Topic Detection and Tracking in English and Chinese

Charles L. Wayne

Department of Defense
Ft. Meade, Maryland, USA
Email: clwayne@nist.gov

Abstract

Topic Detection and Tracking (TDT) refers to automatic techniques for discovering, threading, and retrieving topically related material in streams of data. Newswire and broadcast news are the canonical sources. In 1999, TDT research was extended from English to Chinese, and carefully annotated multilingual corpora were created. Researchers devised clever approaches to the cross-language challenge, and formal performance evaluations yielded very promising results. This paper outlines the 1999 research tasks, corpora, evaluation procedures, technical approaches, and results. The multilingual, multimedia research and evaluations are continuing in 2000 and 2001 under the DARPA TIDES program.

Keywords: Speech; Text; Topic; Segmentation; Tracking; Detection.

1 Introduction

The Defense Advanced Research Projects Agency (DARPA) has sponsored research on Topic Detection and Tracking (TDT) since 1997 and has increased the technical challenges each year. In the 1997 pilot study, systems attacked newswire and manually transcribed broadcast news. In 1998, systems were exposed to the errorful outputs of automatic speech recognition (ASR) technology. In 1999, systems started dealing with cross-language issues. This paper reports on the relevant corpora, research tasks, technical approaches, and results. It consolidates in one place work done at multiple sites and focuses especially on the issues involved in dealing with both English and Chinese language data.

2 Topic Detection and Tracking

TDT encompasses 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 the prototypical situation: Horizontal lines represent incoming streams of news stories from different sources, media, and languages. Each rectangle represents a single story, and each is about the same *event*.

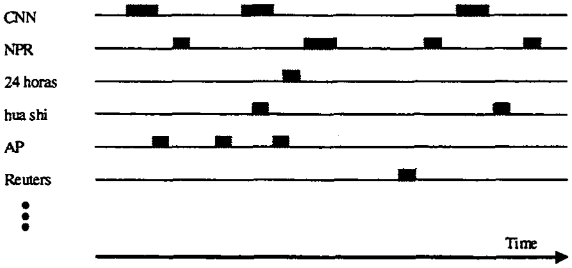


Figure 1 – Stories on One Topic (Event) in Several Media

TDT aims to discover this kind of structure automatically, giving users timely and efficient access to large quantities of information and helping them to keep on top of developments around the world: 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. Alternatively, a user could go into a large archive, find all the stories about a particular event, and see how it evolved.

2.1 Distinguishing Features

Three things make TDT unique.

2.1.1 Definition 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.) Other topic-oriented research deals with *categories* of information (e.g., bombings in general or bombings in some geographical region).

Proceedings of the 5th International Workshop Information Retrieval with Asian Languages

Copyright ACM 1-58113-300-6/00/009 ... $5.00

2.1.2 Discovery of New Events

TDT emphasizes the discovery of *new* events (topics), which no one expected or knew to request. This capability could be quite useful in many applications.

2.1.3 Focus on Linguistic Content

TDT focuses on *content* in order to create broadly applicable language-based technology. Algorithms are allowed to use information that is available in all applications (source, date, and time); but not the titles, keyword lists, et cetera that accompany newswire nor the closed captions that accompany some audio data. (These items could be exploited in real applications, if available.)

2.2 Research Tasks

The TDT 1999 research was factored into five complementary 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*)

This paper concentrates on the first three, as they are the major tasks and have been done in both Chinese and English.

2.3 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.

2.4 Supporting Technology

In addition to using the contents of the original data streams (text or audio in English or Chinese), TDT algorithms are permitted to utilize the outputs of automatic speech recognition (ASR) and machine translation (MT) included in the corpora.

3 Corpora

In order to support TDT research and evaluation, the Linguistic Data Consortium (LDC) produced two multimedia, multilingual corpora (known as TDT2 and TDT3) and *completely* annotated them for specific, randomly chosen topics. The annotations facilitate frequent experimentation and make the corpora extremely valuable for research.

3.1 Sources

Both corpora contain representative data in both Chinese and English from both text and audio news sources. The text comes from a variety of newswire and web sources; the audio, from a mixture of radio and television sources. Table 1 shows the diversity and scale of the two corpora.

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
1872	Total Chinese	12395

Table 1. Sources and Stories for TDT2 and TDT3

TDT2 data are from January to June 1998; TDT3, from October to December 1998. The LDC sampled the various news streams (if available) several times every day in broadcast-size chunks — typically 30 minutes from audio sources or a corresponding number of stories from text sources.

3.2 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
- Automatic transcript (with word/character timing)
- Manual transcript (generally of closed caption quality)

For Chinese sources —
- Machine translation into English

For all sources —
- Origin (medium, source, date, time)
- Story boundaries (with time stamps)
- News / non-news tags
- Topic tags (YES, NO, BRIEF for each story and each topic)

3.3 Topic Annotation

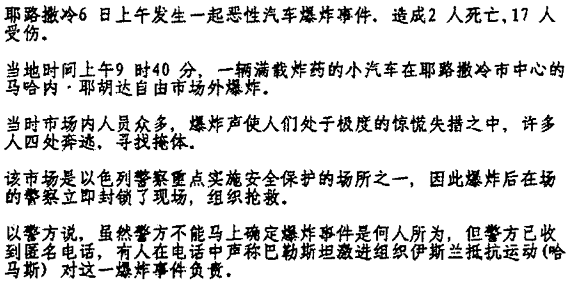
The LDC selected and annotated 100 topics in TDT2, 60 in TDT3. They chose topics (events) at random from the various sources, then read *every* story and labeled it YES, NO, or BRIEF with respect to *each* topic (where BRIEF signifies that a topic occupied less than 10% of a story). For TDT2, the LDC chose the topics from English sources only, then searched for them in both languages. For TDT3, they chose the topics from both English and Chinese sources and made sure that each topic appeared at least four times in each language. Annotators worked in their native languages from text data or manual transcripts of audio data.

3.4 Automatic Transcription

All of the audio files were automatically transcribed by Dragon Systems, using their software, or by NIST, using BBN software. Depending upon the data, the word error rates varied widely, averaging 25-30%. The ASR outputs were created both to save TDT researchers from having to create their own ASR technology and to provide a common basis for comparing TDT algorithms.

3.5 Automatic Translation

Using Systran MT software, the LDC automatically translated all of the Chinese text and Chinese ASR outputs into some approximation of English. To illustrate the accuracy of this process, Figure 2 shows part of a typical story; Figure 3, its translation.



耶路撒冷6 日上午发生一起恶性汽车爆炸事件. 造成2 人死亡,17 人受伤.

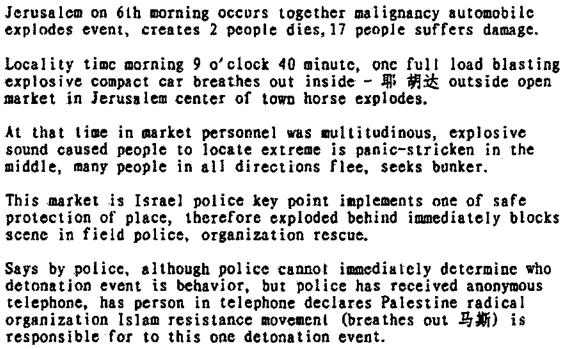
当地时间上午9 时40 分, 一辆满载炸药的小汽车在耶路撒冷市中心的马哈内·耶胡达自由市场外爆炸.

当时市场内人员众多, 爆炸声使人们处于极度的惊慌失措之中, 许多人四处奔逃. 寻找掩体.

该市场是以色列警察重点实施安全保护的场所之一, 因此爆炸后在场的警察立即封锁了现场, 组织抢救.

以警方说, 虽然警方不能马上确定爆炸事件是何人所为, 但警方已收到匿名电话, 有人在电话中声称巴勒斯坦激进组织伊斯兰抵抗运动(哈马斯) 对这一爆炸事件负责.

Figure 2 – Xinhua Story on Bombing in Jerusalem



Jerusalem on 6th morning occurs together malignancy automobile explodes event, creates 2 people dies, 17 people suffers damage.

Locality time morning 9 o'clock 40 minute, one full load blasting explosive compact car breathes out inside - 耶 胡达 outside open market in Jerusalem center of town horse explodes.

At that time in market personnel was multitudinous, explosive sound caused people to locate extreme is panic-stricken in the middle, many people in all directions flee, seeks bunker.

This market is Israel police key point implements one of safe protection of place, therefore exploded behind immediately blocks scene in field police, organization rescue.

Says by police, although police cannot immediately determine who detonation event is behavior, but police has received anonymous telephone, has person in telephone declares Palestine radical organization Islam resistance movement (breathes out 马斯) is responsible for to this one detonation event.

Figure 3 – Systran Translation of Xinhua Story

3.6 Limitations

Despite the many virtues of these corpora, Table 1 shows some limitations: There are a reasonable number and variety of English sources, but only three Chinese sources. The English data contains more audio than text; the Chinese, more text than audio. The amounts of data from particular sources (providers) vary widely. Consequently, one must be somewhat cautious in interpreting the test results.

A new TDT4 corpus with more languages and more diverse sources will be collected this fall. More topics are being annotated within the TDT3 corpus.

3.7 Additional Information

For more information about the contents or the creation of the TDT corpora, please see the LDC's TDT web pages (www.ldc.upenn.edu/Projects/TDT).

4 Evaluation

4.1 Process

In order to focus TDT research, calibrate progress, and provide diagnostic feedback, the National Institute of Standards and Technology (NIST) worked with the sponsor and the research community to define a set of objective performance measures that represented the essence of TDT. NIST then created supporting software (which they shared with the community to aid pre-evaluation research), administered a formal evaluation in December 1999 (wherein sites processed data locally and submitted outputs to NIST), and hosted a technical workshop in February 2000.

4.2 Test Corpus

TDT3 was used for the formal evaluation. TDT2 was available for pre-evaluation research and for learning parameters that would be useful in the test. Sites could also use any other linguistic resources that predated October 1998.

4.3 Formulation as Statistical Detection

To represent the research challenges crisply, NIST formulated each task (not just the detection task) as a classical statistical detection problem. At every decision point (story or potential boundary), a system had to output both a decision and a confidence score.

4.4 Calculations of Performance

4.4.1 DET Curves

From the scores, NIST software produces a *detection error tradeoff* (DET) curve of the sort illustrated in Figure 4 and Figure 5. DET curves, nicely described in [1], show miss and false alarm rates at multiple operating points, making it easy to compare results obtained with different algorithms or under different conditions. (Please note that desirable performance is in the lower left corner of the plot.)

Compared to traditional precision-recall (PR) curves, DET curves have the advantage of avoiding the confounding effect of target richness in different corpora. This occurs because DET curves are based on probability density functions (of scores given target and non-target conditions) and do not depend on the proportion of target and non-target material in a corpus.

4.4.2 Normalized Costs

From the decisions, the software calculates a *normalized cost* that reflects both the overall strength of an algorithm and its ability to set thresholds correctly. While miss and false alarm characterize the accuracy of an algorithm in the abstract, cost reflects its utility in a (hypothetical) application and reduces everything to a single number.

The basic cost function is given in Equation (1), where the target conditions are the presence of a story boundary or of an on-topic story, and their probabilities are a priori.

$$\begin{aligned} C &= C_{Miss} P_{Miss|Target} P_{Target} \\ &+ C_{False\,Alarm} P_{False\,Alarm|NonTarget} (1 - P_{Target}) \end{aligned} \qquad (1)$$

The parameters used in TDT 1999 are given in Table 2.

	C_{Miss}	$C_{FalseAlarm}$	P_{Target}
Segmentation	1.0	0.3	0.30
Tracking	1.0	0.1	0.02
Detection	1.0	0.1	0.02

Table 2 – Parameters for Cost Formula

The resulting cost is normalized as shown in Equation (2).

$$C_{Normalized} = \frac{C}{\min(C_{Miss} P_{Target}, C_{FalseAlarm}(1 - P_{Target}))} \qquad (2)$$

A normalized cost of 0.00 represents perfect performance; 1.00 indicates that no information has been extracted from the source data.

The research challenge is to produce an algorithm that minimizes normalized cost.

4.4.3 Topic-Weighted Results

Because topic difficulty is a major source of variability and the number of on-topic stories varies widely, the software computes the above as *topic-weighted* results (with each topic contributing equally to the overall averages). This improves the reliability of the performance measures.

4.5 Additional Information

For more information about the evaluation process, please see [2], which spells out the research challenges and evaluation procedures clearly and in considerable detail.

5 Research

5.1 Participants

In 1999, eleven academic and industrial research groups (including several volunteers) participated in TDT research and submitted results for one or more of the tasks:

Carnegie Mellon University
National Taiwan University
University of Iowa
University of Maryland
University of Massachusetts
University of Pennsylvania

BBN
Dragon Systems
General Electric
IBM
MITRE

5.2 Approaches

The sites demonstrated considerable creativity, diversity, and sophistication in their approaches, advancing the state of the art in TDT technology in general and applying it (for the first time) to the cross-language challenge.

Due to space limitations, only brief descriptions of the approaches appear below. Much more information can be found in the many papers [3] and presentations [4] given at the TDT Workshop held in February 2000. Still more is appearing in papers that participants are writing for various conferences and journals. As these come out, NIST will post citations at www.nist.gov/TDT/research_links.

6 Key Tasks, Approaches, and Results

6.1 Segmentation

6.1.1 Task

The segmentation task required systems to find story boundaries in audio sources. Systems were given broadcast-sized files of data and had to produce outputs by the end of each broadcast.

6.1.2 Approaches

The most successful system (described in [5]) 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.

6.1.3 Results

Table 3 shows the results obtained by the best system.

English Test	Chinese Test
0.39	0.32

Table 3 – Normalized Segmentation Costs

Techniques developed primarily on English worked well (in fact, better) on Chinese. One cannot draw strong conclusions about language differences, however, since the only Chinese audio source, VOA Mandarin, was of very good quality.

To help interpret these numbers, a system that is always off by 3.5 seconds (for English) or 3.4 seconds (for Chinese) would produce the same normalized segmentation costs as shown in Table 2.

6.2 Tracking

6.2.1 Task

Tracking is essentially query-by-example in IR parlance. The official tracking task required systems to find all subsequent stories about the topic discussed in four example (training) stories. The four training stories were all in English; the test stories, in both English and Chinese. Systems were given true boundary information and had to output decisions as they processed each test story.

6.2.2 Approaches

The most successful system (described in [6]) used logistic regression to combine probabilities obtained from a 2-state topic spotting technique (with a language model built from concatenated training stories) and from a complementary 2-state probabilisitic information retrieval technique (in which the unknown story is assumed to produce the model). This was followed by normalization (with thousands of known off-topic stories) and adaptation (with high scoring test documents added to the training set and the 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.

6.2.3 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.

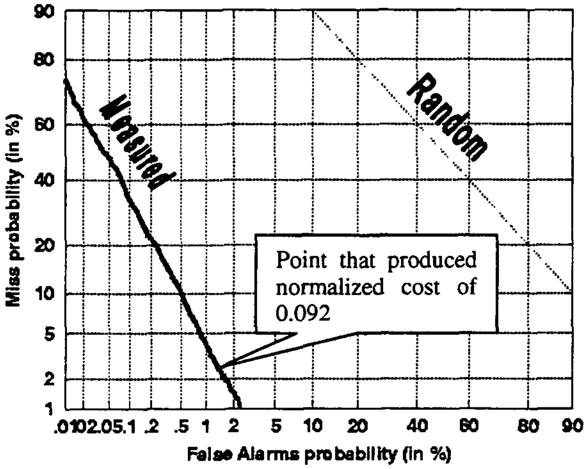


Figure 4 – Tracking Results

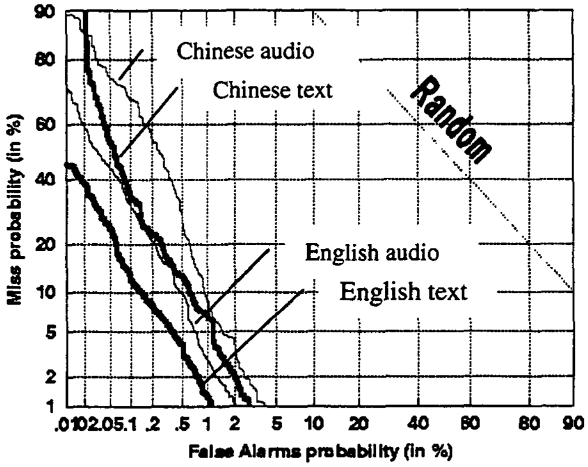


Figure 5 – Tracking Results by Language and Medium

Within each language, the results on text test data (thick lines) are better than the results on audio test data (thin lines). These differences could be due both to content differences (since text stories are longer on average) and/or to ASR output errors. Table 4 shows the associated cost figures.

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

Table 4 – Normalized Tracking Costs Within and Across Languages

Once again, we see that techniques work comparably well in monolingual tasks (training and testing in the same language). However, we also see strong cross-language differences.

6.3 Detection

6.3.1 Task

Topic detection was the most challenging task. It required systems to group incoming stories into topic clusters automatically (i.e., without supervision) and to create new clusters (topics) as needed. Like tracking, the detection task was cross-medium, cross-source, and cross-language. Systems were given a stream of stories in English and Chinese and could wait until the end of 10 files (broadcasts or comparable amounts of text) before announcing decisions. Once made, decisions were irreversible.

6.3.2 Approaches

The most successful system (described in [5]) 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 scores comparable across both documents and topics).

6.3.3 Results

The best system had a normalized detection cost of 0.26.

To provide some insight into the difficulty of doing detection across languages, Table 5 shows the corresponding cost of 0.32 for the second best system – along with the results of unofficial, monolingual side experiments run on that system.

Multilingual	English Only	Chinese Only
0.32	0.23	0.25

Table 5 – Normalized Detection Costs Within and Across Languages

As in monolingual segmentation or tracking, monolingual detection results are reassuringly similar. The official (multilingual) results are noticeably less good, as one might expect.

7 Cross-Language Issues

Since TDT techniques work comparably well on both Chinese and English separately, why do they not work better across languages? A large part of the reason must be translation accuracy, especially for names. Another factor may be cross-language score normalization, which could be improved with more training data, inter alia. And there is certainly room for new ideas, such as symmetrical processing across languages.

7.1 Name Translation

Names provide a great deal of information for linking stories about a particular event. Figure 6 lists all the names that appeared in the Xinhua story (including the portion that did not fit in Figure 2), showing how Systran rendered those names and how an English newswire would have rendered them. It is significant that Systran got only 5 of the 14 names correct. Many of the errors illustrated can be attributed to Systran's incorrect segmentation of the Hanzi character streams.

Xinhua Source	Systran Output	Correct English Translation
耶路撒冷	Jerusalem	Jerusalem
马哈内·耶胡达	breathes out inside - 耶胡达...horse	Machane Yehuda
以色列	Israel	Israel
以	by	Israel
巴勒斯坦	Palestine	Palestine
伊斯兰	Islam	Islam
哈马斯	breathes out 马斯	Hamas
卡哈拉尼	card breathes out Raney	Kahalani
美国	USA	USA
以巴	by Pakistan	Israeli-Palestinian
内塔尼亚胡	In...tower Nepal Asia Hu	Netanyahu
阿拉法特	Alfate	Arafat
约旦河西岸	Jordan Hexi shore	West Bank of Jordan River
拉马	pulls a horse	Ramallah

Figure 6 – Name Translation for Xinhua Story

7.2 Score Normalization

To achieve good results in monolingual TDT tasks, it is necessary to normalize scores across media and sources. Similarly, in cross-language tracking and detection, it is necessary to normalize scores across languages.

Normalization success varied considerably across sites. The best tracking results were obtained by a system [6] that normalized its scores using means and variances estimated separately for English and Chinese on known off-topic stories. The same site achieved the second best normalized detection cost plus the most consistent detection costs within and across languages and media.

7.3 Alternative Translations

All sites ended up using the Systran translations, either because they provided the best results or because doing so was the path of least resistance.

One site devised a relatively simple statistical translation system (described in [6]) and obtained tracking results almost as good as with the Systran output. This suggests that TDT technology could be ported relatively easily to languages for which there is no good translation technology. On the other hand, better quality translations may be needed as TDT algorithms become more sophisticated.

8 Conclusions

TDT represents an important avenue of research with broad application potential and interesting technical challenges. TDT algorithms provide unique capabilities that could be used jointly or separately in various applications and that could be combined with ASR, information retrieval, summarization, and MT technology to satisfy a wide variety of user needs.

TDT performance was quite good on the tracking task, respectable on the segmentation and detection tasks. Systems made creative use of a variety of acoustic, lexical, prosodic, and structural features. They gained robustness by combining scores from different matching algorithms and by normalizing scores to account for source, medium, and language differences.

With minor modifications, algorithms developed for English worked comparably well on monolingual tasks in Chinese. This suggests that TDT technology could be ported to arbitrary languages.

Results were distinctly less good on cross-language tasks (tracking topics in Chinese given only English examples, or detecting that stories in English and Chinese were on the same topic). To get creditable results, researchers found that cross-language normalization was essential (just as cross-medium and/or cross-source normalization was critical within a single language). Even with normalization, there is considerable room for improvement. Accurate name translation would certainly be a fruitful avenue to explore.

Although a great deal of good research can and should be done within a language, there is no substitute for experiments involving multiple languages. This is especially true for TDT, where key applications are inherently cross-language. It was very fortunate that DARPA insisted on this direction and that the LDC was able to create suitable corpora in English and Chinese.

9 Future Directions

TDT research will continue under the DARPA TIDES program, with additional data, more languages, and evolving technical challenges. Additional sites are welcome to participate in the annual evaluations as volunteers. Interested parties are encouraged to contact NIST (Jonathan.Fiscus@nist.gov).

10 Acknowledgements

DARPA provided the key funding, vision, and encouragement that made this work possible. The LDC created the incredibly valuable corpora. NIST provided the essential evaluation infrastructure. Many researchers at many sites helped both to refine the TDT research challenges and to devise creative approaches for attacking them. It was a wonderfully collegial and cooperative venture.

In the preparation of this paper, special thanks are due to Jon Fiscus, who analyzed the evaluation results and provided the cost figures and DET curves used herein, and to Shudong Huang, who provided the MT examples. James Allan, Chris Cieri, George Doddington, and Jon Fiscus all provided helpful comments on portions of the text.

11 Additional Reading

In reading about TDT, one encounters some terminological variations: Officially, the terms TDT1, TDT2, and TDT3 refer to corpora; TDT 1997, TDT 1998, TDT 1999, TDT 2000, and TDT 2001 refer to research periods and associated evaluations. Many authors use TDT1 to refer to TDT 1997, TDT2 to refer to TDT 1998, and TDT3 to refer to TDT 1999 – as those were the corpora used as test material in those years. In addition, groups that did not participate in the official evaluations and workshops sometime use terms like topic, detection, and tracking in non-standard ways.

11.1 First Places to Look

For TDT 1999, the basic procedures are spelled out in [2]. There are 13 papers in [3] and 24 presentations in [4].

For TDT 1998, there are 16 TDT papers in [7].

For TDT 1997, there are 4 TDT papers in [8].

11.2 Other Places to Look

Additional work related to TDT appears in references [9] onward. This is a reasonably representative list, with items grouped by organization. A more complete (and growing) list is at www.nist.gov/TDT/research_links.

12 References

[1] Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech 1998*. 1998. Available at http://www.itl.nist.gov/iaui/894.01/publications.

[2] Doddington, G. The 1999 topic detection and tracking (TDT) task definition and evaluation plan. 1999. http://www.nist.gov/TDT/tdt99/doc/tdt3.eval.plan.99.v2.7.ps.

[3] *Proceedings of the TDT 1999 Workshop*. 2000. http://www.nist.gov/TDT/tdt99/papers.

[4] *Presentations at the TDT 1999 Workshop*. 2000. http://www.nist.gov/TDT/tdt99/presentations.

[5] Franz, M., McCarley, J.S., Roukos, S., Ward, T. and Zhu, W.-J. Segmentation and detection at IBM: Hybrid statistical models and two-tiered clustering. In *Proceedings of the TDT 1999 Workshop*. 2000. http://www.nist.gov/TDT/tdt99/papers.

[6] Leek, T., Jin. H., Sista, S. and Schwartz, R. The BBN crosslingual topic detection and tracking system. In *Proceedings of the TDT 1999 Workshop.* 2000. http://www.nist.gov/TDT/tdt99/papers.

[7] *Proceedings of DARPA Broadcast News Workshop.* 1999. http://www.nist.gov/speech/publications/darpa99/index.htm

[8] *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop.* 1998. http://www.nist.gov/speech/publications/darpa98/index.htm.

[9] Cieri, C. Multiple annotation of reuseable data resources: Corpora for topic detection and tracking. In *Actes des 5es Journees internationales d'analyse statistique des donnees textuelles*, Rajman, M. and Chappelier, J., eds. 2000, volume 1.

[10] Cieri, C., Graff, D., Liberman, M., Martey, N. and Strassel, S. Large multilingual broadcast news corpora for cooperative research in topic detection and tracking: The TDT2 and TDT3 corpus efforts. In *Proceedings of the Second International Language Resources and Evaluation Conference.* 2000.

[11] Strassel, S., Graff, D., Martey, N. and Cieri, C. Quality control in large annotation projects involving multiple judges: The case of the TDT corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference.* 2000.

[12] Beeferman, D., Berger, A. and Lafferty, J. Statistical models for text segmentation. *Machine Learning, Special Issue on Natural Language Learning*, Cardie, C. and Mooney, R., eds., 1999, 34(1-3), pp. 177-210.

[13] Yang, Y., Carbonell, J., Brown, R. and Pierce, T. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems, Special Issue on Applications of Intelligent Information Retrieval.* 1999, 14(4), pp. 32-43.

[14] Yang, Y., Ault, T., Pierce, T. and Lattimer, C. Improving text categorization methods for event tracking. In *Proceedings of SIGIR 2000.* 2000, pp 65-72.

[15] Yang, Y., Pierce, T., and Carbonell, J. A study on retrospective and on-line event detection. In *Proceedings of SIGIR 1998.* 1998, pp. 28-36.

[16] Yang, Y., Ault, T. and Pierce, T. Combining multiple learning strategies for effective cross validation. In *Proceedings of International Conference on Machine Learning (ICML'00).* 2000, pp 1167-1182.

[17] van Mulbregt, P., Carp, I., Gillick, L., Lowe, S., Yamron, J. Text segmentation and event tracking on broadcast news via a hidden markov model approach. In *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio.* 1999, pp. 90-95.

[18] Yamron, J., Carp, I., Gillick L., Lowe, S. and van Mulbregt, P. A hidden markov model approach to text segmentation and event tracking. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing.* 1998, 1: pp. 333-336.

[19] McCarley, J.S. and Franz, M. Influence of speech recognition errors on topic detection. In *Proceedings of SIGIR 2000.* 2000, p. 342.

[20] Dharanipragada, S., Franz, M., McCarley, J.S., Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. Statistical models for topic segmentation. In *Proceedings of ICSLP 2000.* 2000.

[21] Dharanipragada, S., Franz, M., McCarley, J.S., Papineni, K., Roukos, S. and Ward, T. Story segmentation and topic detection for recognized speech. In *Proceedings of Eurospeech 1998.* 1998.

[22] Shriberg, E., Stolcke, A., Hakkani-Tur, D. and Tur, G. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 2000, 32(1-2).

[23] Tur, G., Hakkani-Tur, D., Stolcke, A. and Shriberg, E. Integrating prosodic and lexical cues for automatic topic segmentation. To appear in *Computational Linguistics.*

[24] Papka, R. and Allan, J. Topic detection and tracking: Event clustering as a basis for first story detection. In *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, Croft, B., ed. Kluwer Academic Publishers, 2000, pp 97-126.

[25] Fukumoto, F. and Suzuki, Y. Event tracking based on domain dependency. In *Proceedings of SIGIR 2000.* 2000, pp 57-64.

[26] Clifton, C. and Cooley, R. TopCat: Data mining for topic identification in a text corpus. In 3^{rd} *European Conference on Principles and Practices of Knowledge Discovery in Databases.* 1999.

[27] Carthy, J. and Smeaton, A. The design of a topic tracking system. In *Proceedings of BCS-IRSG 2000.* 2000, pp. 84-93.

[28] Hatch, P., Stokes, N., and Carthy, J. Topic detection, a new application for lexical chaining? In *Proceedings of BCS-IRSG 2000.* 2000, pp. 94-103.