

THE TOPIC DETECTION AND TRACKING PHASE 2 (TDT2) EVALUATION PLAN

George Doddington
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025

INTRODUCTION

The purpose of the TDT2 project is to advance the state of the art in Topic Detection and Tracking. The general TDT task domain is to be explored and technology is to be developed in the context of an evaluation-drive R&D paradigm, in which key technical challenges are defined and supported by formal evaluations. This document presents these formal task definitions and the performance measures and procedures to be used to direct the research and evaluate technical capabilities and research progress.

The TDT2 project addresses multiple sources of information, including both text and speech. These sources are namely newswires and radio and television news broadcast programs. The information flowing from each source is modeled as a sequence of stories (and non-stories). These stories may provide information on one or more topics. The technical challenge is thus to identify and to follow the topics being discussed in these stories.

1. TOPICS

In the initial TDT study, conducted during 1996 and 1997, techniques were explored for detecting the appearance of new topics and for tracking the reappearance and evolution of them. Early on in this study, the notion of a topic was modified and sharpened to be an "event", meaning something that happens at some specific time and place. For example, the eruption of Mount Pinatubo on June 15th, 1991 is consider to be an event, whereas volcanic eruption in general is not. Events might be unexpected, such as an airplane crash, or expected, such as a political election.

In the TDT2 project, the notion of a topic as an event will be broadened. Stories will be considered to be "on topic" whenever the story is *directly* connected to the associated event. So, for example, a story on the search for survivors of an airplane crash, or on the funeral of the

crash victims, will be considered to be a story on the crash event. (This is different from the TDT pilot study, where consequential events were considered to be separate events.) Obviously there must be limits to this inclusiveness. (For example, stories on FAA repair directives that derive from a crash investigation probably would not be considered to be stories on the crash event.) As part of this effort to broaden the notion of a topic, topics will also include coherent and topical news foci, even when there is no clear underlying event. (The issue of how to define the notion of a topic is very difficult issue, one which has not been fully resolved and for which there exists no perfect solution.)

2. THE CORPUS

A corpus of text and transcribed speech is being developed to support the TDT2 project. This corpus will span the first half of 1998, January through June, and will include approximately 40,000 stories. There will be 6 sources, including 2 newswires, 2 radio programs and 2 television programs.¹ The corpus will include transcriptions of the radio and TV stories in addition to recordings of the audio signal. Two distinct transcriptions of the audio sources will be provided. These are namely a manual transcription (produced using closed captioning and program transcription data) and an automatic transcription (provided by Dragon Systems).

¹ For newswire sources each story is clearly delimited by the newswire format. For radio and TV sources, however, the segmentation of the broadcast audio into stories is not so clear. Segmentation of audio sources will be performed with an eye toward the TDT task domain, and as a guide, audio sources will be segmented into stories so that each "story" discusses a single topic. It turns out that closed captioning services provide just such a kind of segmentation, and therefore closed captioning practice will be followed.

The transcriptions will be annotated and provided in sgml format. Segment boundary times will be provided for the audio sources, and the resulting segments will be labeled according to whether they are judged as “stories” or “non-stories”.

Only a sub-sample of data from each source will be included in the TDT2 corpus. Each of these samples will be a continuous recording of 20 contiguous stories for newswire sources and of 30- or 60-minute broadcasts for audio sources. Each of these recordings will be stored in its own separate file. (Audio broadcasts will be subdivided, if necessary, into recordings of 30-minute duration, to limit the duration of any specific recording to be no more than 30 minutes.)

A set of approximately 100 target topics will be identified to support the TDT2 research effort. These topics will span the period of data collection uniformly and will also span a spectrum of topic types. Candidate topics will be defined by random selection of stories. Each selected story will serve as an implicit definition of one topic. (This assumes that each story discusses only a single topic.) This candidate set will be winnowed by discarding stories that aren't about topics, as they are defined. Then each remaining implicitly defined topic will be sharpened with a brief textual identification of the topic (in one sentence or less). The TDT2 corpus will be completely annotated with respect to these topics, so that each story in the corpus is appropriately flagged for each target topic it discusses.

The TDT2 corpus will be divided into three parts for research management purposes. The first third of the corpus (i.e., the data collected in Jan/Feb 1998) will comprise the *training* set. This data may be used without limit for research purposes. The middle third of the corpus (Mar/Apr 1998) will comprise the *development test* set. This data will be freely available for testing TDT algorithms, but its use should be restricted to diagnostic purposes (rather than direct corpus-based training purposes). The last third of the corpus (May/June 1998) will comprise the *evaluation test* set. This data will be reserved for final formal evaluation of performance on the TDT tasks at the end of 1998.

The Linguistic Data Consortium (LDC)² is preparing the TDT2 corpus and will make it available to the TDT2 research participants in phases, as early as possible, in order to accelerate the research effort. The corpus will also be made available to the research community at large when it is completed.

3. THE TASKS

The TDT2 project is concerned with the detection and tracking of topics. The input to this process is a stream of stories. This stream may or may not be pre-segmented into stories, and the topics may or may not be known to the system (i.e., the system may or may not be trained to recognize specific topics). This leads to the definition of three technical tasks to be addressed. These are namely the segmentation of a news source into stories, the tracking of known topics, and the identification of unknown topics.

3.1 The Story Segmentation Task

The story segmentation task is defined to be the task of segmenting the stream of data from a source into its constituent stories. The source may be either an audio signal (for radio and TV) or text (for newswire). Segmentation of audio signals may be performed directly on the audio signal itself or on the various textual transcriptions of the audio signal. Segmentation of newswire text will be performed on the concatenation of the story texts alone, without the auxiliary header and format information.

3.2 The Topic Tracking Task

The topic tracking task is defined to be the task of associating incoming stories with topics that are known to the system. A topic is “known” by its association with stories that discuss the topic. Thus each target topic is defined by one or more stories that discuss it. To support this task, a set of training stories is identified for each topic to be tracked. The system may train on the target topic by using all of the stories in the corpus, up through the most recent training story. The tracking task is then to correctly classify all

² The Linguistic Data Consortium
3615 Market Street, Suite 200
Philadelphia, PA, 19104-2608, USA.
Phone: 215/898-0464
Fax: 215/573-2175
ldc@ldc.upenn.edu
<http://www.ldc.upenn.edu>

subsequent stories as to whether or not they discuss the target topic.

3.3 The Topic Identification Task

The topic identification task is defined to be the task of detecting and tracking topics not previously known to the system. It is characterized by a lack of knowledge of the topic to be detected. Therefore the system must embody an understanding of what a topic is, and this understanding must be *independent of topic specifics*. In the topic identification task, the system must detect new topics as the incoming stories are processed.³ The system must then proceed to associate input stories with those topics. Thus this process identifies a set of topics, as defined by their association with the stories that discuss them.

4. THE EVALUATION

To assess TDT application potential, and to calibrate and guide TDT technology development, TDT task performance will be evaluated formally according to a set of rules for each of the three TDT tasks. The general approach to evaluation will be in terms of classical detection theory, in which performance is characterized in terms of two different kinds of errors. These errors are namely misses (in which the target is not detected when it is present) and false alarms (in which the target is falsely detected when it is not present). In this framework, different topics will be treated independently of each other and a system will have separate outputs for each of the target topics.

All sources must be processed together, and processing must be in chronological order. This implies that synchronization information must be supplied to control source selection. This synchronization information will be a simple list of source file names which gives the chronological ordering of the source files. Each source file will contain an uninterrupted recording of contiguous source data, and it will be assumed that there is no temporal overlap between different source files.⁴

³ Decisions may be deferred for a limited period, however.

⁴ It is unlikely that data in different source files will never overlap. This assumption is justified,

Here are the evaluation rules for the three tasks:

4.1 The Story Segmentation Task

Segmentation algorithms will be evaluated in terms of their ability to correctly locate the boundaries between stories. Each file of source data is processed separately, but segmentation output may not be deferred until the end of each file. A certain amount of look-ahead will be allowed, however, and this deferral period is a primary task parameter. (Longer deferral periods should allow better segmentation performance.)

A primary task parameter will be this deferral period, N_d . For text sources N_d is the number of words of deferral allowed, and for audio sources N_d is the time (in seconds) of deferral allowed. The values to be used in the TDT2 project are 100, 1000 and 10,000 words, for text, and 30, 300 and 3,000 seconds for audio.

Segmentation systems under evaluation must record segmentation decisions in an output file, one record for each hypothesized story boundary. The first record in this file will contain three fields which specify information that applies globally to the whole file. These 3 fields will contain:

System Source N_d

where

System is an alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)

Source is the filename of the source file being processed.

N_d is the allowed deferral period, in words for text files and in seconds for audio files.

Each subsequent data record in the file will identify a hypothesized boundary. These records will have only 1 field and will contain:

Boundary

where

Boundary is a hypothesized boundary. For text files, **Boundary** is the index number of the first word in the hypothesized segment, in the range {1, 2, . . .}. For audio files,

however, because of the great simplification provided, coupled with the minor temporal disruption due to the limited time duration represented by each file.

Boundary is the time of the beginning of the segment $\{0.0, \dots\}$. (It isn't necessary to output the beginning of the first segment.) The hypothesized **Boundary** points must occur in chronological order.

Segmentation performance will be measured using a modified version of an error metric suggested by John Lafferty.⁵ This method avoids dealing with boundaries explicitly by measuring the probability that two sentences are correctly classified as to whether they belong to the same story.

Three modifications have been made to Lafferty's error measure:

1. The unit of distance has been changed from the sentences that Lafferty used. TDT will use words for text sources and time (in seconds) for audio sources.
2. The boundary test (as to whether the two sentences/words/times belong to the same story) is made at a fixed distance rather than a probabilistic distance.
3. The error measure is split into miss and false alarm probabilities, so as to represent and evaluate the segmentation task as a formal detection task.

For text sources, distances are measured in terms of words, and the boundary test is made at a separation of k words. Choice of k is a critical consideration in order to produce a meaningful and sensitive evaluation. For the TDT2 project, k will be chosen to be 50 words, and the miss and false alarm probabilities are computed as:

$$\mathbf{P}_{\text{Miss}} = \sum_{s \in \{\text{seg}\}} \left\{ \sum_{i=1}^{N_s-k} \{ \delta_{\text{top}}(i, i+k) \cdot (1 - \delta_{\text{ref}}(i, i+k)) \} \right\} / \sum_{s \in \{\text{seg}\}} \left\{ \sum_{i=1}^{N_s-k} \{ 1 - \delta_{\text{ref}}(i, i+k) \} \right\}$$

$$\mathbf{P}_{\text{False Alarm}} = \sum_{s \in \{\text{seg}\}} \left\{ \sum_{i=1}^{N_s-k} \{ (1 - \delta_{\text{top}}(i, i+k)) \cdot \delta_{\text{ref}}(i, i+k) \} \right\} / \sum_{s \in \{\text{seg}\}} \left\{ \sum_{i=1}^{N_s-k} \delta_{\text{ref}}(i, i+k) \right\}$$

where the summation is over all the words in the entire story text of all the source files in the corpus and where

⁵ "Text Segmentation Using Exponential Models", by Doug Beeferman, Adam Berger, and John Lafferty.

$$\delta_{\text{sys}}(i, j) = \begin{cases} 1 & \text{when words } i \text{ and } j \text{ are deemed by } \text{sys} \text{ to be within the same story} \\ 0 & \text{otherwise} \end{cases}$$

For audio sources, distances are measured in terms of time, and the boundary test is made at a separation of Δ seconds. Choice of Δ is a critical consideration in order to produce a meaningful and sensitive evaluation. For the TDT2 project, Δ will be chosen to be 15 seconds, and the miss and false alarm probabilities are computed as:

$$\mathbf{P}_{\text{Miss}} = \sum_{s \in \{\text{seg}\}} \left\{ \int_{t=0}^{T_s-\Delta} \{ \delta_{\text{top}}(t, t+\Delta) \cdot (1 - \delta_{\text{ref}}(t, t+\Delta)) \} \right\} / \sum_{s \in \{\text{seg}\}} \left\{ \int_{t=0}^{T_s-\Delta} \{ 1 - \delta_{\text{ref}}(t, t+\Delta) \} \right\}$$

$$\mathbf{P}_{\text{False Alarm}} = \sum_{s \in \{\text{seg}\}} \left\{ \int_{t=0}^{T_s-\Delta} \{ (1 - \delta_{\text{top}}(t, t+\Delta)) \cdot \delta_{\text{ref}}(t, t+\Delta) \} \right\} / \sum_{s \in \{\text{seg}\}} \left\{ \int_{t=0}^{T_s-\Delta} \delta_{\text{ref}}(t, t+\Delta) \right\}$$

where the integration is over the entire duration of all the source files in the corpus and where

$$\delta_{\text{sys}}(t_1, t_2) = \begin{cases} 1 & \text{when times } t_1 \text{ and } t_2 \text{ are deemed by } \text{sys} \text{ to be within the same story} \\ 0 & \text{otherwise} \end{cases}$$

In summary, the evaluation of story segmentation may be conducted under a total of up to 12 different conditions. This number is the product of 4 source conditions and 3 deferral periods:

- ◆ Four source conditions:
 - ◇ **newswire text**
 - ◇ audio – **manual transcription**
 - ◇ audio – **automatic transcription**
 - ◇ audio – **sampled data signal**
- ◆ Three segmentation decision deferral periods:

For text sources, in words:

100 1000 10,000

For audio signals, in seconds:

30 300 3,000

4.2 The Topic Tracking Task

Each topic is to be treated separately and independently. In training the system for any particular target topic, allowable information includes the training set and topic flags *for that target topic only*. (No information is given on any other target topic).

A primary task parameter is the number of stories used to define ("train") the target topic, N_t . Evaluation will be conducted for five values of N_t , namely $\{1, 2, 4, 8, 16\}$. In training, N_t will count just the number of YES tags for the target topic, and stories marked BRIEF for the target

topic will be excluded from training. The division of the corpus between training and test will depend on the topic. Specifically, the test set for a particular target topic will be all data (for all sources) that follow the $(N_t)_{Max}^{th}$ training story for the topic. The target topic training data will be the last N_t training stories marked YES for the target topic. More precisely, the training data will be all of the stories up to and including the $(N_t)_{Max}^{th}$ story that discusses that topic, but excluding all stories prior to the $[(N_t)_{Max} - N_t + 1]^{th}$ training story for the topic that are marked YES or BRIEF for the target topic).

Topic Tracking will be performed under two conditions. These are namely that 1) story boundaries are given, and 2) story boundaries are not given. (This will apply only to the test data. Story boundaries will always be supplied for training data.)

The Topic Tracking task is to hypothesize points in the source stream where the target topic is discussed. Topic tracking systems will perform this task by outputting information about these hypothesized points to a file, one record for each putative discussion of the target topic, written in ASCII format⁶. The first record in this file will contain four fields which specify information that applies globally to the whole file. These 4 fields will contain:

System Boundaries N_t Topic

where

System is an alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)

Boundaries is either YES or NO, where YES indicates that story boundaries are supplied to the system being tested and NO indicates that they are not.

⁶ Records will be separated by newline characters, and fields within a record will be separated by white space. The appearance of a # character in a record signals a comment. The first record in a file that begins with a # will be used as a title record. This title record will be displayed along with the evaluation results for that file. Other than this special use as a title, the remainder of a record will be ignored whenever a # character is encountered.

N_t is the number of stories used to train the system to the target topic.

Topic is an index number in the range {1, 2, . . . ~100} which indicates the target topic being tracked.

Each subsequent data record in the file will identify the point in the source stream that discusses the target topic, and a measure of the confidence in the identification. These records will have 4 fields and will contain:

Source Pointer Decision Score

where

Source is the filename of the source file being processed.

Pointer indicates where in the source file the specified topic is being discussed. Topics are associated with a *specific word* for text sources and a *specific time* for audio sources. Thus for text sources **Pointer** is the index number of the specified word, in the concatenation of all story texts for the source file (in the range {1, 2, . . .}). And for audio sources **Pointer** is the specified time, in seconds.

Decision is either YES or NO, where YES indicates that the system believes that the story being processed discusses the target topic, and NO indicates not.

Score is a real number which indicates how confident the system is that the story being processed discusses the associated topic. More positive values indicate greater confidence.

Before identification performance may be evaluated, the system output information must be associated with a story. The evaluation system performs this function by mapping {**Source**, **Pointer**} to the corresponding story, according to the story boundary information provided to the evaluation system. In cases where multiple output records map a story to a given target topic, the record with the highest score will be used. In cases where no output record maps a story to a given target topic, that story will match no target topic.

The topic tracking system may adapt to the test data as it is processed, but only in an unsupervised mode. Supervised feedback is not allowed. (Evaluating over a set of N_t 's provides essentially equivalent information.)

Topic tracking decisions must be made by the end of the current source file. (Decisions may be deferred to the end of the source data file being processed, but no further.)

In calculating performance, those stories tagged as **BRIEF** for the target topic will not be included in the error tally.

In summary, the evaluation of topic tracking may be conducted under a total of up to 40 different conditions. This number is the product of 4 source conditions, 2 boundary conditions, and 5 training conditions:

- ◆ Four source conditions:
 - ◇ **newswire text**
 - ◇ audio – **manual transcription**
 - ◇ audio – **automatic transcription**
 - ◇ audio – **sampled data signal**
- ◆ Two story boundaries conditions:
Given Not given
- ◆ Five different training conditions (# of training stories):
1 2 4 8 16

4.3 The Topic Identification Task

The topic identification task is simply to associate together the stories that discuss each topic. Topic identification will use a whole (2-month) sub-corpus as input. However, identification performance will be evaluated only on those stories which discuss one of the predefined target topics and, therefore, only on those stories for which truth can be ascertained. (In order to ascertain truth, and thus support a meaningful evaluation, it must be assumed that each story discusses, and therefore may be associated with, at most one topic.⁷)

The topic identification system must process the input source data in chronological order (according to the sequence listing of source data files). However, the topic identification system is allowed to defer its identification of topics (and its association of stories with them) until a certain amount of subsequent source data is processed. This deferral period is a primary task

⁷ The assumption that each story discusses only one topic is reasonable for the vast majority of stories. It is being used because it simplifies the task and the evaluation.

parameter. (The greater the deferral, the better will be decisions regarding both the identification of new topics and the association of stories with them.)

A primary task parameter will be this deferral period, N_f . This is the number of subsequent source files that may be processed before committing to an association of a story with a topic. Evaluation will be conducted for three values of N_f , namely 1, 10 and 100.

Topic Identification will be performed under two conditions. These are namely that 1) story boundaries are given, and 2) story boundaries are not given.

The Topic Identification task is to detect topics and then to hypothesize points in the source stream where they are discussed. Topic Identification systems will perform this task by recording information about these hypothesized points in a file, one record for each putative discussion of a topic, written in ASCII format⁸. The first record in this file will contain three fields which specify information that applies globally to the whole file. These 3 fields will contain:

System Boundaries N_f

where

System is an alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)

Boundaries is either YES or NO, where YES indicates that story boundaries are supplied to the system being tested and NO indicates that they are not.

N_f is the deferral period allowed before a decision must be made.

Each subsequent data record in the file will identify a topic, the point in the source stream

⁸ Records will be separated by newline characters, and fields within a record will be separated by white space. The appearance of a # character in a record signals a comment. The first record in a file that begins with a # will be used as a title record. This title record will be displayed along with the evaluation results for that file. Other than this special use as a title, the remainder of a record will be ignored whenever a # character is encountered.

that discusses it, and a measure of the confidence in the identification. These records will have 5 fields and will contain:

Topic	Source	Pointer
	Decision	Score

where

Topic is an index number in the range {1, 2, . . .} which uniquely indicates the topic.

Source is the filename of the source file being processed.

Pointer indicates where in the source file the specified topic is being discussed. Topics are associated with a *specific word* for text sources and a *specific time* for audio sources. Thus for text sources **Pointer** is the index number of the specified word, in the concatenation of all story texts for the source file (in the range {1, 2, . . .}). And for audio sources **Pointer** is the specified time, in seconds.

Decision is either YES or NO, where YES indicates that the system believes that the story being processed discusses the associated topic, and NO indicates not.

Score is a real number which indicates how confident the system is that the story being processed discusses the associated topic. More positive values indicate greater confidence.

Before identification performance may be evaluated, the system output information must be associated with a story. The evaluation system performs this function by mapping {**Source**, **Pointer**} to the corresponding story, according to the story boundary information provided to the evaluation system. In cases where multiple output records map a story to a given target topic, the record with the highest score will be used. In cases where no output record maps a story to a given target topic, that story will match no target topic.

Topic identification performance will be evaluated by measuring how well the stories belonging to each of the target topics match the stories that the system has assigned to the corresponding system-defined topics. This presents a problem, because no correspondence is given between target topics and system-defined topics. This correspondence is determined by associating each system-defined topic with (at

most) one target topic. This will be accomplished by associating each target topic with the system-defined topic that best matches it. The best match is defined to be the one with the smallest number of misses. (In case of ties, the hypothesized topic with the smallest number of false alarms will be the best match.)

Evaluation of topic identification may be conducted under a total of up to 24 different conditions. This number is the product of 4 source conditions, 2 boundary conditions, and 3 deferral periods:

- ◆ Four source conditions:
 - ◇ **newswire text**
 - ◇ audio – **manual transcription**
 - ◇ audio – **automatic transcription**
 - ◇ audio – **sampled data signal**
- ◆ Two story boundaries conditions:
Given Not given
- ◆ Three different decision deferral periods (# of source files to defer):
1 10 100