that should be italicized. Probably appears only in broadcast and is JGI-provided.

COPYRIGHT

Surrounds a trailing copyright notice that was put on most if not all broadcast stories.

Comments

The conversion to SGML was setup so that the information critical to the TDT task is always surrounded by <P>...</P> tags. Anything outside those tags is extraneous and would not normally be available, with the following exceptions:

- In broadcast stories, the paragraph boundaries, punctuation, capitalization, and font selection would not be known.[1]
- For newswire, the author, dateline, and headline would all be available since it is transmitted by Reuters. It is likely that the date and time would also be specified.

Note that the speaker tag has been carefully setup to follow the <P>...</P> convention. For example:

<SP>BILL TUSH</SP>
<P>And <I>The Postman Always Rings Twice</I> in 1946.</P>

The information that was spoken is between <P> and </P>. Obviously the italics tags, the punctuation, and the capitalization were inferred by the transcriber.

Note that the broadcast news also has editorial comments by the transcriber. For example, "[Unintelligible]" occurs occasionally, as does "[Off camera]" and so on. Text between square brackets is not part of the transcription and should not be used as input to TDT tasks.

Event Definitions

1) Aldrich Ames

SHORT: Ames spy case

LONG: On February 22, 1994 Aldrich Ames (a CIA official) and his wife Maria were arrested and charged with spying for Russia. It was determined that, among other things, Ames passed along the names of CIA agents working in Russia to the KGB, and many were subsequently killed. Ames eventually accepted a plea bargain in which he got a life sentence, but his wife was sentenced to just over 5 years.

2) Carlos the Jackal

SHORT: The arrest of 'Carlos the Jackal'

LONG: "Carlos the Jackal" (Ilich Ramirez Sanchez) has been accused of masterminding several notorious crimes. The French government claimed on August 15, 1994, to have captured him (apparently there were many false claims in the past). He had been wanted for over 20 years.

Anything related to his arrest--including the verification of his identify, his being transported to prison, and so on—are relevant. Stories only about his past activities are not relevant. (Note that there doesn't seem to be much discussion about this except for the few days around the arrest.)

3) Carter in Bosnia

SHORT: Jimmy Carter negotiated cease fire in Bosnia (December 25, 1994)

LONG: Jimmy Carter plays the elder statesman in Bosnia (like he did in Haiti) and successfully negotiates a Christmas time cease fire between Bosnians, Croats and Serbs. The cease fire starts on Christmas day and holds for a while.

Do not confuse this event with other cease fires in Bosnia. This is the only Bosnia activity involving Carter.

---

[1] While use of this information is not allowed as an objective, it was nonetheless used during the pilot study.

4) Cessna on White House

SHORT: The light plane (Cessna) that slammed into the base of the White House.

LONG: On September 12, 1994, a 172 Cessna crashed onto the South Lawn of the US White House. It was flown by Frank Eugene Corder who died in the crash.

The event starts with the crash of the plane. Stories which talk about the event itself, figuring out the identity of the pilot, and efforts to find the motivation for what Corder did, are all relevant. Discussion of white house security concerns spawned by this event are not relevant otherwise.

5) Clinic Murders (Salvi)

SHORT: The murder of clinic workers in Brookline, MA by John Salvi

LONG: On December 30, 1994, a gunman entered two abortion clinics in Brookline, MA and shot 7 people, killing 2. He was arrested the next day after attempting a similar attack on a clinic in Norfolk, VA, and identified as John Salvi, from Hampton, NH.

Relevant stories include reports of the original incident (which, of course, do not mention Salvi's name) and reports of the identification and arrest of Salvi. The trial and conviction are relevant.

6) Comet into Jupiter

SHORT: Comet collision with Jupiter

LONG: The large comet Shoemaker-Levy split into fragments and began raining down on the planet Jupiter on July 16, 1994, generating tremendous scientific interest. It took about a week for all of the fragments to finally disappear.

This is an interesting event because it was completely predicted months in advance. In our data, stories on the collision appear as early as February 1994, describing the preparation of scientists and monitoring equipment.

7) Cuban riot in Panama

SHORT: Cuban refugees riot in Panamanian detention camp. (December 8, 1994)

LONG: Cuban refugees riot in Panamanian detention camp, protesting living conditions and affirming their desire for release. There are casualties.

Not to be confused with related Cuban refugee events in US or in Guantanamo, nor with the general plight of Cuban refugees.

8) Death of Kim Jong Il (N. Korea)

SHORT: Death of Kim Jong Il and his son's succession.

LONG: Kim Il Sung, 82-year-old and long-time leader of Korea, died on Friday morning, July 8, 1994, of a heart attack. The US had been in the midst of negotiations with North Korea to prevent North Korea from becoming a nuclear power. His son is Kim Jong Il, also known as "The Dear Leader". Kim Jong Il was named successor in 1980.

For this event, we are interested in stories which discuss the death of Kim Il Sung and his son's assumption of the power of head of state. The event here is both Kim Jong Il's death and the confirmation that his son is taking control. Any story which discusses Jong's death or the question of the assumption of power, is relevant. Once it becomes clear who has the power of head of state in North Korea, no more stories about that question are relevant. Stories about the funeral are relevant since they are about Kim Jong Il's death.

9) DNA in OJ trial

SHORT: DNA blood analysis in the OJ Murder Trial.

LONG: Prosecutors in the OJ murder trial call DNA experts to testify linking blood on the murder scene with OJ's blood by DNA fingerprinting, a relatively new forensic technique.

Stories that relate directly to the trial and the introduction, presentation, discussion, or counter-arguments directly relating to DNA blood tests should be included. Excluded are background stories on DNA testing, other events in the OJ trial, and other blood testimony not involving DNA.

[Note this event will always be problematic because all of the events in the OJ trial are interwoven. Pre-trial DNA discussions, other non-DNA blood evidence, etc. are confounders. Please do your best.]

10) Haiti ousts observers

SHORT: Haiti throws out 101 human rights observers.

LONG: There were elections in Haiti and Jean-Bertrand Aristide won. However, in 1991 there was a military coup that overthrew Aristide. There was a good deal of complaining about the resulting situation, and human rights observers claimed to have documented more than 3,000 political murders since the coup.

On July 11, 1994, the generals who were ruling Haiti told the human rights observers (about 100 of them) that were there that they had to get out by the following Wednesday. This announcement is the starting point of this event.

They did leave, and the event effectively ends on the 13th, though stories will report it through the 14th. Eventually, the US got rid of the generals and installed Aristide back in power. However, only stories that refer to the actual ousting of the 100 or so UN human rights observers are relevant. The US invasion of Haiti is a different event.

11) Hall's copter (N. Korea)

SHORT: Bobby Wayne Hall's helicopter going down in North Korea (December 17, 1994)

LONG: A Helicopter with two people, including Bobby Wayne Hall, is shot down over North Korea where Hall is captured and held for 17 days, and accused of spying. After he is released, he returns to the US and received a hero's welcome. The incident threatens to worsen relations with North Korea.

The event must be primarily about Hall's capture, ordeal, release.

12) Humble, TX, flooding

SHORT: Flooding in Humble, Texas

LONG: In mid October (around the 19th), there was some severe flooding northeast of Houston "in the Humble Kingwood area", closing route US 59. The storm had been going for about 4 days by that point: 24-30 inches.

Things talking about the flooding and problems with the cleanup are relevant. That includes the oil spill and fires that were caused by the flooding. Federal disaster relief stories are also relevant.

13) Justice-to-be Breyer

SHORT: Stephen Breyer's nomination to the Supreme Court

LONG: Clinton nominated Stephen Breyer to serve on the Supreme Court of the US. His confirmation hearings were begun on July 12, 1994. He was confirmed and was ceremonially sworn into office on August 12, 1994, though the actual swearing-in happened 9 days earlier. He underwent the "new justice" rituals on September 30, 1994.

Relevant stories should discuss the senate confirmation hearings up through his actually joining the Supreme Court. Human interest stories about Breyer are not relevant. There is not likely to be anything interesting after September 30th.

Note that stories about the upcoming hearing (an expected event) are relevant.

14) Karrigan/Harding

SHORT: Nancy Kerrigan assault

LONG: Figure skater Nancy Kerrigan was attacked after competing at the US National Skating Championships on January 6, 1994. A plot to get her out of the Championships, and the Winter Olympics scheduled later in the year, is exposed. The mastermind found to be the husband of Tonya Harding, another US figure skater.

Relevant to this event are stories on the attack on Kerrigan, the investigation that eventually identifies the perpetrators, the arrest of Harding's husband and his accomplices, and their trial and sentencing. This event does not include Harding's subsequent trouble with the US Figure Skating Association, her fight to go to the Olympics, or anything about the performance of Harding or Kerrigan at the Olympics.

15) Kobe Japan quake

SHORT: A severe earthquake rocks Kobe Japan on 17-January-1995.

LONG: In the pre-dawn hours on Jan 17, 1995, a severe earthquake shook the city of Kobe in Southern Japan (population approx. 1.5 million). The quake measured 7.2 in the Richter Scale and caused approx. 5000 deaths, 30,000 injuries, damaged many buildings, highways and train lines. The Kobe-Osaka corridor was particularly hard hit.

The event includes early reports of damage (totaling less than the final tally), location and nature of quake, rescue efforts, consequences of the quake, and on-site reports. Specifically excluded are reports of other quakes, including an earlier one in Northern Japan.

16) Lost in Iraq

SHORT: Iraq jails two Americans who inadvertently cross the border from Kuwait.

LONG: On March 13, 1995, William Barloon and David Daliberti, two Americans visiting friends in the demilitarized zone between Kuwait and Iraq, inadvertently crossed the border into Iraq and were immediately arrested (the first reports did not appear until March 17). They were subsequently tried and sentenced to 8 years in prison. About 4 months later the were released into the custody of a Bill Richardson, a US Congressman.

Relevant stories cover the description of the arrest, conviction, and appeal of the two Americans, the trip to Baghdad by their wives to plea for their release, and negotiations by the US and the possible linking of their release to the lifting of sanctions.

17) NYC Subway bombing

SHORT: The December 1994 bombing on a New York City subway car

LONG: A gasoline bomb goes off on a New York City subway car on December 21, 1994, injuring about 50 people, including the bomber himself. He is quickly identified as Edward Leary, a "computer expert". The bombing is described as part of an extortion attempt.

Anything about the original incident, background on Leary, his identification and arrest, and the trial and conviction are relevant.

18) OK-City bombing

SHORT: The bombing of the federal building in Oklahoma City.

LONG: On April 19, 1995 a large, fertilizer-based bomb hidden in a rented truck outside the federal building in Oklahoma City exploded, destroying the building and killing over 150 people. Two suspects, Timothy McVeigh and Terry Nichols, were subsequently arrested and charged with the crime.

Relevant stories cover: 1) the bombing itself, including the rescue effort and the search for bodies; 2) the investigation, leading to the identification and arrest of McVeigh and Nichols; and 3) court maneuverings by defense and prosecution.

19) Pentium chip flaw

SHORT: Pentium chip flaw

LONG: On November 22, 1994 Intel announced that in the summer of 1994 it had found a flaw in the floating point unit of its Pentium chips. They argued that the flaw was inconsequential for most users, and initially refused to replace the defective chips.

This story actually broke earlier, but didn't hit the mainstream press until November. Relevant stories include the initial report, Intel's attempts at damage control and their evolving decisions about how to satisfy users. Also relevant are discussions about the seriousness of the flaw, and the reactions of computer manufacturers (who actually sold computers with the defective chips and therefore had PR problems of their own), and the effect on Intel's business and stock.

20) Quayle lung clot

SHORT: Hospitalization of Dan Quayle, blood clot in lung

LONG: On November 29, 1994, Dan Quayle was admitted to an Indiana hospital complaining of chest pains. The diagnosis was initially a single blood clot in his lung, but he was later found to have 3 clots.

The story covers his hospitalization, treatment, and release from the hospital about 1 week later. This is short, focused event on a major health issue surrounding an important political figure. Relevant stories discuss his hospitalization and treatment.

21) Serbians down F-16

SHORT: The downing of the F-16 by the Bosnian Serbs and the subsequent rescue of the pilot.

LONG: On June 2, 1995, an American F-16 was shot down by a Bosnian Serb missile. The Serbs claimed to have captured the pilot, though it turned out later that they did not. The pilot, Captain Scott O'Grady was rescued by the Marines on June 8th.

All stories which discuss the downing of the pilot, the struggle to find out what happened, Serbian claims about the downing and the pilot, any counterclaims or discussion of what really happened, the discovery that O'Grady was hiding in the woods, his rescue, and his return to the US are relevant. Discussion of how O'Grady managed to stay hidden, reports of parades, appearances, and so on, after he returns home as a "hero" are also relevant.

22) Serbs violate Bihac

SHORT: The Bosnian Serb's violation of the UN-designated safe area around Bihac.

LONG: The Bosnian Serbs launched an offensive on Bihac around November 11, 1994. Bihac was declared a safe haven by the U.N. at some point earlier (not sure when, but probably not in the second half of 1994).

Stories discussing the Serb attack on Bihac and international reaction are relevant. Stories related to other safe havens in the former Yugoslavia are not relevant.

Pre-attack stories about the related troop movements, fears of an attack, and warnings are relevant.

23) Shannon Faulkner

SHORT: Shannon Faulkner's efforts to be admitted into the Citadel.

LONG: Shannon Faulkner applied to the Citadel, an all-male, state-supported, military school in South Carolina. She was accepted (she didn't indicate her gender), but when they found out she was female, they told her she couldn't come in. She sued. This happened back in 1993 (I think). On July 22, 1994, a federal court ruled that the Citadel had to allow her. Faulkner made plans to go, but on August 13th and appeals court temporarily blocked her admission, though she was allowed to take classes.

The starting point for this event is the court decision allowing her to attend. The ending point is the appeals court rejecting her attendance. The actual hearing for that appeal (scheduled for December) and events after that are not relevant to this event. However, stories in July which refer to the upcoming July 22nd court ruling are relevant.

24) USAir 427 crash

SHORT: Crash of USAir flight 427.

LONG: On September 8, 1994, US Air flight 427 crashed as it was approaching Pittsburgh (from Chicago). 131 passengers were reported on board and they were all killed.

Stories reporting the event are relevant, as are any stories which discuss efforts to determine the cause of the crash. Stories which talk solely about passengers are not relevant. Stories which mention this crash in passing in the context of a larger story (e.g., general aviation safety stories) are not relevant.

25) WTC Bombing trial

SHORT: The trial of Sheik Omar Abdel-Rahman et al on January 9, 1995.

LONG: Sheik Omar Abdel-Rahman is accused of conspiring and plotting a series of bombings targeted at key US buildings and sites, including the World Trade Center Bombing. The focus of this event is his trial, along with 11 co-conspirators, starting January 9, 1995 in New York City. The Sheik is a blind Muslim Cleric, previously accused of plotting to murder Egyptian President Mubarak, and inciting the assassination of radical Rabbi Meier Kahane.

Stories about the Mubarak and Kahane incidents without reference to the bombing trial are not relevant. Stories about the trial, both before and after its start are relevant, including prosecution arguments ("inciting radical muslims to overthrow the US Government" and ordering the bombings), and defense ones (a hapless blind cleric, prosecuted because of his Muslim beliefs). to the trial, etc.

Labeling Instructions

Your task is to label a corpus of news stories as fast as possible without making errors.

The corpus consists of 15,863 stories covering a wide variety of topics drawn from nine broadcast news sources plus one newswire. All of the stories are from 1 July 1994 through 30 June 1995.

Twenty-five labels relating to news events have been defined. Most stories will match none of the label definitions; some will match one; a very few will match more than one. Your task is to apply all relevant labels to each story.

Most of the defined events happened unexpectedly, and the relevant stories will all follow the event. However, some events are expected (e.g., court decisions), and relevant stories will appear both before and after the actual event.

When you see a reference to an event, please check YES (relevant) or if less than 10% of the story discusses the event, check MAYBE[2] (relevant but minor).

The definition of each event includes a helpful, but incomplete summary of what happened. It may also indicate some of the things that would be relevant or irrelevant. Neither list is complete.

You may take advantage of the summary information associated with each story, but cannot rely on that to be complete. Labelable events may be mentioned in the middle of long stories, and stories from certain news sources may end with forward pointers or summaries of "other news".

[2] The “MAYBE” labels were renamed as “BRIEF” after the labeling process was completed, in order to conform to and better represent the meaning of the tag.