Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation

Charles L. Wayne
Department of Defense
Ft. Meade, MD 20755-6514
clwayne@nist.gov

Abstract

Topic Detection and Tracking (TDT) refers to automatic techniques for locating topically related material in streams of data such as newswire and broadcast news. DARPA-sponsored research has made enormous progress during the past three years, and the tasks have been made progressively more difficult and realistic. Well-designed corpora and objective performance evaluations have enabled this success.

Introduction

This paper has two goals: To report on the substantial progress being made in Topic Detection and Tracking research and to explain the vital roles that common corpora and formal evaluation have played in that success.

Basic Idea

Topic Detection and Tracking (TDT) refers to a variety of automatic techniques for discovering and threading together topically related material in streams of data such as newswire and broadcast news.

Figure 1 illustrates a prototypical situation. Horizontal lines represent incoming streams of news stories from different sources, media, and languages. Each rectangle represents a single story, and all are about the same event. TDT aims to discover this kind of structure automatically.

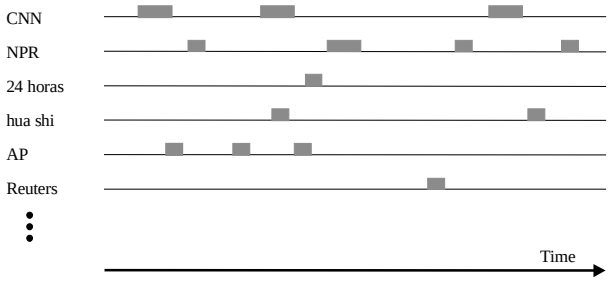

Figure 1: Stories on One Topic (Event) in Several Media

Value

Such automatic discovery and threading could be quite valuable in many applications where people need timely and efficient access to large quantities of information. Systems could alert users to new events and to new information about old events. By examining one or two stories, a user could decide whether to pay attention to the rest of an evolving thread. Similarly, a user could go to a large archive, find all the stories about a particular event, and learn how it evolved.

Related Technology

TDT intersects with, but goes well beyond the traditional concerns of Information Retrieval, Information Management, and Data Mining – particularly in TDT's emphasis on discovering *new* information and its focus on specific *events* rather than subject matter categories.

TDT builds upon, but does not include research on, automatic speech recognition (ASR) technology to convert speech to text and machine translation (MT) technology to convert text from one language to another.

Concise History

The basic idea for TDT originated in 1996, when the Defense Advanced Research Projects Agency (DARPA) realized that it needed technology to determine the topical structure of news streams without human intervention.

In 1997, a pilot study laid the essential groundwork, producing a small corpus and establishing feasibility. During 1998 and 1999, TDT research blossomed, with new and more challenging tasks, many more participating sites, and considerably larger multilingual corpora (adding ASR data in 1998 and Chinese data in 1999).

TDT research is continuing under the new DARPA program known as TIDES (Translingual Information Detection, Extraction, and Summarization). New sites are welcome to participate in the annual TDT evaluations conducted by the National Institute of Standards and Technology (NIST).

Research Paradigm

TDT research follows the focused research paradigm that has powered a variety of successful DARPA Human Language Technology research efforts over the past 13 years. The paradigm features formal research tasks, common (shared) data, and common evaluations.

The formal task definitions focus research on core technical challenges that could benefit many applications. The corpora capture real world challenges, enabling innovative research and objective performance evaluation. The common evaluations make it possible to compare different approaches for solving core technical challenges and to establish periodic performance benchmarks.

Outline of Paper

This paper describes the TDT research tasks and corpora, discusses the research and evaluation conducted in 1999, and outlines the plans for 2000 and beyond.

Since TDT research has been very much a community effort – aided by the strong collaboration of researchers, evaluators, corpus creators, and sponsors – the term “we” below encompasses a great many individuals.

Research Tasks

Overall Goal

TDT research aims to devise powerful, broadly useful, fully automatic algorithms for determining the topical structure of human language data. These algorithms must be source, medium, domain, language, and application independent.

Technical Tasks

We factored TDT into five technical *tasks*:

- Finding topically homogeneous regions (*segmentation*)
- Finding additional stories about a given topic (*tracking*)
- Detecting and threading together new topics (*detection*)
- Detecting new topics (*first story detection*)
- Deciding whether stories are on the same topic (*linking*)

Figure 2 illustrates the basic notion behind each task:

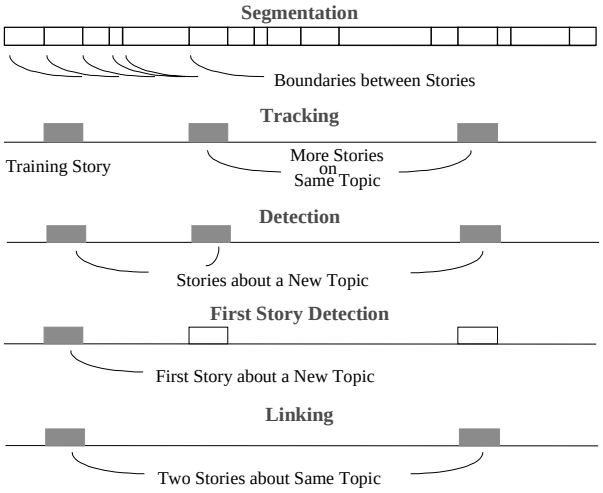

Figure 2: TDT Tasks (matching top line in Figure 1)

Algorithms addressing the first four tasks could be used jointly or separately in a variety of applications. Linking is a more fundamental task that potentially serves the other four. All complement (hence, could be combined with) traditional Information Retrieval technology.

Distinguishing Features

Several things set TDT apart.

Meaning of "Topic"

TDT defines "topic" to mean a specific *event or activity* plus directly related events or activities. (For instance, the Oklahoma_City_Bombing topic includes the destruction of the federal building in 1995, the memorial services, the state and federal investigations, the prosecution of Timothy McVeigh, et cetera.)

This definition sets TDT apart from other topic-oriented research that deals with *categories* of information (e.g., bombings in general). It also makes the TDT problem tractable. Dealing with categories would have required an arbitrary categorization scheme; slower, more subjective annotation; and different technical tasks.

Discovery of New Events

Another thing that sets TDT apart from other research is its emphasis on discovering *new* events (topics) — events that no one expected or knew how to request.

Focus on Linguistic Content

While the TDT corpora contain a great deal of ancillary information, TDT algorithms are allowed to employ only the *content* of the data — plus information about source, date, and time that would generally be available in applications. (In the case of audio data, content includes ASR output and whatever else can be extracted automatically from audio signals.)

In real applications, one would naturally exploit all available knowledge sources. For example, newswire has headers; television news contains video information and possibly closed captions. However, since we are trying to develop general-purpose language-based algorithms, we exclude these other aids, which would dilute and defocus the effort.

Evaluation Methodology

To calibrate progress and provide diagnostic feedback to researchers, we worked out objective procedures for evaluating algorithm accuracy for each of the TDT tasks. This also helped to clarify the research goals.

Specification

TDT evaluation details evolved over the past three years, as we became more experienced, algorithms more capable, and the sponsor increased the research challenge. The latest evaluation plan is (Doddington, 1999).

To represent the research challenges in simple terms, it formulates each task as a classical statistical detection problem, where at each decision point (story or putative boundary), a system must output both an actual decision and a confidence score.

Calculation of Performance

From the scores, NIST software produces *detection error tradeoff* (DET) curves (Martin et al., 1997) like the one illustrated in Figure 4. DET curves show the tradeoffs between miss and false alarm rates at multiple operating points and make it easy to compare results obtained with different algorithms or under different conditions. (We chose miss-false alarm in lieu of precision-recall to emphasize the importance of minimizing errors and to avoid the confounding effects of target richness in different corpora.)

From an actual decision, NIST software calculates a *normalized cost* from the miss and false alarm rates associated with that decision, predetermined costs for misses and false alarms, plus an a priori probability for the target condition (e.g., a story being on topic). A normalized cost, ranging from 0.00 (perfect) to 1.00 or more, reflects both the overall strength of an algorithm and its ability to set thresholds correctly.

Because the number of on-topic stories varies widely and topic difficulty is a major source of variability, NIST generally computes *topic-weighted* results (wherein each topic contributes equally to the overall averages) to improve the reliability of the performance measures.

Common Corpora

Good corpora are *extremely* important for both research and evaluation. With known ground truth, researchers can run many experiments inexpensively, exploring new ideas, comparing results, and learning from one another.

DARPA has done the world an *enormous* service by funding the creation of many very valuable corpora (including the TDT corpora) and making them available via the Linguistic Data Consortium (LDC). Though often expensive to produce, these corpora are absolutely essential for research. They are a bargain in the long run, frequently reused in other research projects and enabling them to start up quickly.

TDT Corpora

In the case of TDT, there are now three corpora: TDT1, created by the pilot study participants in 1997; TDT2 and TDT3, created by the LDC to support the 1998 and 1999 evaluations. All are available from the LDC.[1]

TDT1 contains newswire plus high quality transcripts of news broadcasts, all in English. TDT2 and TDT3 contain Chinese as well as English, audio as well as text, ASR outputs plus closed caption quality transcripts from audio data, and MT outputs from Chinese.

All three corpora are *completely* annotated in terms of selected topics (events) — 25 topics for TDT1, 100 topics for TDT2, 60 topics for TDT3. This is an important difference between the TDT corpora and those created by other communities, and it is part of what has enabled TDT's rapid progress.

TDT1 contains 15,863 stories produced by Reuters and CNN from July 1994 through June 1995. TDT2 covers January – June 1998; TDT3, October – December 1998. Table 1 gives an idea of the rich diversity of data in TDT2 and TDT3 – plus its scale – 116,012 stories in all:

TDT2 Stories	Sources	TDT3 Stories
12760	AP Worldstream	7338
11795	NY Times News Service	6871
2913	PRI The World	1575
8214	VOA English News Service	3948
2153	ABC World News Tonight	1012
15785	CNN Headline News	9003
	MSNBC News w/ Brian Williams	683
	NBC Nightly News	846
11286	Xinhua News Service	5153
5170	Zaobao WWW News Service	3871
2265	VOA Mandarin News Service	3371
53620	Total English	31276
18721	Total Chinese	12395

Table 1 – Distribution of Stories in TDT2 and TDT3

Corpus Contents

TDT2 and TDT3 contain the following types of information:

For text sources —
- Text body
- Ancillary data (e.g., titles, subject categories, slug lines, bylines, filing location)

For audio sources —
- Audio signal
- ASR transcript
- Manual transcript (of closed caption quality, for stories and sometimes other material; of high quality from certain sources)

For Chinese sources —
- MT translation into English

For all sources —
- Origin (medium, source, date, time)
- Boundaries (with time stamps)
- News / Non-News tags
- Part of speech tags
- Name tags
- Topic tags (YES, NO, BRIEF for chosen topics)

Depending on its type, the information is encoded in SGML markup, tables, et cetera. The text data is in both reference form (as received with story boundaries added) and tokenized form (with one word per line like ASR output and no metadata).

Corpus Creation Process

Wayne (1998) explains how TDT1 was created. Here is a brief summary of how TDT2 and TDT3 were created. More detailed information appears in Cieri et al. (2000) and Strassel & Graff (2000).

Sources

We selected a diverse set of sources — including text (newswire, web) and speech (radio, television) from multiple languages (English, Chinese) — for which the LDC was able to negotiate data rights.

Sampling

The LDC sampled the various data streams several times every day (for sources that were available that often) and recorded each sample in a separate file.

For audio sources, a file consists of a whole broadcast (typically 30-minutes). For text sources, a file consists of as many stories as needed (approximately 20) to get a comparable amount of text.

Segmentation and Categorization

For text sources, the LDC decided whether or not each unit of text was a valid news story. For audio sources, they first identified where boundaries lay, then decided whether each intervening stretch of data was a news story.

Commercials, previews (as in "Coming up next ..."), and lists (e.g, of currency prices) that are not part of larger stories were classified as "non-news."

Manual Transcription

When manual transcripts of audio data were available (e.g., closed captions for certain television broadcasts, scripts or good quality transcripts for other shows), the LDC included them in the corpus. When no manual transcripts were available, the LDC produced manual transcripts of roughly the same quality as closed captions.

[1] http://www.ldc.upenn.edu/Projects/TDT

Automatic Transcription

All of the audio files were automatically transcribed by Dragon Systems, using their software, or by NIST, using BBN software. Word error rates varied widely. The average error rates were approximately 25-30%.

Machine Translation

The LDC converted all of the Chinese text (including ASR outputs) to "English" using Systran MT software.

Automatic Labeling

On the text and on the ASR and MT outputs, the LDC ran automatic part-of-speech tagging software and BBN ran name tagging software.

Topic Selection

Since it was not feasible to name and locate every topic, the LDC selected 100 topics for TDT2 and 60 for TDT3. Lead annotators chose stories at random from the various sources and wrote topic descriptions from suitable ones.

For TDT2, they chose topics from English stories only, subsequently searching for them in Chinese. For TDT3, they made sure that each topic appeared at least 4 times in both English and Chinese sources.

This is a typical topic description:

Hurricane Mitch

Seminal Event

WHAT: Hurricane Mitch forms over warm ocean waters, killing thousands and causing millions of dollars in damage.

WHERE: The Caribbean and surrounding areas, particularly Honduras, Nicaragua and Central America.

WHEN: Mitch forms in late September 1998, and lasts through the month of October.

Topic Explication

Hurricane Mitch was the most destructive Atlantic hurricane since 1780, killing over 10,000 people in Central America and leaving millions homeless. ON TOPIC: coverage of the disaster itself; estimates of damage and reports of loss of life; relief efforts by the Red Cross and other aid organizations; impact of the hurricane on the economies of the effected countries.

Rule of Interpretation Rule 4: Natural Disasters

Examples – tornado, snow and ice storms, floods, droughts, mud-slide, volcanic eruptions. The event would include causal activity (El Nino, in many cases this year) and direct consequences. The topic would also include; the declaration of a Federal Disaster Area, victims and losses, rebuilding, any predictions that were made, evacuation and relief efforts.

Figure 3 – Description of TDT3 Topic 2
Plus General Rule of Interpretation 4

Topic Annotation

Annotators read each story and labeled it as YES, NO, or BRIEF with respect to each topic (with BRIEF signifying that the topic was mentioned, but occupied less than 10% of the story).

Additional Topic Annotation

To support research on first story detection, additional topics were partially annotated (with special attention to finding the first stories). To support research on story linking, pairs of stories were annotated as to whether or not they mentioned the same topic (without an explicit topic selection step).

TDT 1999

This section outlines the scope, procedures, approaches, and results of the TDT 1999 research effort. Allan et al. (2000) provides additional detail. Various participants will publish papers with considerably more detail.

Overview

Participants

Eleven academic and industrial research sites:

BBN
Dragon Systems
General Electric
IBM
MITRE
Carnegie Mellon University
National Taiwan University
University of Iowa
University of Maryland
University of Massachusetts
University of Pennsylvania

participated in TDT 1999.

As a minimum, each site conducted research and submitted formal evaluation results for one or more of the tasks. Most sites also participated in two dry runs, four meetings, and various e-mail discussions. They also attended the TDT 1999 Workshop held in February 2000.

Administration

NIST organized and disseminated the test material, which included the TDT3 corpus plus appropriate index files for each task. Sites had three weeks to run the material through their systems and submit their results to NIST.

Test Conditions

The evaluation procedures are spelled out clearly and in considerable detail in (Doddington,1999).

For every task, the test material consisted of both text and speech data. Systems could use either the actual speech data or the ASR output. (All systems used ASR output; some added measurements from the audio signal.)

For the first three tasks, the test material consisted of both English and Chinese. For tracking and detection, systems could use either the actual Chinese text/ASR output or the Systran version. (All systems ended up using the Systran MT output, but several experimented with their own versions of rough MT.)

For every task except segmentation, systems were given true (known) boundaries.

For every task except tracking, systems were allowed to defer their decisions briefly.

Segmentation

Task

The segmentation task required systems to find all story boundaries. The systems ran on audio sources only, and could wait until the end of a broadcast to output decisions.

Approaches

The most successful system combined maximum entropy and decision tree models fed by various source-specific features, including speaking rate (TV announcers speak faster at the beginning of stories than at the end), sentence length (longer at the beginning of stories), position in the show (when commercial breaks appear at predictable times), and word/character n-grams.

Other systems employed Bayes classifiers, various lexical cues (pre and post boundary trigger words plus words appearing on both sides of a boundary), pause durations, and changing energy levels.

Results

The best system had normalized segmentation costs of 0.39 for English and 0.32 for Mandarin. These figures are equivalent to fixed offset errors of 3.5 seconds and 3.4 seconds. The English cost was 19% lower than the lowest cost obtained in 1998 on TDT2.

A side experiment found that segmentation results were quite similar for manual (closed caption) and automatic (ASR) transcripts.

Tracking

Task

The topic tracking task required systems to find all later stories about a topic, given one or more training stories about that topic. This is similar to query-by-example in IR parlance.

For each topic, systems were given four training stories in English plus a large number of test stories in English and Chinese containing an unspecified number of on-topic stories. Systems had to output decisions on a story-by-story basis.

Approaches

The most successful system used logistic regression to combine probabilities from a topic spotting technique (with a language model built from concatenated training stories) and an information retrieval technique (in which the unknown story produces the model) followed by normalization (with thousands of known off-topic stories) and adaptation (with high scoring test documents added to the training set and parameters re-estimated).

Other systems used cosine-vector similarity measures, word feature vectors (with and without stop words, sometimes heavily pruned), name recognition, tf*idf weighting (where idf may be adapted incrementally), Rocchio classification (with positive and negative examples), k-Nearest Neighbor (kNN) clustering, language models based on Hidden Markov Models, source and language-dependent normalization, plus various score combination methods.

Results

The best system obtained a normalized tracking cost of 0.092. Figure 4 shows the corresponding DET plot. Figure 5 shows how the results depend on language and medium.

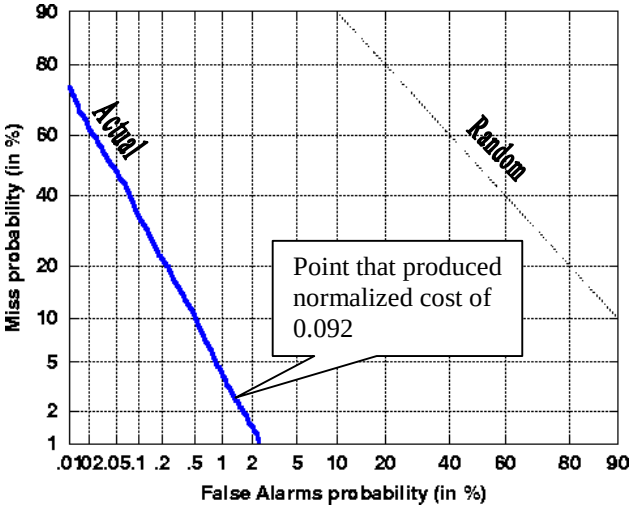

Figure 4 – Tracking Results for Best System

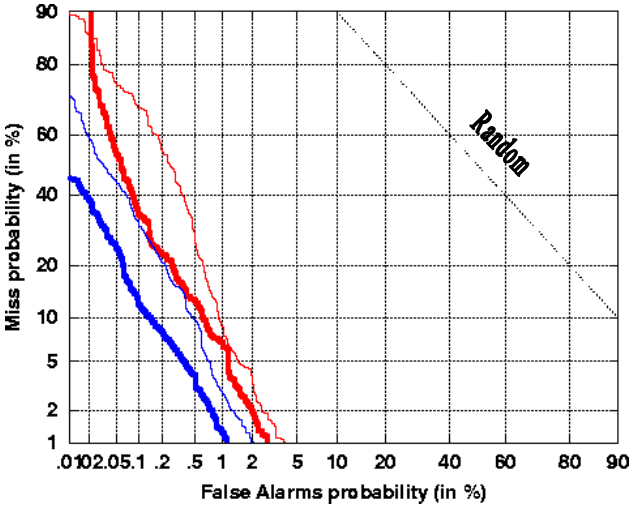

Figure 5 – Tracking Results for Best System
Broken Out by Language and Medium
(Thicker curves for text sources, thinner for audio.
Left pair for English, right pair for Chinese.)

The results are better on text sources than audio sources. This is not surprising, and may be due to content differences (text stories are longer on average) and/or ASR output errors.

Since the training samples were in English, the English results are understandably better than the Chinese results. On the other hand, monolingual experiments using Chinese training and test data produced results comparable to those using English training and test data:

	English Test	Chinese Test
English Training	0.077	0.111
Chinese Training	0.115	0.080

Table 2 – Normalized Tracking Costs
Within & Across Languages

Although cross-language results are not as good as monolingual results, it appears that the basic technology is portable to diverse languages.

We should be somewhat cautious in interpreting these results, however, since the test data contained more audio than text in English, more text than audio in Chinese.

Looking only at the English results, we see an impressive 72% reduction in normalized tracking costs from 1998 to 1999.

Detection

Task

The topic detection task required systems to group incoming stories into topic clusters, creating new clusters (topics) as needed. It was basically unsupervised clustering with limited look ahead.

Systems were given a stream of stories in English and Chinese and could wait until the end of 10 files (broadcasts or a comparable amount of text) before announcing decisions. Once made, decisions were irreversible.

The NIST evaluation software took the system outputs (arbitrary topic IDs), matched as many as possible with the 60 known topics, and calculated performance. (Most clusters did not match a known topic and were ignored.)

Approaches

The most successful system compared each incoming document with all existing clusters using a symmetric Okapi formula and (depending on a threshold) either added the story to the closest cluster or started a new one. It took advantage of the 10-file deferral period to first form microclusters, calculate a source-dependent idf, and then rescore.

Other systems used other IR matching methods or topic spotting language models and normalized twice (to make comparable scores across both documents and topics).

Results

The best system had a normalized detection cost of 0.26. (There are no English-only results for this system to compare to the best 1998 results.)

The second best system had a cost of 0.32. In side experiments, it obtained an English-only cost of 0.23 and a Chinese-only cost of 0.25. This is reassuringly similar to the monolingual segmentation and tracking results in showing that language makes little difference for monolingual tasks.

First Story Detection

Task

The first story detection task required systems to find the first (and only the first) story about a topic. Systems were given a stream of stories in English and could wait until the end of 10 files before having to output decisions.

Performance was calculated for 180 labeled topics, the 60 fully annotated topics plus 120 partially annotated to include the first story. (As in the detection task, most outputs were necessarily ignored in the scoring process.)

Approaches

The most successful system used an incremental vector space model and compared new stories to previously formed clusters.

Another system (a close second) compared each new story to all previous stories (rather than clusters).

Results

The best system obtained a normalized first story detection cost of 0.70. The corresponding DET curve is shown in Figure 6.

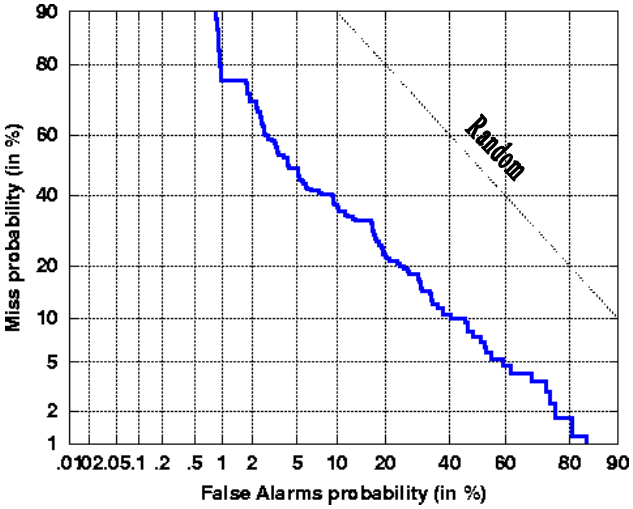

Figure 6 – First Story Detection Results for Best System

Although not very impressive, this performance may be adequate for some applications: According to the DET curve, first story detection software would let a person reduce his reading by 90% and still find half of the new topics as soon as they appear.

In side experiments, Jin showed (Allan et al., 1999) that first story detection performance can be predicted from tracking performance (where the high false alarm rate end of one curve influences the high miss rate end of the other). This suggests that, to get good first story detection results, tracking needs to be improved substantially or other methods need to be found.

Link Detection

Task

The story link detection task required systems to decide whether two stories discuss the same topic. Systems were presented with 21,600 pairs of English stories and asked to say Yes or No.

The test material consisted of 180 stories x 120 comparison stories for each, with roughly half of the comparisons addressing the same topic, the other half not.

Approaches

The most successful system used a cosine similarity measure to compare the two stories under test with stop words, stemming, binary term vectors, incrementally updated tf*idf values, and a time-based penalty (to lower the score the more time there is between the stories).

Other systems used variants on the same basic idea.

Results

The best system obtained a normalized link detection cost of 0.095. The comparable human performance is 0.055.

Lessons Learned

TDT 1999 produced significant progress in all dimensions, including especially good results for tracking.

The most pleasant lesson was that TDT techniques can function well in languages very different from English. Monolingual tracking is as good in Chinese as in English, and translingual tracking works moderately well. Translingual detection needs more work.

Translingual performance is impacted by translation errors. Although Systran MT outputs produced the best results in TDT 1999, less expensive techniques (which could be ported to other languages easily) worked almost as well.

Combining scores from different algorithms proved to be a big win.

Various combinations of lexical, prosodic, and structural features demonstrated value in specific tasks.

Names (of people, organizations, and locations) helped.

Boundaries from automatic segmentation seem almost as useful as true boundaries.

There was considerable variability in detection performance across topics, but it was system specific.

Normalization – across sources, languages, and topics – is essential.

Through demonstrations and discussions at the workshop, we are beginning to see how TDT technology can be integrated into TIDES and other applications.

TDT 2000

TDT 2000 will be much like TDT 1999. The languages will be English and Chinese again. The corpus will be TDT3 with 60 new topics to provide fresh challenges. The tasks will be variants of the five TDT 1999 tasks modified to raise the bar a bit further: Systems will not be given true boundaries, but may use automatic boundaries to be provided with the corpus. For the tracking task, systems will be given only single stories for training.

A dry run will be conducted during the summer, the official TDT 2000 evaluation will be conducted in October, and the results will be discussed at the TDT Workshop to be held in November after TREC.

New participants are welcome, and may choose any or all of the tasks. Interested parties should contact NIST[2] as soon as possible.

DARPA will also begin experimenting with ways to use TDT technology inside the TIDES Portal, an experimental interface to multiple data streams.

TDT 2001

TDT 2001 will be a further evolution. How the tasks will change will depend upon the results obtained in 2000. The number of languages is expected to grow, and there will be a broader range of sources for languages other than English. A new corpus, called TDT4, will be created to serve these needs using data to be recorded in 2000.

[2] Jonathan.Fiscus@nist.gov

Conclusions

TDT is an important area of research, addressing central application needs. It presents new and interesting technical challenges.

The enormous progress demonstrated anew the virtue of formal research task definitions, common data, and common evaluations. Clearly defined technical tasks made it possible to move forward. Representative, accurately labeled corpora made it possible to conduct meaningful research and to evaluate performance. Common, objective evaluations showed researchers which techniques worked best and allowed them to make meaningful improvements.

Acknowledgements

DARPA provided critical funding and strong support for this work, including the all-important corpora. A crew of talented people at the LDC spent thousands of hours creating the corpora. George Doddington played a pivotal role in problem definition and evaluation specification. Jon Fiscus did a superb job of administering the evaluation and analyzing its results. James Allan was a benevolent and skilled Coordinator for the TDT Evaluation Community. A great many other people, including inventive researchers at diverse sites, contributed to the signal success of TDT 1999. It was a wonderfully collaborative team effort.

Jon Fiscus provided the DET curves and cost figures used in this paper. Chris Cieri, George Doddington, James Allan, and Jon Fiscus made helpful suggestions regarding the paper itself. The author thanks them all.

References

Allan, J.; Caputo, D.; Gildea, D.; Hoberman, R.; Jin, H.; Lavrenko, V.; Rajman, M.; & Wayne, C. (1999). Topic Based Novelty Detection. In Final Report of the Johns Hopkins University Summer Workshop, 1999. Available at http://www.clsp.jhu.edu/ws99/projects/tdt.

Allan, J. et al. (2000). Proceedings of the TDT 1999 Workshop, February 2000. Available at http://www.nist.gov/TDT.

Cieri, C.; Graff, D.; Liberman, M.; Martey, N.; & Strassel, S. (2000). Large Multilingual Broadcast News Corpora for Cooperative Research in Topic Detection and Tracking: The TDT2 and TDT3 Corpus Efforts. In Proceedings of Language Resources and Evaluation Conference, Athens, Greece, May 2000.

Doddington, G. (1999). The 1999 Topic Detection and Tracking (TDT) Task Definition and Evaluation Plan. Available at http://www.nist.gov/TDT.

Martin, A.; Doddington, G.; Kamm, T.; Ordowski, M.; & Przybocki, M. (1997). The DET Curve in Assessment of Detection Task Performance. In Proceedings of Eurospeech, September 1997. Available at http://www.nist.gov/TDT.

Strassel, S. & Graff, D. (2000). Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora. In Proceedings of Language Resources and Evaluation Conference, Athens, Greece, May 2000.

Wayne, C. (1998). Topic Detection & Tracking: A Case Study in Corpus Creation & Evaluation Methodologies. In Proceedings of Language Resources and Evaluation Conference, Granada, Spain, May 1998.