The Topic Detection and Tracking (TDT) Pilot Study Evaluation Plan

Introduction

The purpose of the Topic Detection and Tracking Study is to advance and accurately measure the state of the art in TDT and to assess the technical challenges to be overcome. At the beginning of this study, the general TDT task domain was explored and key technical challenges were clarified. This document defines these tasks and the performance measures to be used to assess technical capabilities and research progress.

Background

The TDT study is intended to explore techniques for detecting the appearance of new, unexpected topics and for tracking the reappearance and evolution of them. During the first portion of this study, the notion of a "topic" was modified and sharpened to be an "event", meaning some unique thing that happens at some point in time. The notion of an event differs from that of a topic both in temporal localization and in specificity. For example, the eruption of Mount Pinatubo on June 15th, 1991 is consider to be an event, whereas volcanic eruption in general is considered to be a topic. Events might be unexpected, such as the eruption of a volcano, or expected, such as a political election.

The TDT study assumes multiple sources of information, for example various newswires and various news broadcast programs. The information flowing from each source is assumed to be divided into a sequence of stories, which may provide information on one or more events. The general task is to identify the events being discussed in these stories, in terms of the stories that discuss them. Stories that discuss unexpected events will of course follow the event, whereas stories on expected events can both precede and follow the event.

The Corpus

A corpus of text and transcribed speech has been developed to support the TDT study effort. This study corpus spans the period from July 1, 1994 to June 30, 1995 and includes nearly 16,000 stories, with about half taken from Reuters newswire and half from CNN broadcast news transcripts. The transcripts were produced by the Journal of Graphics Institute (JGI). The stories in this corpus are arranged in chronological order, are structured in SGML format, and are available from the Linguistic Data Consortium (LDC).[1]

A set of 25 target events have been defined to support the TDT study effort. These events span a spectrum of event types and include both expected and unexpected events. They are documented on the UMass TDT web site. The TDT corpus was completely annotated with respect to these events, so that each story in the corpus is appropriately flagged for each of the target events discussed in it. There are three flag values possible: YES (the story discusses the event), NO (the story doesn't discuss the event), and BRIEF (the story mentions the event only briefly, or merely references the event without discussion. Flag values for all events are available in "tdt-corpus.judgments".[2]

---

[1]

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

[2] Only values of YES and BRIEF are listed, thus reducing the size of the judgment file by two orders of magnitude. (The vast majority of stories have flag values of NO for all events.)

The Tasks

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

The Tracking Task

The tracking task is defined to be the task of associating incoming stories with events known to the system. An event is defined ("known") by its association with stories that discuss the event. Thus each target event is defined by a list of stories that discuss it.

In the tracking task a target event is given, and each successive story must be classified as to whether or not it discusses the target event. To support this task the study corpus will be divided into two parts, with the first part being the training set and the second part being the test set. (This division is different for each event, in order to have appropriate training and test sets.) Each of the stories in the training set will be flagged as to whether it discusses the target event, and these flags (and the associated text of the stories) will be the only information used for training the system to correctly classify the target event. The detection task is to correctly classify all of the stories in the test set as to whether or not they discuss the target event.

A primary task parameter is the number of stories used to define ("train") the target event, N_t. The division of the corpus between training and test will be a function of the event and the value of N_t. Specifically, the training set for a particular event and a particular value of N_t will be all of the stories up to and including the N_t^{th} story that discusses that event. The test set will be all subsequent stories.

The Detection Task

The detection task is characterized by the lack of knowledge of the event to be detected. In such a case, one may wish to retrospectively process a corpus of stories to identify the events discussed therein, or one may wish to identify new events as they occur, based on an on-line stream of stories. Both of these alternatives are supported under the detection task.

Retrospective Event Detection

The retrospective detection task is defined to be the task of identifying all of the events in a corpus of stories. Events are defined by their association with stories, and therefore the task is to group the stories in the study corpus into clusters, where each cluster represents an event and where the stories in the cluster discuss the event. It will be assumed that each story discusses at most one event. Therefore each story may be included in at most one cluster.[3]

On-line New Event Detection

The on-line new event detection task is defined to be the task of identifying new events in a stream of stories. Each story is processed in sequence, and a decision is made whether or not a new event is discussed in the story, after processing the story but before processing any subsequent stories). A decision is made after each story is processed. The first story to discuss an

[3] While it is reasonable that a story will typically discuss a single event, this is not always the case. In addition to multifaceted stories, there are also overlapping events. For example, in the case of the TDT study's corpus and target events, there are 10 stories that have a YES or BRIEF tag for more than one event. One of these (story 8481) has a YES tag for two events (namely *Carter in Bosnia* and *Serbs violate Bihac*). Nonetheless, the assumption that each story discusses only one event will be used, because it is reasonable for the large majority of stories and because it vastly simplifies the task and the evaluation.

event should be flagged YES. If the story doesn't discuss any new events, then it should be flagged NO.

The Segmentation Task

The segmentation task is defined to be the task of segmenting a continuous stream of text (including transcribed speech) into its constituent stories. To support this task the story texts from the study corpus will be concatenated and used as input to a segmenter. This concatenated text stream will include only the actual story texts and will exclude external and internal tag information. The segmentation task is to correctly locate the boundaries between adjacent stories, for all stories in the corpus.

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. In these evaluations, there will be numerous conditions and questions to be explored. Among these are:

- How does performance vary when processing different sources and types of sources?
- How does selection of training source and type affect performance?

In general evaluation will be in terms of classical detection theory, in which performance is characterized in terms of two different kinds of errors, namely misses (in which the target event is not detected) and false alarms (in which the target event is falsely detected). In this framework, different events will be treated independently of each other and a system will have separate outputs for each of the target events. Here are the evaluation rules:

The Tracking Task

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

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 event and will exclude the BRIEF tags. However, the full classification of each story in the training set (i.e., YES, BRIEF, and NO) may be used in training.

All of the stories in the test set must be processed, but there will be two evaluations – one over all of the test data (for each value of N_t), and one over a fixed set of test data, namely the test set corresponding to $(N_t)_{Max} = 16$. (Tabulating performance for a fixed test set as N_t varies will allow a more stable comparison of performance across the various values of N_t, because the test set will be the same for all values of N_t.)

The test stories are to be processed in chronological order, and decisions must be made "on line". That is, the detection decision for each test story must be output before processing any subsequent stories. Decisions may not be deferred.

The event 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.)

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

There will be a TDT tracking trial for each story, and for each event, and for each value of N_t. This will make a total of about 1,000,000 trials. This number is derived by multiplying the number of target events (25) by the average number of test stories (assumed to number about 8,000) by the number of values of N_t (5). Of these trials, less than one percent will be type I (true) trials.

For each trial, there will be two outputs – first, a logical detection indication, YES or NO, indicating whether the system believes that the story discusses the target event; and second, a score indicating how confident the system is in this decision. This confidence indicator will be used to compute a detection error trade-off (DET) between misses and false alarms.

The trials for the tracking task will be recorded in a file in ASCII format, one record per trial, with trials separated by newline characters and with fields in a record separated by white space. Each record will have 5 fields in the following format:

Event N_t Story Decision Score

where

Event is an index number in the range {1, 2, . . . 25} which indicates the target event being detected.

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

Story is the TDT corpus index number in the range {1, 2, . . . 15863} which indicates the story being processed.

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

Score is a real number which indicates how confident the system is that the story being processed discusses the target event. More positive values indicate greater confidence. (Large negative numbers indicate lack of confidence, while large positive number indicate high confidence.)

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.

The Detection Task

The detection task will used the entire TDT study corpus as input. However, detection performance will be evaluated only on those stories which discuss *only one* of the 25 target events and which are flagged as such with a YES flag for that story. There are 1131 such stories.[4]

Retrospective Event Detection

System output for the retrospective event detection task will be the clustering information necessary to associate each of the stories with a cluster. (Each story is constrained to appear in only one cluster.) This information will be recorded in a file in ASCII format, one record per story, with records separated by newline characters and with fields in a record separated by white space. Each record will have 5 fields in the following format:

Cluster N_t Story Decision Score

where

Cluster is an index number in the range {1, 2, . . .} which indicates the cluster (event) affiliation of the story.

N_t is the number of stories used to train the system to the event. (Since this is a detection task, N_t is identically zero, but it is retained in the output format so as to maintain format uniformity across different tasks.)

[4] There are a total of 1382 event flags and 1372 flagged stories. (10 stories were flagged by two events.) However, 240 of these stories were flagged as BRIEF, and one was flagged as YES by two events.

Story is the TDT corpus index number in the range {1, 2, . . . 15863} which indicates the story being processed.

Decision is either YES or NO, where YES indicates that the system believes that the story being processed discusses the cluster event, and NO indicates not. (Again, Decision should always be YES since the story is a member of its cluster, but it is retained in the output format so as to maintain format uniformity across different tasks.)

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

The performance of retrospective detection will be evaluated by measuring how well the stories belonging to each of the target events match the stories belonging to the corresponding cluster. This presents a problem, because it is not known which of the clusters corresponds to a particular target event. Thus it is necessary to associate each target event with (exactly) one cluster to determine this correspondence. This will be accomplished by associating each target event with the cluster that best matches it. The degree of match between an event and a cluster is defined to be the number of stories that belong to both the event and the cluster.

On-line New Event Detection

The on-line new event detection task is to output a new event flag each time a story discusses a new event. Since evaluation is performed only over the set of target events, the small number (25) of type I trials presents a problem for estimating performance. This problem is addressed by artificially changing the corpus so as to multiply the number of type I trials by N_{skip} (with N_{skip} = 10). This is done in the following way:

- The corpus will be processed once after deleting all stories with BRIEF event tags.
- The corpus will be processed a second time after further deleting the first story which discusses each of the target events.
- The corpus will be processed N_{skip} - 1 more times, each time further deleting the subsequent first (next) story which discusses each of the target events, until the first N_{skip} stories discussing each of the target events have been skipped.

System output for the on-line detection task will be a declaration for each story processed. This output is to indicate whether or not the story discusses a new event. This information will be recorded in a file in ASCII format, one record per story, with records separated by newline characters and with fields in a record separated by white space. Each record will have 6 fields in the following format:

Event N_t Story Decision Score j_{skip}

where

Event is an event index number. (Since there is no event affiliation for the on-line detection task, Event is set to zero, but it is retained in the output format so as to maintain format uniformity across different tasks.)

N_t is the number of stories used to train the system to the event. (Since this is a detection task, N_t is identically zero, but it is retained in the output format so as to maintain format uniformity across different tasks.)

Story is the TDT corpus index number in the range {1, 2, . . . 15863} which indicates the story being processed.

Decision is either YES or NO, where YES indicates that the system believes that this story is the first to discuss the event which the story discusses, and NO indicates not.

Score is a real number which indicates how confident the system is that the story being processed is the first to discuss the event. More positive values indicate greater confidence.

j_{skip} is the number of initial stories that have been skipped for each of the target events, in the range $\{0, 1, \ldots N_{skip}\}$.

The Segmentation Task

Segmentation will be evaluated in two different ways. First, segmentation will be evaluated directly in terms of its ability to correctly locate the boundaries between stories. Second, segmentation will be evaluated indirectly in terms of its ability to support event tracking and preserve event tracking performance.

For the segmentation task, all of the TDT study corpus will be reserved for evaluation purposes. This means that any material to be used for training the segmentation system must come from sources other than the TDT study corpus. Also, the nature of the segmentation task is that the segmentation is performed on a single homogeneous data source. Therefore, for the purpose of evaluating the segmentation task, segmentation shall be performed on separately on the two separate sub-streams of the TDT corpus – one comprising just the Reuters stories, and the other comprising just the CNN stories. In addition, the segmentation task must be performed without explicit knowledge of the source of the text, whether from newswire or transcribed speech.

Direct Evaluation of Segmentation

Segmentation will be evaluated directly using a modification of a method suggested by John Lafferty.[5] This is an ingenious method that avoids dealing with boundaries explicitly. Instead, it measures the probability that two sentences drawn at random from the corpus are correctly classified as to whether they belong to the same story. For the TDT study, the calculation will be performed on words rather than sentences.[6] Also, the error probability will be split into two parts, namely the probability of misclassification due to a missed boundary (a "miss"), and the probability of misclassification due to an extraneous boundary (a "false alarm"). These error probabilities are defined as

$$\mathbf{P}_{\mathbf{Miss}} = \sum_{i=1}^{N-k} \delta_{hyp}(i, i+k) \cdot \left(1 - \delta_{ref}(i, i+k)\right) \Big/ \sum_{i=1}^{N-k} \left(1 - \delta_{ref}(i, i+k)\right)$$

$$\mathbf{P}_{\mathbf{False\ Alarm}} = \sum_{i=1}^{N-k} \left(1 - \delta_{hyp}(i, i+k)\right) \cdot \delta_{ref}(i, i+k) \Big/ \sum_{i=1}^{N-k} \delta_{ref}(i, i+k)$$

where the summations are over all the words in the corpus and where

$$\delta(i, j) = \begin{cases} 1 & \text{when words i and j are from the same story} \\ 0 & \text{otherwise} \end{cases}$$

---

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

[6] There are several reasons for using words rather than sentences. First, there will likely be less debate and fewer problems in deciding how to delimit words than how to delimit sentences. Second, the word seems like a more suitable unit of measurement, because of the relatively high variability of the length of sentences.

Choice of k is a critical consideration in order to produce a meaningful and sensitive evaluation. For the TDT study corpus, k will be chosen to be 250 words, which is about half of the average number of words in a story.

Indirect Evaluation of Segmentation

Segmentation will be evaluated indirectly by measuring event tracking performance on stories as they are defined by automatic segmentation means. A straightforward three-step procedure will be used:

1. Segment the whole corpus using the segmentation system under test.
2. Using an event tracker that has been evaluated on the TDT corpus previously, run this system on the auto-segmented corpus. Follow the standard event tracking rules, with the following exceptions:
   - Train the event tracking system on the original correctly segmented stories.
   - Begin evaluation on the first auto-segmented story which follows the last training story and which is disjoint from it.
3. Evaluate the event tracker results and compare these results with results on the original correctly segmented stories.

The evaluation is complicated by the fact that there are no event flags for the auto-segmented stories. This problem will be solved by creating scores for the original stories from those computed for the auto-segmented stories.[7] The evaluation will then be performed on the original stories using these synthetic scores. The synthetic score for each original story will be a weighted sum of the scores for all overlapping auto-segmented stories, where the weight prorates each score according to how many words the overlapping story contributes:

$$\text{Score}_{orig(i)} = \sum_{j \in overlap(i)} w_{ij} \cdot \text{Score}_{autoseg(j)} \Big/ \sum_{j \in overlap(i)} w_{ij}$$

where

w_{ij} = the # of words in the jth auto-segmented story that overlap with the ith original story.

The output record format will be the same as for the conventional event tracking task. The decision will be computed in the standard way and will be based on the synthetic score.

Speech Tracking – the TWA 800 crash event

In addition to the TDT study corpus, an additional corpus will be processed to explore the tracking task for different representations of speech, including machine recognition of speech. The corpus consists of CNN recordings and spans the period during 1996 when the TWA 800 crash occurred. This corpus contains a total of 1400 stories, of which 18 discuss the TWA crash. Three different representations of the speech will be processed. These are:

---

[7] There are two possible ways of solving this problem – either by mapping the event flags for the original stories onto the auto-segmented stories, or by mapping the decisions on the auto-segmented stories onto the original stories.

Mapping event flags onto the auto-segmented stories might seem to represent the actual application scenario more accurately. Mapping scores was chosen, however, to facilitate a clearer comparison with results on the original stories and to avoid conceptual and mechanical difficulties involved in mapping the event flags.

- The JGI transcripts
- Closed captioning taken from the CNN broadcast
- Speech recognition output provided by CMU