THE 2002 TOPIC DETECTION AND TRACKING (TDT2002) TASK DEFINITION AND EVALUATION PLAN

1. INTRODUCTION

The purpose of the TDT project is to advance the state of the art in Topic Detection and Tracking. The task domain is multilingual human language. This domain is to be explored and technology is to be developed in the context of an evaluation-guided R&D paradigm in which key technical challenges are defined and supported by common informative evaluations. This document defines the tasks, the performance metrics, and the evaluation procedures to be used to direct the research and to evaluate research progress and technical capabilities.

TDT addresses multiple sources of information, including both text and speech. These sources are namely newswires, radio and television news broadcast programs, and WWW sources. The source languages are English, Mandarin and Arabic. The information flowing from each source is modeled as a sequence of stories. These stories provide information on many topics. The general technical challenge is to identify and to follow the topics being discussed in these stories.

2. TOPIC DEFINITION

In the initial TDT study, conducted during 1996 and 1997, the notion of a topic was limited 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 considered 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 second TDT project, TDT2, the definition of a topic was broadened to include, in addition to the triggering event, other events and activities that are directly related to it. This definition was retained for the third project, TDT3, and will continue for TDT2002. For this year's project:

> ***A topic is defined to be a seminal event or activity, along with all directly related events and activities***.

A story will be considered to be "on topic" whenever it discusses events and activities that are *directly* connected to that topic's seminal 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. Obviously there must be limits to this inclusiveness. (For example, stories on FAA repair directives that derive from a crash investigation would not be considered to be stories on the crash event.) Topic boundaries are subject to interpretation, so the Linguistic Data Consortium (LDC) has created guidelines to improve agreement and consistency of topic labeling.[1]

3. THE CORPUS

LDC[2] is providing three corpora to support TDT research. These are namely the TDT Pilot corpus, the TDT2 corpus and the TDT3 corpus. These corpora are collections of news, including both text and speech, from a number of sources and languages.

The TDT2 corpus spans the first six months of 1998 and is annotated for 200 topics in English, 20 of which have also been annotated in.

The TDT3 corpus spans Oct-Dec 1998 consists of English, Mandarin and Arabic data. There are 120 annotated topics for the English and Mandarin data. A subset of these topics are also annotated against the Arabic data.

Each story in the TDT2 and TDT3 corpora is tagged according to whether it discusses each of the defined topics. These story-topic tags (Tag[story,topic]) assume a value of YES if the story discusses the target topic, BRIEF if that discussion comprises less than 10% of the story, or otherwise NO (the default tag) if the story does not discuss the topic. A complete transcription of audio source data is also provided. These transcriptions include non-news material in addition to news stories. (Non-news stories include commercials and list-type reports such as sports scores and financial data.) Accordingly, stories are labeled either as NEWS or MISCELLANEOUS or, in case there exists no transcription for a story, as UNTRANSCRIBED.

The story type label may **not** be used by systems as side information; systems are expected to process all stories regardless of the story type. However, systems will not be judged using any NON-NEWS stories. The evaluation code will disregard any decisions/scores output for NON-NEWS stories.

The LDC provides multiple representations of the source data for the TDT2 and TDT3 corpora. For audio sources, there are three different representations:

1. the audio sampled data signal
2. a manual transcription of the audio signal
3. a transcription produced automatically by an automatic speech recognition (ASR) system.

For Arabic and Mandarin sources, there are two different text representations:

1. the original language character source stream, which is either source text, or a manual or ASR transcription of speech.[3]
2. An English translation produced automatically from the original language source stream.

---

[1] Determining the limits of TDT topics is often extremely difficult and arbitrary. The question is where to draw the line on including (or excluding) "related" events. The LDC have facilitated this task considerably by identifying certain general types of topics and creating specific boundary determination rules for each of those topic types. These rules are contained in a web-accessible document that LDC uses to instruct and guide annotation. This document may be found at the following URL: http://morph.ldc.upenn.edu/TDT/Guide/manual.front.html.

[2] The Linguistic Data Consortium Phone: 215/898-0464
Email: ldc@ldc.upenn.edu Fax: 215/573-2175
URL: http://www.ldc.upenn.edu/

[3] The ASR transcription includes whitespace delimitation of words. Original source text and manual transcriptions do not.

For complete details on the TDT2 and TDT3 corpora, refer to URL http://www.ldc.upenn.edu/TDT/.

3.1 CORPORA RESOURCES FOR TDT 2002

The 2002 TDT Evaluation will make use of the TDT-Pilot, TDT2 and TDT3 corpora. All three are designated as training resources. Systems may make use of these corpora in any way.

The 2002 TDT Evaluation corpus will be the TDT4 corpus due to be released for the evaluation. The evaluation corpus will be shipped to the participating research sites as specified by the TDT schedule[4].

Participants may supplement their training sources with any other data, however, all additional data must predate the evaluation corpus which begins October 1, 2000.

4. THE TASKS

There are five TDT tasks defined for the 2002 evaluation. They are namely the segmentation of a news source into stories, the tracking of known topics, the detection of unknown topics, the detection of initial stories on unknown topics, and the detection of pairs of stories on the same topic. Of these five tasks, the topic tracking task and the link detection task are considered to be "primary". All sites that choose to participate in the evaluation will be required to perform at least one of these primary tasks, and one or both of these tasks should be the primary focus of sites' TDT research. That is because these tasks represent core technology that is broadly applicable to many different TDT applications.

4.1 THE STORY SEGMENTATION TASK

The TDT story segmentation task is defined to be the task of segmenting the stream of data from a source into its constituent stories. Since original text sources are supplied in segmented form, this task applies only to audio sources. Segmentation of audio signals may be performed directly on the audio signal itself or on the various textual transcriptions of the audio signal. Segmentation is to be performed only in the original (untranslated) language of the source.

4.2 THE TOPIC TRACKING TASK (PRIMARY)

The TDT 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 it. Thus each target topic is defined by one or more stories that are "on" (i.e., that discuss) the topic. To support this task, a small set of on-topic 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.[5] The tracking task is then to classify correctly all subsequent stories as to whether or not they discuss the target topic.

[4] http://www.nist.gov/speech/tests/tdt/tdt2002/sched.htm

[5] Those stories in the training set for each topic that are not identified as being on-topic *are not necessarily off-topic*. As an option, a small number of these stories will be identified and certified as being off-topic. The selected off-topic stories will be specifically those stories that appear to be most similar to the on-topic stories, as determined by a computer algorithm. This information will be provided in order to determine whether such information is useful in improving topic tracking performance.

4.3 THE TOPIC DETECTION TASK

The TDT topic detection 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 detection task, the system must detect new topics as the incoming stories are processed. 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.

4.4 THE FIRST-STORY DETECTION TASK

The TDT first-story detection task is defined to be the task of detecting, in a chronologically ordered stream of stories from multiple sources (and in multiple languages), the first story that discusses an event. This task may be viewed as being essentially the same as the topic detection task. The principal difference is in what the detection system outputs.

4.5 THE LINK DETECTION TASK (PRIMARY)

The TDT link detection task is defined to be the task of determining whether two stories discuss the same topic. Thus, just as in the topic detection task, the system must embody an understanding of what a topic is, and this understanding must be *independent of topic specifics*. The link detection task, however, does not deal with topics explicitly. Thus links are not constrained to segregate stories into a set of orthogonal topics, and there is no presumption that each story discusses one and only one topic.

5. THE EVALUATION

In order to inform TDT research, to guide TDT technology development, and to assess TDT application potential, TDT task performance will be evaluated according to a set of rules for each of the five TDT tasks. Evaluation methodology, parameters and procedures remain unchanged for TDT2002.

All of the TDT tasks are cast as detection tasks. Detection performance is characterized in terms of the probability of miss and false alarm errors (P_{Miss} and P_{FA}). These error probabilities are then combined into a single detection cost, C_{Det}, by assigning costs to miss and false alarm errors:

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{non-target}$$

where

- C_{Miss} and C_{FA} are the costs of a Miss and a False Alarm, respectively,
- P_{Miss} and P_{FA} are the conditional probabilities of a Miss and a False Alarm, respectively, and
- P_{target} and $P_{non\text{-}target}$ are the *a priori* target probabilities ($P_{non\text{-}target} = 1 - P_{target}$).

C_{Det} is the bottom-line representation of TDT task performance that is used to judge TDT systems. This cost measure is often a reasonable measure of application value, and consideration of the application can provide appropriate values for the relative costs of misses and false alarms and the target probability. Because these values vary with the application, C_{Det} will be normalized so that $(C_{Det})_{Norm}$ can be no less than one without extracting information from the source data. This is done as follows:

$$(C_{Det})_{Norm} \quad = \quad C_{Det} \bigg/ \min\left(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{non-target}\right)$$

Thus the absolute value of $(C_{Det})_{Norm}$ is a direct measure of the value (i.e., relative cost) of the TDT system.

There are two reasonable methods of estimating detection error probabilities, called story-weighted and topic-weighted. The story-weighted method assigns equal weight to each decision for each story and accumulates errors over all topics. The topic-weighted method accumulates errors separately for each topic and then averages the error probabilities over topics, with equal weight assigned to each topic. **The topic-weighted method will be used exclusively in TDT2002**, because it provides better estimates of performance.[6]

Input data to the TDT systems comprises the various news sources. This source stream is presented to the processing systems in chronological order, and the various sources are presented together, interwoven so as to preserve chronological order. Source file sequencing will be controlled by means of a list of chronologically ordered source file names. Each source file will contain an uninterrupted sample of source data. It is assumed that there is no temporal overlap between different source files.[7] TDT tasks require processing of all data, including both news and non-news stories. The computation of task performance, however, will exclude system output on non-news stories.

For all five TDT tasks, the system may use knowledge of the source of the data and knowledge of the time of the stories. Specific evaluation rules for the five tasks now follow.

5.1 STORY SEGMENTATION EVALUATION

Segmentation algorithms will be evaluated in terms of their ability to locate correctly the boundaries between stories. Only audio sources will be evaluated. Each file of source data is to be processed separately. Three choices for the type of source data to be processed are shown in Table 1.

Table 1. **Input data alternatives for the story segmentation task (audio sources only)**

Data Type
manual transcription
ASR transcription *(required)*
Original audio data

The segmentation test set will include stories from all sources, English, Mandarin and Arabic. Evaluation, however, will be conditioned on source language.

Segmentation must be performed as the data is being processed, but a certain amount of look-ahead will be allowed before boundaries must be output. Deferral of output is allowed because look-ahead presumably will improve segmentation performance. The deferral period, N_d, is a primary task parameter. For source data in text form, N_d is the maximum deferral period in words (for English and Arabic) or characters (for Mandarin). For source data in audio form, N_d is the maximum deferral period in seconds. The deferral parameter values to be used in TDT2002 are shown in Table 2. Decisions may not be deferred beyond the end of the current source file. (Deferral beyond the current source file is unnecessary because every source file begins and ends on a story boundary.)

Table 2. **Maximum deferral periods for the story segmentation task.**

Text		Audio
English/ Arabic (words)	**Mandarin** (characters)	**English, Mandarin and Arabic** (seconds)
100	150	30
1,000	1,500	300
10,000 *(required)*	**15,000** *(required)*	3,000

Each research site is encouraged to study as many of these conditions as may be productively done. However, all sites that perform the segmentation evaluation are required to perform at least one evaluation under common shared conditions for the story segmentation task. These conditions are given in Table 3.

Table 3. **Required Story Segmentation Conditions**

Parameter	Value
Source language	English, Mandarin and Arabic
Source data type	ASR transcription
Maximum decision deferral period	10,000 words for English and Arabic 15,000 characters for Mandarin

Segmentation performance will be judged by determining how well computed story boundaries agree with reference boundaries. This agreement will be judged with an evaluation interval that is

[6] A major source of variance in error probability estimates is the topic. Therefore, because of the high variability in the number of stories per topic, it is important to reduce the contribution of topic variance by equalizing the contribution of different topics.

[7] It is certain that data in different source files will overlap occasionally. The assumption of no overlap is made, however, because of the great simplification provided. The loss of strict temporal order is judged minor and insignificant, because the time duration represented by each source file is a small fraction of a single day.

swept through the input data.[8] The evaluation interval is chosen to be long enough to include all computed boundaries that might reasonably be associated with a true reference boundary, but short enough to exclude unreasonable associations and multiple reference boundaries (i.e., whole stories). Evaluation is performed by sweeping the evaluation interval through the input source stream and judging the correctness of the segmentation at each position of the interval:

- If there is both a computed boundary and a reference boundary within the interval, then segmentation is judged correct.
- Likewise, if there is neither a computed nor a reference boundary within the interval, then segmentation is judged correct.
- However, a miss is declared if there is no computed boundary within an interval that contains a reference boundary,
- And a false alarm is declared when a computed boundary exists within an interval that doesn't contain a reference boundary.

The duration of the evaluation interval depends on the type of input data being processed. The value to be used for the TDT2002 evaluation is shown in Table 4.

Table 4. **Evaluation interval for story segmentation evaluation**

Data Type	Interval length
Transcription (Mandarin)	75 characters
Transcription (English/Arabic)	50 words
original audio data	15 seconds

Segmentation performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. The evaluation cost parameters to be used for the TDT2002 evaluation are given in Table 5. Choice of P_{target} was based on an analysis of the stories in the TDT2 training corpus. (Note that P_{target} is also related to the length of the evaluation interval.)

Table 5. **Story segmentation evaluation cost parameters**

Parameter	Value
P_{target}	0.3
C_{Miss}	1.0
C_{FA}	0.3

5.2 TOPIC TRACKING EVALUATION

Tracking algorithms will be evaluated in terms of their ability to detect which stories are on-topic and which are not. Topics are to be tracked individually, and each topic is to be treated separately and independently. In training the system for a particular target topic, allowable information includes the training set and topic tags *for that target topic only*. During the evaluation of each target topic, no information is given on any other topic. Evaluation will be over the whole extent of the evaluation corpus.

For each topic tracked, the primary system output will be a DECISION (YES/NO) for each story processed, indicating whether the story is judged to be on-topic or not. In addition to this primary output, the system must also produce a SCORE that indicates the confidence that the story is on-topic for the target topic. This SCORE may then be used to explore the trade-off between miss and false alarm errors.

There will be three alternative choices for the form of the source data to be processed. These are shown in Table 6.

Table 6. **Input data alternatives for the topic tracking task**

Source Data Types
text sources and manual transcription of the audio sources
text sources and ASR transcription of the audio sources
text sources and the sampled data signal for audio sources

A primary task parameter is N_t, the number of stories used to define the target topic. The maximum value of N_t, $(N_t)_{Max}$, will be 4. All of these on-topic training stories are tagged YES for the target topic. The evaluation corpus will be divided chronologically into a different training set and test set for each target topic. The training set will comprise the first part of the corpus, up through the last training story tagged YES for the target topic. In addition, the training data will include story identification of the N_t stories that serve to define the target topic. The test set will comprise the remainder of the corpus that follows. Note that the target topic is defined using the *last* N_t on-topic training stories in the training set for the target topic.

Note that training stories that are not in the set of on-topic stories are not necessarily off-topic. As an option, a small number of these stories, N_n, will be identified and certified as off-topic. These certified off-topic stories will be specially selected by a search algorithm provided by LDC to be among the most similar off-topic stories in the training set.

Another important issue is which language(s) are to be used for topic training. For the TDT 2002 Evaluation, topic training may be performed on English, Mandarin or Arabic sources, but not two at once. Evaluation will be supported for three values of N_t and two values of N_n for each training language. The various possible combinations are shown in Table 7.

[8] The method used here derives from a method developed by Lafferty, et al. ("Text Segmentation Using Exponential Models", by Doug Beeferman, Adam Berger, and John Lafferty).

Table 7. **Language training conditions for topic tracking evaluation**

	English	Mandarin	Arabic
N_t – the # of on-topic stories	1	1	1
	2	2	2
	4	4	4
	Variable, 1-4	Variable, 1-4	Variable, 1-4
N_n – the # of off-topic stories	0	0	0
	2	2	2

Note that there is also a variable N_t condition in Table 7. This condition simulates application situations in which the number of training stories is not controlled. It also represents a conceptually more difficult task, because optimum decisions require that the decision score be normalized across different values of N_t.

Note that there is no knowledge of which cross-language stories in the training data are on-topic, because training is monolingual. Nevertheless, sites are permitted to use the cross-language portion of the training data for training purposes, albeit *without* on-topic story tags being given. Note also that stories not tagged as on-topic are not guaranteed to be off-topic, even in the training data for the given training language.

Independent of the language training condition, the *test* set for each topic will include *all* sources for *all* languages. Evaluation of topic tracking performance will be conditioned on source language. This conditional analysis of tracking output will provide performance results as a function of language and will measure performance separately for cross-language and same-language training/test conditions.

The evaluation of topic tracking performance will support three story boundary conditions, as shown in Table 8.[9] When story boundaries are not provided, tracking output must also identify that portion of the input source data to which the DECISION and SCORE apply. This identification is in terms of word/character indices (for text type sources) or time in seconds for sampled data signals. No restriction is imposed on the temporal quantization of such outputs. (E.g., outputs need not be synchronous with automatically determined story boundaries.) This is handled in evaluation by averaging the tracking output (DECISION and SCORE) over the true extent of the stories and then using these averaged values for evaluation.

Table 8. **Story boundary conditions for topic tracking evaluation**

Story boundary conditions
Reference story boundaries (manually determined correct boundaries)
Automatic story boundaries[10] (automatically determined errorful boundaries)
No story boundaries (the system must provide its own boundaries)

The topic tracking system must conform to the following rules:

- The topic tracking system must process the input source data in chronological order.
- Tracking output must be made as each story is processed. (When story boundaries are not provided, tracking output must be completed by the end of the current source file.)
- Unsupervised adaptation is allowed, using information obtained as the input data are processed.
- No look-ahead is allowed.

Each research site is encouraged to study as many of the above conditions as may be productively done. However, all sites that perform the tracking evaluation are required to perform either the Basic Required condition or the Alternate Challenge condition. These conditions are shown in Table 9.

Topic tracking performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. The evaluation cost parameters to be used for the TDT2002 evaluation are given in Table 10. Choice of P_{target} was based on an analysis of the stories in the TDT2 training corpus.

Table 9. **Required Topic Tracking Conditions**

Parameter	Value
Basic Required Conditions	
Topic training language	English
Source language	All
On-topic training stories	1
Off-topic training stories	0
Source data type	text sources and manual transcription of audio sources
Story boundaries	Reference boundaries

[9] Reference story boundaries will *always* be provided for the training data, regardless of the story boundary test condition.

[10] These automatically-determined story boundaries will be provided with the evaluation test data.

Alternate ("Challenge") Conditions	
Topic training language	English
Source language	All
On-topic training stories	4
Off-topic training stories	2 and 0[11]
Source data type	text sources and ASR transcription of audio sources
Story boundaries	Automatic boundaries

Table 10. **Topic tracking evaluation cost parameters**

Parameter	Value
P_{target}	0.02
C_{Miss}	1.0
C_{FA}	0.1

5.3 Topic Detection Evaluation

Topic detection algorithms define topics by story affiliation. That is, each topic is defined as a set of stories. Thus topic detection algorithms will be evaluated in terms of their ability to cluster together same-topic stories. Evaluation will be over the whole extent of the evaluation corpus.

The primary detection system output will be a TOPIC_ID for each story. This index will uniquely identify the topic that the story discusses. In addition to this primary output, the system must also produce a SCORE that indicates the system's confidence that the subject story is in fact on-topic for the identified topic. This score may then be used to explore the trade-off between miss and false alarm errors.

There will be three alternative choices for the form of the source data to be processed. These are the same alternatives as for topic tracking and are shown in Table 6.

Another important issue is whether the different languages are to be processed together or separately. For the TDT2002 evaluation, three different conditions will be supported, as shown in Table 11.

Table 11. **Language conditions for topic detection evaluation**

Language Conditions
English only
Mandarin only
Arabic only
English, Mandarin and Arabic together *(required)*

[11] Those sites choosing the challenge conditions must actually perform two evaluations, namely one that uses the certified off-topic training stories and one, for contrast, that doesn't.

Topic detection may be performed both with and without knowledge of story boundaries, as shown in Table 8. When story boundaries are not provided, detection output must also identify that portion of the input source data to which the TOPIC_ID and SCORE apply. This identification is in terms of word/character indices (for text type sources) or time in seconds for sampled data signals. No restriction is imposed on the temporal quantization of such outputs. (E.g., outputs need not be synchronous with automatically determined story boundaries.) Story assignment is handled in evaluation by accumulating votes for all TOPIC_ID's over the true extent of each story (in terms of # of words/characters/seconds, as appropriate) and then assigning the dominant TOPIC_ID to the reference story. In case of ties, the voting will be retabulated using score-weighted votes.

The topic detection system must process the input source data in chronological order. However, the topic detection system may defer its identification of topics (and its association of stories with them) until a limited amount of subsequent source data is processed. This deferral period, N_f, is a primary task parameter and is the number of source files, including the source file being processed, for which processing may be completed before committing to and outputting story-topic associations. (The greater the deferral, presumably the better will be decisions regarding both the detection and identification of new topics.) The deferral parameter values to be used in the TDT2002 are shown in Table 12.

Table 12. **Maximum decision deferral periods for topic detection evaluation in number of source files.**

# of source files, including the one being processed
1
10 *(required)*
100

Each research site is encouraged to study as many of the above conditions as may be productively done. However, all sites that perform the topic detection evaluation are required to perform at least one evaluation under common shared conditions for the topic detection task. These conditions are given in Table 13.

Table 13. **Required Topic Detection Conditions**

Parameter	Value
Source Language	English, Mandarin and Arabic together
Source data type	text sources and the TDT3 automatic transcription of audio sources
Maximum decision deferral period	10 source files
Story boundaries	Reference boundaries

Topic detection performance will be evaluated by measuring detection performance separately for each topic, in terms of miss and false alarm errors. Performance will be evaluated only on the set of predefined topics for which the evaluation corpus has been annotated. (These are the only topics for which "truth" is known.) In order to tabulate detection performance, the reference topics must be mapped to the appropriate system output topics to which they correspond. This mapping will be done automatically by choosing for each reference topic the system output topic that it matches best. The best matching system output topic is defined to be that which produces the lowest evaluation cost. This mapping process will result in a favorable and unavoidable bias in the results, because the mapping yields best (lowest cost) performance by construction, whereas some other mapping might be deemed more appropriate by other (subjective) criteria.

Evaluation of topic detection performance will be conditioned on source language. This conditional analysis of detection performance will thus provide performance results as a function of language. Language-conditioned analysis will be performed *after* topic mapping, however. Thus, since the reference topic is mapped only to one (the "best") system output topic, high detection performance will require that system output topics include topic stories from all languages.

Topic detection performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. The detection cost parameters to be used for the TDT2002 evaluation are given in Table 14. Choice of P_{target} was based on an analysis of the stories in the TDT2 training corpus.

Table 14. **Topic detection evaluation cost parameters**

Parameter	Value
P_{target}	0.02
C_{Miss}	1.0
C_{FA}	0.1

It is assumed that each story discusses at most one topic. Stories that are tagged as discussing more than one of the target topics will not be used in computing topic detection performance.[12]

5.4 FIRST-STORY DETECTION EVALUATION

The first story detection task is logically the same as the topic detection task.[13] The evaluation, however, focuses on the specific aspect of detection associated with novel information. To aid in first-story detection research, the TDT3 corpus has been augmented with first-story annotation for an additional 120 topics. Evaluation will be over the whole extent of the evaluation corpus.

The primary first-story detection system output will be a DECISION (YES/NO) for each story processed, indicating whether the story is judged to be the first story of a topic or not. In addition to this primary output, the system must also produce a SCORE that indicates the confidence that the story is a first story. This SCORE may then be used to explore the trade-off between miss and false alarm errors.

The conditions for performing first-story detection are the same as for topic detection, except that the source language is limited to English. There are the same three alternative choices for the form of the source data, as shown in Table 6. There are the same story boundary conditions, as shown in Table 8. And there are the same decision deferral conditions, as shown in Table 12.

Each research site is encouraged to study as many of the above conditions as may be productively done. However, all sites that perform the first-story detection evaluation are required to perform at least one evaluation under common shared conditions for the task. These conditions are given in Table 15.

Table 15. **Required First-Story Detection Conditions**

Parameter	Values
Source Language	English only
Source data type	text sources and the TDT3 automatic transcription of audio sources
Maximum decision deferral period	10 source files
Story boundaries	Automatic boundaries

First-story detection performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. The evaluation cost parameters to be used for the TDT2002 evaluation are given in Table 16. Choice of P_{target} was based on an analysis of the stories in the TDT2 training corpus.

Table 16. **First-story detection evaluation cost parameters**

Parameter	Value
P_{target}	0.02
C_{Miss}	1.0
C_{FA}	0.1

5.5 LINK DETECTION EVALUATION

Link detection algorithms will be evaluated in terms of their ability to determine whether specified pairs of stories discuss the same topic (i.e., are "linked"). Evaluation will be over the whole extent of the evaluation corpus.

The primary link detection system output will be a DECISION (YES/NO) for each pair of stories processed, indicating whether the stories are judged to be on the same topic or not. In addition to this primary output, the system must also produce a SCORE that indicates the system's confidence that the story-pair is linked. This SCORE may then be used to explore the trade-off between miss and false alarm errors.

[12] The assumption that each story discusses only one topic is reasonable for a large majority of stories and greatly simplifies the definition of the detection task and its evaluation. Some stories will likely be tagged as on-topic for more than one of the target topics, however. These stories will be excluded in computing topic detection performance. Such stories will be retained as part of the story stream, however, and will not be eliminated from the corpus during topic detection processing.

[13] Knowing when a story is the first story on a topic implies knowing when a story is *not* a first story.

There will be three alternative choices for the form of the source data to be processed. These are the same alternatives as for topic tracking and are shown in Table 6.

Another important issue is the choice of languages to be processed. For the link detection task, TDT2002 will use English, Mandarin and Arabic source data. Evaluation of link detection will be conditioned on source language. This conditional analysis of link detection performance will tabulate performance separately for cross-language and same-language story pairs.

Table 17. **Language conditions for link detection evaluation**

Language Conditions
English, Mandarin and Arabic together *(required)*

Link detection may be performed both with and without knowledge of story boundaries. For TDT2002, however, evaluation will be supported only for systems that use the given reference story boundaries and produce output with respect to them.

The link detection system must process the input source data in chronological order. However, the link detection system may defer its identification of story links until a limited amount of subsequent source data is processed. This deferral period, N_f, is a primary task parameter and is the number of source files, including the source file being processed, for which processing may be completed before making story-story link decisions. (The greater the deferral, presumably the better will be the link decisions.) The deferral parameter values to be used in TDT2002 are shown in Table 12.

Each research site is encouraged to study as many of the above conditions as may be productively done. However, all sites that perform the link detection evaluation are required to perform at least one evaluation under common shared conditions for the link detection task. These conditions are shown in Table 18.

Table 18. **Required Link Detection Conditions**

Parameter	Value
Source Language	English, Mandarin and Arabic
Source data type	text sources and the ASR transcription of audio sources
Maximum decision deferral period	10 source files
Story boundaries	Reference boundaries

While link detection systems are expected to be capable of making link decisions for all pairs of stories, evaluation of all such decisions is neither practical nor necessary. Therefore evaluation will be limited to a subset of story pairs sufficient to provide reliable estimates of P_{Miss} and P_{FA}. This will keep system output files to a manageable level. System output must be limited to the specified story pairs. The story pairs for which output is required will be listed in chronological order, ordered primarily according to the second (i.e., the most recent or newest) story of the pair and secondarily according to the first (i.e., the oldest) story of the pair. This will facilitate coordination of system output with chronological processing and deferral requirements. The maximum deferral period is with respect to the source file containing the most recent of the pair of stories.

Link detection performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. For the link detection task, P_{target} is the probability that a pair of stories chosen at random discuss the same topic. The evaluation cost parameters to be used for the TDT2002 evaluation are given in Table 19.

Table 19. **Link detection evaluation cost parameters**

Parameter	Value
P_{target}	0.02
C_{Miss}	1.0
C_{FA}	0.1

A IMPLEMENTATION DETAILS

A.1 SOURCE DATA FORMATS

Source data is organized into a sequence of files in chronological order. Each source file contains continuous news content from a single source. For newswire sources, the average length of a file is 20 stories. Newswire source files contain text data in the form of an untagged text stream. For audio sources, the file duration is one program, typically 30 or 60 minutes. Audio source files contain either sampled data that represents the audio waveform[14], or text data in the form of an untagged text stream.

There are four different kinds of text source files. These are namely text sources, manually transcribed audio, automatically transcribed audio, and a machine translation (into English) of the Non-English transcripts. The format of the text source files depends on the kind of text represented and is shown in Table 20.

Table 20. **Text source files formats**

		Text Sources	Audio Sources: Manual Transcripts	Audio Sources: Automatic Transcripts
English	**Original Language**	ASCII, 1 word per record	ASCII, 1 word per record	ASCII, 1 word per record
Mandarin	**Original Language**	GB, 1 character per record	GB, 1 character per record	GB, 1 word per record
Mandarin	**Machine Translation**	ASCII, 1 word per record	ASCII, 1 word per record	ASCII, 1 word per record
Arabic	**Original Language**	UTF-8, 1 word record	UTF-8, 1 word record	UTF-8, 1 word record
Arabic	**Machine Translation**	ASCII, 1 word per record	ASCII, 1 word per record	ASCII, 1 word per record

The various evaluation tasks will have several input files supplied for the evaluation. The first is a task specific index file that contains the names and supplies the sequencing of the source files to be processed. The task specific index files are described in the following sections. The second index file contains auxiliary side information that systems are permitted to take advantage of. The auxiliary information index file is an ASCII text file with one record per file. The format of this file is given in Table 21.

[14] Sampled data files are structured as SPHERE files. SPHERE software and documentation is available at the following URL: http://www.itl.nist.gov/div894/894.01/software.htm.

Table 21. **Auxiliary information index file**

Record Structure	Record 1-N: <Filename> <Source> <Language> <DateStamp>
Field Descriptions	**Filename**: The root source file name, without file type extensions or directory information. (e.g. 19980612_1931_2033_NYT_NYT)
	Source: Name of the data source. (e.g. ABC_WNT, CNN_HDL)
	Language: Language (ENGLISH, MANDARIN or ARABIC) of the original transcript.
	DateStamp: Date and time when the data collection began for this source. The format is: "YYYYMMDD HH:MM:SS" (e.g. "19980612 12:30:14")

A.2 STORY SEGMENTATION I/O FORMATS

Input data: The segmentation test will be directed by an index file containing a list of source files to be segmented. Systems must process the source files in order of occurrence. For the segmentation task, there will be three different index files, corresponding to audio data that is manually transcribed, automatically transcribed, and untranscribed audio. (Machine-translated source files will not be supported for the segmentation task.) The first data record in the index file will contain:

Record Structure	Header Record: # <Task> <PointerType>
Field Descriptions	**Task**: An indication of the TDT task to be run. For story segmentation the field will contain **SEGMENTATION.**
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for segmentation of source data in text form, or **TIME** for segmentation performed directly on the audio signal

Each subsequent data record in the index file will identify a source file to process. These records will have only field:

Record structure	Record 2…: <Source_file>
Field Descriptions	**Source_file**: The file name of the source data file to be processed. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition.

Output data: 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 that specify information that applies globally to the whole file, and the balance of the file will contain boundary decision records. These 3 fields will contain:

Record Structure	Header Record: <System> <N_d> <PointerType>
Field Descriptions	**System**: An alphanumeric character string that uniquely identifies the system being tested.
	N_d: The maximum allowed deferral period, in words for English and Arabic text files, in characters for Mandarin text files, and in seconds for audio files.
	PointerType: The type of boundary indicated in the balance of the file. The possible values are **RECID** or **TIME.**

Each subsequent data record in the file will identify a hypothesized boundary. These records will have two fields and will contain:

Record Structure	Record 2…: <Source_file> <Boundary>
Field Descriptions	**Source_file**: The file name of the source data file to be processed. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition.
	Boundary: a hypothesized boundary. For text files, Boundary is the index number of the first RECID in the hypothesized segment, in the range {1, 2, . . .}.[15] For audio files, Boundary is the time of the beginning of the segment, in the range {0.0, . . .}. (It isn't necessary to output the beginning of the first segment.) The hypothesized Boundary points must occur in chronological order.

A.3 TOPIC TRACKING I/O FORMATS

Input data: Each tracking test will be directed by an experiment control file. This file will document the conditions of the experiment and will contain a list of index files for that experiment, one index file per topic. Topic tracking index files contain a list of topic training stories, a list of certified off-topic training stories, followed by a list of source files for which the target topic is to be tracked.

The **experiment control file** structure is as follows. First comes a header record that defines the experimental condition for which the control file is intended. The header record contains 4 fields:

[15] For English and Arabic text sources, the RECID's are equivalent to words. For manually transcribed Mandarin text sources, RECIDs are equivalent to characters. For ASR transcribed Mandarin sources, RECIDs are equivalent to words.

Record Structure	Header Record: # <SourceType> <TrainLang> <TestLang> <N_tValues>
Des cri	**SourceType**: The information source types, either "nwt+bnasr" or "nwt+bnman".
	TrainLang: The language of the training stories, either "eng" for English, "man" for Mandarin, or "arb" for Arabic
	TestLang: The language of the test source files, either 'mul,nat' for multilingual texts, native transcriptions or 'mul.eng' for multilingual texts, English translations (for non-English languages)
	N_tValue: The value of N_t. For TDT2002, this field will contain one of the following four values: "1", "2", "4", or "V" (for variable).

Each subsequent data record in the experiment control file will contain the filename of a topic tracking index file, one filename per newline-separated record:

Record Structure	Record 2…: <Topic_tracking_index_file>
Field Descriptions	**Index_file**: The filename of a topic tracking index file.

The **topic tracking index** file structure is as follows. First comes a header record:

Record Structure	Header Record: # <Task> <PointerType> <Topic=N>
Field Descriptions	**Task**: An indication of the TDT task to be run. For topic tracking the field will contain **TRACKING**.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for segmentation of source data in text form, **TIME** for segmentation performed directly on the audio signal, or DOCNO for decisions based on corpora story numbers, (as indicated by the 'docno' attribute of each story).
	Topic=N: Declares the topic id number for which the test is to be run. (This is for documentation only. Allowable information for topic training is restricted to the indices of the training stories that follow.)

Then come $(N_t)_{Max}$ records containing the topic training stories, in chronological order. (There will be fewer than $(N_t)_{Max}$ records if there are fewer than $(N_t)_{Max}$ on-topic stories in the corpus.) Only the last N_t of these stories are to be used for training. When the N_t training condition is "variable N_t", all of the topic training stories are to be used. For this condition, special topic tracking index files will be provided that contain a variable number of training stories.

Record Structure	Record 2…($(N_t)_{Max}$ +1): # Topic_training_story <Story_ID> <Source_file> <Begin> <End>
Field Descriptions	**Story_ID**: A character string story identifier. (This is the TDT3 corpus “docno”.)
	Source_file: The file name of the source data file containing the training story. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition and the source.
	Begin: The word index (or time, if the source file contains sampled audio data) of the beginning of the story.
	End: The word index (or time, if the source file contains sampled audio data) of the end of the story.

Then come N_n records containing the off-topic training stories:

Record Structure	Record ($(N_t)_{Max}$ +2)…($(N_t)_{Max}$ + N_n + 1): # Off-topic_training_story <Story_ID> <Source_file> <Begin> <End>
Field Descriptions	**Story_ID**: A character string story identifier. (This is the TDT3 corpus “docno”.)
	Source_file: The file name of the source data file containing the training story. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition and the source.
	Begin: The word index (or time, if the source file contains sampled audio data) of the beginning of the story.
	End: The word index (or time, if the source file contains sampled audio data) of the end of the story.

Finally come the records that identify the source files to process. These records will have two fields:

Record Structure	Record ($(N_t)_{Max}$ + N_n + 2)…: <Source_file> <Begin>
Field Descriptions	**Source_file**: The file name of the source data file to be processed. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition.
	Begin: the word/character index (or time, if the source file contains sampled audio data) of the point in the source at which processing is to begin.[16]

Output 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. The first record in this file will contain five fields that specify information that applies globally to the whole file. These 5 fields will contain:

Record Structure	Header Record: <System> <Boundaries> <N_t> <Topic> <PointerType>
Field Descriptions	**System**: An alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)
	Boundaries: 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_t: The number of stories used to train the system to the target topic.
	Topic: An index number in the range {1, 2, . . . ~100} which indicates the target topic being tracked.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for textual source data or **TIME** for source data in audio form.

Each subsequent data record in the file will identify the beginning point in the source stream of a judgment about whether the target topic is being discussed, along with an associated decision and confidence. This decision and confidence will apply to all subsequent source data until the point specified by the next output data record. These records will have 4 fields and will contain:

[16] Processing begins at the very beginning of almost all source files. However, when the source file includes the end of the training data, tracking doesn’t begin until after the last topic training story.

Record Structure	Record 2…: <Source_file> <Pointer> <Decision> <Score>
Field Descriptions	**Source_file**: The filename of the source file being processed. When the pointer type (via the header field designation) is 'DOCNO', this field is ignored, however a non-space data element must exist as a placeholder.
	Pointer: Indicates where in the source file the subject discussion commences.[17] For textual source data, Pointer is either the beginning RECID in the source file (in the range {1, 2, . . .}) or the story's DOCNO[18] attribute from the corpus. For source data in audio form, Pointer is the beginning time, in seconds.
	Decision: Either **YES** or **NO**, where **YES** indicates that the system believes that the source being processed does in fact discuss the target topic. **NO** indicates not.
	Score: A real number that indicates how confident the system is that the source being processed discusses the associated topic. More positive values indicate greater confidence.

A.4 TOPIC DETECTION I/O FORMATS

Input data: The topic detection test will be directed by an index file containing a list of source files for which topics are to be detected and tracked. Systems must process the source files in order of occurrence. The index files will follow this format:

Record Structure	Header Record: # <Task> <PointerType>
Field Descriptions	**Task**: An indication of the TDT task to be run. For topic detection the field will contain **DETECTION.**
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for source data in text form or **TIME** for audio data.

Each subsequent data record in the file will identify a source file to process. These records will have only one field:

Record Structure	Record 2…: <Source_file>
Field Descriptions	**Source_file**: The filename of the source data file being processed.

[17] Output records must be ordered so that each successive record's value of Pointer is greater than that of its predecessor, thus indicating the termination of the range of the predecessor record's Decision and Score.

[18] When DOCNO is used as pointer type, the decision and score apply only to the specified story. Therefore, DOCNO inventories must match the evaluation corpora's DOCNO inventory. Omitted DOCNO are assigned the decisions NO with a score if –9e99.

Output data: The Topic Detection task is to detect topics and then to hypothesize points in the source stream where they are discussed. Topic Detection 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 format. The first record in this file will contain four fields that specify information that applies globally to the whole file. These four fields will contain:

Record Structure	Header Record: <System> <Boundaries> <N_f> <PointerType>
Field Descriptions	**System**: An alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)
	Boundaries: 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: The maximum deferral period allowed before a decision must be made.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for text stream segmentation or **TIME** for audio segmentation.

Each subsequent data record in the file will identify a topic, the point in the source stream that discusses it, and a measure of the confidence in the identification. This decision and confidence will apply to all subsequent source data until the point specified by the next output data record. These records will have 5 fields and will contain:

Record Structure	Record 2…: <Topic> <Source_file> <Pointer> <Decision> <Score>
Field Descriptions	**Topic**: An index number in the range {1, 2, . . .} which uniquely indicates the topic.
	Source_file: The filename of the source data file being processed.
	Pointer: Indicates where in the source file the subject discussion commences. For textual source data, Pointer is the beginning RECID in the source file (in the range {1, 2, . . .}). For source data in audio form, Pointer is the beginning time, in seconds.
	Decision: Either **YES** or **NO**, where **YES** indicates that the system believes that the source being processed does in fact discuss the target topic. **NO** indicates not.
	Score: A real number that indicates how confident the system is that the source being processed discusses the associated topic. More positive values indicate greater confidence.

A.5 FIRST-STORY DETECTION I/O FORMATS

Input data: The first story detection test will be directed by an index file containing a list of source files for which the first story of new topics are detected. Systems must process the source files in order of occurrence. The file format is identical to the detection index file, with the only exception being the designation of the task. The index files will follow this format:

Record Structure	Header Record: # <Task> <PointerType>
Field Descriptions	**Task**: An indication of the TDT task to be run. For first-story detection the field will contain **FIRST_STORY.**
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for source data in text form or **TIME** for audio data.

Each subsequent data record in the file will identify a source file to process. These records will have only one field:

Record Structure	Record 2…: <Source_file>
Field Descriptions	**Source_file**: The filename of the source data file being processed.

Output data: The first story detection task is to determine whether or not each processed story is the first story of a new topic. The first story detection system will perform this task by outputting one record for each putative decision to a file. The first record in this file will contain four fields that specify information that applies globally to the whole file. These four fields will contain:

Record Structure	Header Record: <System> <Boundaries> <N_f> <PointerType>
Field Descriptions	**System**: An alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)
	Boundaries: 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: The maximum deferral period allowed before a decision must be made.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for text stream segmentation or **TIME** for audio segmentation.

Each subsequent data record in the file will identify a putative decision, the point in the source stream of that decision, and a measure of the confidence in the decision. This decision and confidence will apply to all subsequent source data until the point specified by the next output data record. These records will have 4 fields and will contain:

Record Structure	Record 2…: <Source_file> <Pointer> <Decision> <Score>
Field Descriptions	**Source_file**: the filename of the source data file being processed.
	Pointer: Indicates where in the source file the new subject commences. For textual source data, 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, . . .}). For source data in audio form, Pointer is the specified time, in seconds.
	Decision: Either **YES** or **NO**, where **YES** indicates that the system believes that the source being processed does in fact discuss a new topic. **NO** indicates not.
	Score: A real number that indicates how confident the system is that the source being processed discusses a new topic. More positive values indicate greater confidence.

A.6 LINK DETECTION I/O FORMATS

Input data: The story link detection test will be directed by an index file containing a list of source files to process followed by a list of story pairs for which story link decisions are to be made. Systems must process each story pair independently, and in order of occurrence. The index files will follow this format:

Record Structure	Header Record: # <Task>
Field Descriptions	**Task**: An indication of the TDT task to be run. For link detection the field will contain **LINK_DETECTION.**

Next is the list of source files to process, in chronological order:

Record Structure	Record 2…N_{source}+1: # source_file <Source_file>
Field Descriptions	**Source_file**: The file name of the source data file to be processed. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition.

Each subsequent data record in the file will identify a pair of story ids to process. The first story id will be the first (i.e., oldest) story, and the second story id will be the second (i.e., most recent) story. The list will be sorted by the first column, and then by the second column. Thus the second (and most recent) story in each subsequent record will be in monotonically increasing chronological order. These records will have only two fields:

Record Structure	Record N+3…: <Story_ID_1> <Story_ID_2>
Field Descriptions	**Story_ID_1**: The character string story identifier of the first story. This identifier is the source file and the TDT3 corpus docno, separated by a colon.
	Story_ID_2: The character string story identifier of the second story. This identifier is in the same format as for the first story, namely the source file and the TDT3 corpus docno, separated by a colon.

Output data: The story link detection task is to determine whether or not each story pair is on the same topic, or "linked". The story link detection system will perform this task by outputting one record for each story pair presented in the index file. The order of the decisions in the output file *must* match the index file exactly. The first record in this file will contain two fields that specify information that applies globally to the whole file. These two fields will contain:

Record Structure	Header Record: <System> <N_f>
Field Descriptions	**System**: an alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)
	N_f: The maximum deferral period allowed before a decision must be made.

Each subsequent data record in the file will identify a putative decision on a story pair, and a measure of the confidence in the decision. These records will have 4 fields and will contain:

Record Structure	Record 2…: <Story_ID_1> <Story_ID_2> <Decision> <Score>
Field Descriptions	**Story_ID_1**: The character string story identifier of the first story. (This is the TDT3 corpus "docno".)
	Story_ID_2: The character string story identifier of the second story. (This is the TDT3 corpus "docno".)
	Decision: Either **YES** or **NO**, where **YES** indicates that the system believes that the story pair is linked. **NO** indicates not
	Score: A real number that indicates how confident the system is that the story pair is linked. More positive values indicate greater confidence.