6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

RESULTS OF THE 1999 TOPIC DETECTION AND TRACKING EVALUATION IN MANDARIN AND ENGLISH

Jonathan G. Fiscus and George R. Doddington

National Institute of Standards and Technology

ABSTRACT

The National Institute of Standards and Technology (NIST) administered the second open evaluation of Topic Detection and Tracking (TDT) technologies in 1999. The TDT project supports development of technologies that automatically organize event-related news stories. The program leverages expertise in core technologies, Automatic Speech Recognition (ASR), Document Retrieval (DR), and Machine Translation (MT) to build the TDT technologies.

The 1999 TDT project extended the 1998 TDT project in two dimensions, first by adding Mandarin Chinese audio and text sources and second by adding two new evaluation tasks. Through experimental controls and conditioned analysis of system performance, the 1999 evaluation yielded numerous insights into the effects of multilingual texts on TDT technologies. Three notable generalizations arise from the evaluation: (1) English and Mandarin story segmentation performance is similar, (2) cross-lingual topic tracking performance is 44% worse than monolingual tracking, and (3) multilingual topic detection performance is 37% worse than monolingual topic detection.

1. INTRODUCTION TO TDT

The TDT project is a DARPA-sponsored evaluation-driven research program to advance the state of the art in technologies that automatically organize event-related stories from continuously expanding information streams. [8] Users of such information are confronted with an overwhelming amount of information, and technology is needed to reduce the amount of human labor needed to digest it.

The 1999 TDT project was the second open evaluation. The program began in 1997 with a small pilot study, and the first open evaluation occurred in 1998 [3]. The third open evaluation will occur in 2000 as an integral part of the DARPA Translingual Information Detection and Summarization (TIDES) program [7]. Details of the latest evaluation are available on the NIST TDT Website http://www.nist.gov/TDT [2].

The 1999 TDT project included all of the tasks from the previous year's evaluation: Topic Tracking, Topic Detection, and Story Segmentation and expanded the project along three dimensions.

First, the project was made multilingual by expanding the corpora to include Mandarin Chinese newswire, web-based news, and broadcast news information streams. The multilingual aspect to the evaluation was seen as a natural extension of the technology. However, the introduction of Chinese, and the implied need for Machine Translation, added considerable complexity to the TDT systems. In order to reduce the amount of system re-engineering, the LDC provided the output of a commercially available Mandarin-to-English translation system, Systran, on the Mandarin transcripts [4]. Evaluation participants had the choice to work on the Mandarin text directly or to use the supplied translations.

The second expansion was the addition of two new evaluation tasks: Link Detection and First Story Detection (FSD). In brief, the link detection task evaluates a primitive function that answers the question: "Are these two documents 'linked' by a common topic?" The FSD task evaluates a system's ability to detect if a new story discusses a previously unseen topic.

The third expansion of the TDT was to include two new English broadcast news sources. While the content of the new sources is unlikely to be markedly different from previous sources, the story segmentation performance on unseen sources is of interest to the research community.

TDT evaluations involve many details that required extensive collaborative efforts to develop. The product of this collaboration was "The 1999 Topic Detection and Tracking Task Definition and Evaluation Plan" [1]. The evaluation plan serves the vital role of precisely communicating the expected system function and evaluation criteria.

1.1 TDT Topics

Research in the TDT project is predicated on finding and organizing broadcast news stories based on topics. The definition of "topic" is a fundamental issue. Rather than treat topics as categorization scheme, like most other previous research, the TDT community defined a topic to be "a seminal event or activity, along with all directly related events and activities" [1]. 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.

1.2 TDT Evaluation Criterion

TDT tasks are cast as detection tasks. Detection tasks view performance as a tradeoff between two error types: missed detections and false alarms. Such systems have many operating points, so TDT evaluates system performance both by error tradeoff and by a selected single operating point.

TDT system evaluation follows the precedent defined for the Text Independent Speaker Recognition community [5]. Task performance is measured in two ways: (1) the Detection Error Tradeoff (DET) Curve and (2) the Normalized Detection Cost Function

Decision error tradeoff curves are graphical depictions of the tradeoff between miss and false alarm probabilities. [1] As is evident in Figure 2, the axes are warped by the normal deviant function. Thus, the straightness of lines indicates normally distributed detection scores.

The normalized cost function distills performance into a single number. The evaluation plan defines the cost function's formula. Essentially, the normalized cost function is a linear combination of the costs associated with miss detection and false detection. This metric is scaled so that a cost of 0.0 is perfect and cost of 1.0 is the best score achievable by saying "NO" for all detections.

1.2 TDT Evaluation Corpus

The LDC's TDT3 [4] corpus was used for the 1999 TDT evaluation. The LDC paper discusses in detail the TDT corpora, its content, structure and the extensive annotation performed by the LDC.

The evaluation corpus in 1999 was approximately double the size of the 1998 [3] evaluation corpus. Table 1 summarizes the salient differences between the evaluation corpus used in each year.

	1998	1999	
Language	English	English	Mandarin
LDC Corpus	TDT2	TDT3	
Evaluation epoch	May-June 1998	Oct-Dec 1998	
Sources	2 Newswire 4 TV/Radio	2 Newswire 6 TV/Radio	1 Newswire 2 Radio/Web
Hours Audio	255	475	121
Number of Stories	19K	31K	12K
Full Topic Annotation	34	60	
Partial Topic Annotation	0	120	0
Link Annot.	0	14K pairs	0

Table 1 Comparison of 1998 and 1999 TDT Evaluation Corpora

The 1999 evaluation epoch was 50% longer in duration, however the additional two English sources and three Mandarin sources effectively double the number of stories. The LDC exhaustively annotated 60 new topics against the corpus, partially annotated an additional 120 English topics to support the first story detection evaluation, and annotated 14K story pairs to support the link detection evaluation.

2. 1999 TDT EVALUATION

Eleven research groups participated in the evaluation: GTE Internetworking's BBN Technologies (BBN), Carnegie Mellon University (CMU), Dragon Systems (Dragon), General Electric (GE), IBM's T. J. Watson Laboratory (IBM), MITRE Corporation (MITRE), National Taiwan University (NTU), University of Pennsylvania (UPenn), University of Iowa (UIowa), University of Maryland (UMd), and University of Massachusetts (UMass).

The evaluation project supported five evaluation tasks in 1999: Topic Tracking, Topic Detection, Story Segmentation, and two new tasks, First Story Detection and Link Detection. The following sections describe each of the tasks, the results of the primary evaluation conditions, and interesting post-evaluation subset conditioned analysis.

Each evaluation task has a defined "primary" evaluation condition that all task participants are required to run. It is for these primary conditions that most of the cross-site comparisons are made since all systems have been run under the same experimental conditions.

The primary evaluation condition is "global" average that convolves many factors (e.g., broadcast language, medium, machine translation or recognition, etc.). The research community is interested in assessing the effects of these factors so NIST conditions their analysis, (i.e., computes performance based on factors), by dividing system results into subsets for each factor. These post-evaluation subset conditioned analyses highlight the factors' effects on performance.

2.1. Topic Tracking Task

The topic tracking task evaluates technologies that associate incoming stories with topics "known" to the system. Topics are known to the system by means of a small set of on-topic stories from the initial part of the evaluation corpus. Since each topic is independently defined, these "training epochs" vary from topic to topic and contain many off-topic stories and potentially additional unlabelled on-topic stories. The tracking task requires systems to classify correctly all stories after the training epoch as to whether or not they are on- or off-topic.

The primary evaluation condition was specified as topic tracking with four English topic training stories, testing on multilingual newswire and automatic transcriptions of broadcast news texts, with human generated story boundaries.

Figure 1 is a bar chart that summarizes the tracking performance for each of the test participants.

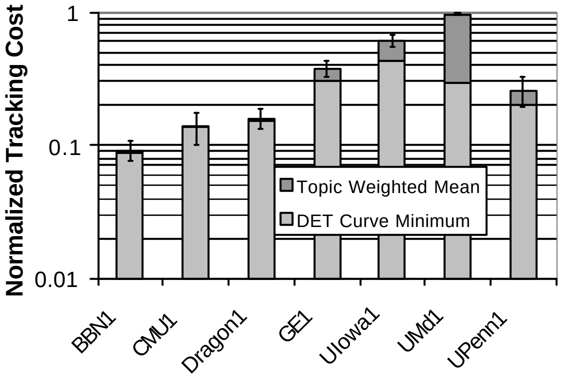

Figure 1 1999 Topic Tracking Primary Systems

The chart shows three statistics for each site.

First, the height of a bar is the topic-weighted normalized tracking cost associated the decisions made by the system, either "YES", a story is on-topic or "NO", it is not. The lowest cost of 0.092 was achieved by the BBN1 system and the next lowest cost of 0.14 by the CMU system.

The second statistic is the 95% confidence bar associated with the mean. The width of the confidence intervals show the per topic variability. Unexpectedly, the higher tracking costs do not imply larger topic performance variability. For instance, the GE system’s confidence bars are smaller that CMU's even though CMU's costs are lower.

The third statistic, the minimum DET point, is shown by the shaded part of a bar. The height of the sub-bar indicates the best score a system could get if the decision threshold was set to coincide with the optimal point of the DET curve. The DET curves in Figure 2 were used to derive these statistics. The difference between the two bars indicates how well the system's decision threshold was set. Most sites did well at choosing their thresholds. For the UMd system, this statistic shows their system performed markedly better than the actual decision costs suggest.

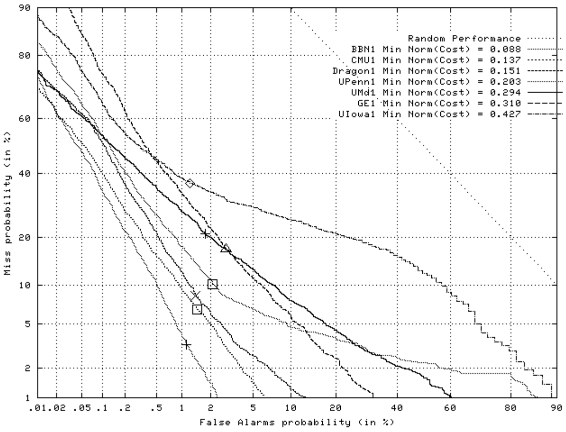

Figure 2 1999 Topic Tracking Primary System DET Curves

Post-evaluation subset conditioned analysis yielded Figure 3, system performance conditioned on language and data source type, newswire vs. broadcast news automatic transcripts.

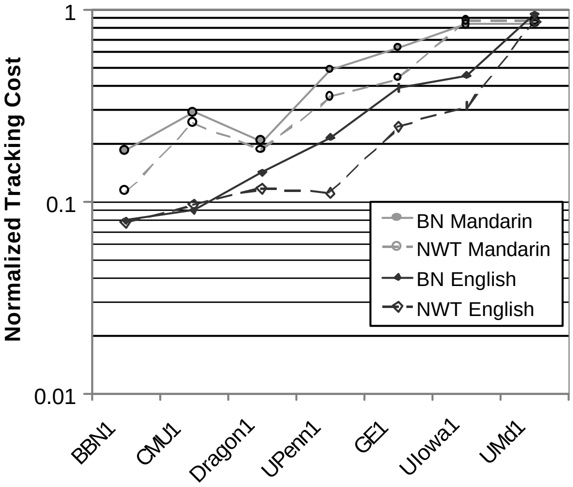

Figure 3 Primary systems conditioned analysis by broadcast type and language

There is a clear trend for topic tracking in Mandarin data to be harder than in English. However, primary systems were trained using only English training stories so the degradation is expected. Figure 4 compares the BBN system performance when it was trained using three different conditions: English stories, Mandarin stories and both English and Mandarin stories.

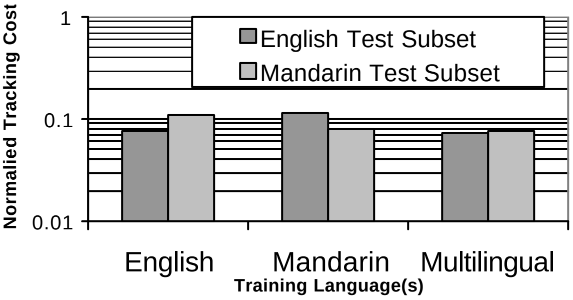

Figure 4 BBN Primary system trained on 4 English stories, 4 Mandarin stories, or 4 English + 4 Mandarin stories

The graph in Figure 4 shows that for the BBN system, the cross-language degradation as about the same, 44% relative, regardless of the training language. There is a slight improvement for both languages when both training from both languages is used for topic training.

2.2. Topic Detection Task

The topic detection task evaluates technologies that detect novel, previously unknown, topics. As in the tracking task, topics are defined by associating together stories that discuss the topic; however, topic detection systems are not given a priori knowledge of the topic. Therefore the system must embody an understanding of what constitutes a topic, and this understanding must be independent of topic specifics.

The primary evaluation condition was specified as topic detection on multilingual newswire and automatic transcriptions of broadcast news texts using a 10-source file decision deferral (an amount of look ahead time) period with human generated story boundaries. Figure 5 is a bar chart summarizing the performance for all the participants except NTU[1]. IBM achieved the lowest detection cost of 0.26.

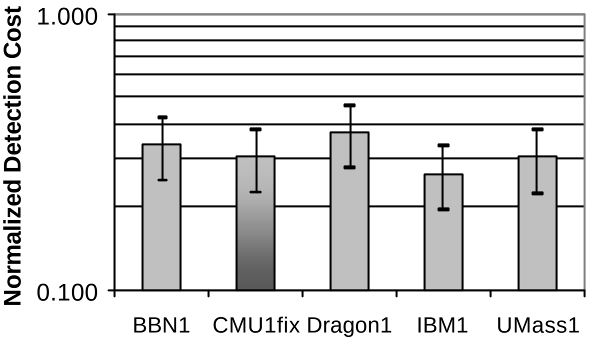

Figure 5 1999 Primary Topic Tracking Primary Systems

Like the tracking bar chart, the graph shows the topic weighted topic detection scores and the 95% confidence intervals associated with the means. Topic detection costs are typically high than the tracking costs. This is expected since topic detection can be thought of as an unsupervised training variation of the tracking task. Topic performance appears to be more widely varied as compared to tracking.

Results of a post-evaluation conditioning on source language and type are shown in Figure 6.

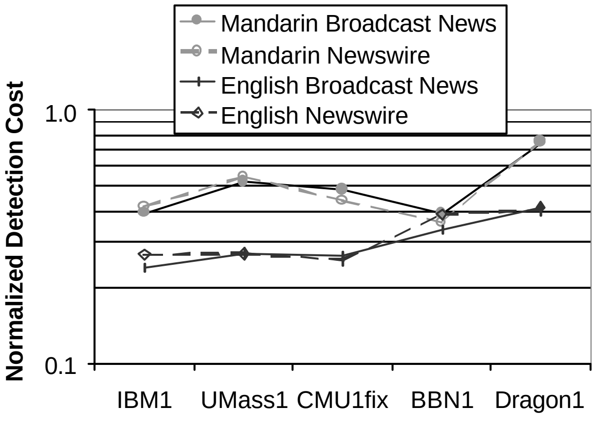

Figure 6 Primary detection system conditioned analysis by broadcast type and language

As with the tracking evaluation, there is a marked difference in performance between English and Mandarin test subsets. This is true for all participants except BBN. The BBN system was able to build multilingual topic clusters where other systems had more difficulty or built essentially monolingual clusters.

It was relatively easy to compare the 1998 test set to the 1999 test set since the monolingual BBN system was essentially unchanged from the previous year. The BBN monolingual detection costs were virtually identical in comparing the two year's English test sets, so we can conclude that the topic difficulty, for topic detection, was roughly identical between years. Since test sets between years are roughly identical, the UMass monolingual system improvement of 43% is mostly due to system improvement rather than test set variability.

2.3. Story Segmentation Task

The story segmentation task evaluates technologies that segments, (or divides), an incoming stream of text into TDT-style stories. In TDT, a story is a "topically cohesive segment of news that includes two or more DECLARATIVE independent clauses about a single event." [1] The notion of story explicitly excludes commercials from being stories, and therefore systems are not evaluated on inter-commercial story boundaries.

A new feature to this year's evaluation was story segmentation of Mandarin data. Since both Mandarin and English story segmentation are enabling technologies for deployable TDT systems, segmentation systems must do both languages using their native orthographies. The primary evaluation condition was to segment the English and Mandarin broadcast news data[2] using the full source file for decision deferral.

The left graph in Figure 7 summarizes the 1999 story segmentation performance. IBM had the lowest English segmentation cost of 0.39 and the lowest Mandarin segmentation cost of 0.32.

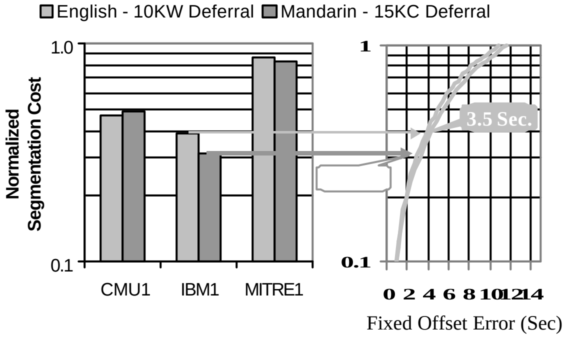

Figure 7 1999 Story Segmentation Primary Systems

[1] As a first time participant, NTU ran their topic detection on a monolingual Mandarin condition.

[2] Newswire texts are not used for the story segmentation evaluation. Newswires routinely define units of texts as stories so there is no need for automated technologies.

From a detection theoretical standpoint, detection costs are a natural way to express system performance. However, humans do better with tangible concepts like story boundaries being incorrect by a fixed offset. The right graph in Figure 7 shows the segmentation cost of a hypothetical system that missplaces each story boundary by a fixed time offset[3]. Even though the English and Mandarin segmentation costs are different, when we compare the costs to fixed boundary offsets, the fixed offset times are 3.5 and 3.4 respectively.

Two new broadcast news sources were added to this year's evaluation, NBC Nightly News and MSNBC. The segmentation performance for these sources tested the systems' ability to segment shows not represented in the training data. Figure 8 shows system performance as a function of broadcast source. While the performance is lower than most of the data sources, the performance degradation is not significant.

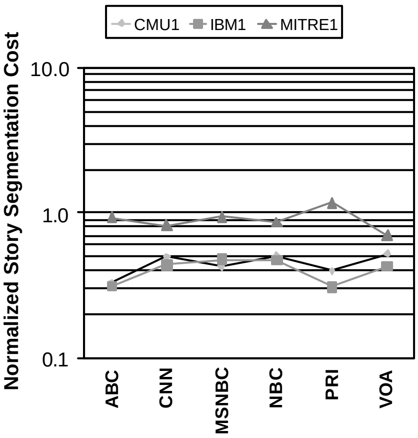

Figure 8 Primary segmentation system performance as a function of English source

Both CMU and IBM participated in last year's evaluation. For 1999, CMU made no substantive changes to their system, so their performance on this year's data can be compared to last year's data to infer test set difficulty. For last year's data, CMU achieved a 0.49 segmentation cost and this year they achieved a 0.47 segmentation cost on the 1999 English subset, only a 3% relative difference. However, IBM's 30% relative improvement (segmentation costs of 0.55 to 0.39) suggests they made appreciable improvements to their system.

2.4. First Story Detection Task

The first story detection task (FSD) evaluates technologies that detect, or find, the first story, and only the first story, in a stream to discuss a topic. This special case of the topic detection task focuses on the specific aspect of topic detection associated with novel information detection.

CMU and the Univ. of Mass. participated in this first formal evaluation of FSD. Their normalized FSD costs were 0.74 and 0.81 respectably. Figure 9 is the DET plot for comparing the two systems.

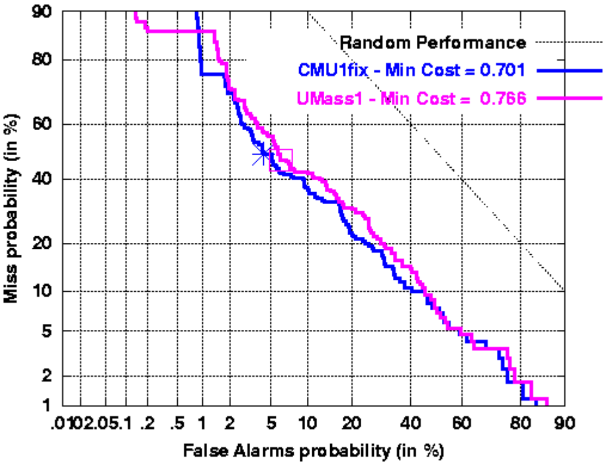

Figure 9 1999 Primary First Story Detection DET Curves

The FSD costs are appreciably higher than for most of the other tasks and are the theoretically bounded [6] by current topic tracking performance.

2.5. Link Detection Task

The link detection task evaluations technologies that detect when pairs of stories discuss the same topic (i.e. the stories are "linked" by a common topic). In its simplest evaluation condition, this task answers the YES/NO question: "do these two stories discuss the same topic." This task can be thought of as a core capability from which topic tracking and topic detection systems can be built.

Both CMU and the Univ. of Mass. participated in this year's evaluation. Their scores were 0.11 and 0.10 respectively for actual decisions, (0.096 and 0.088 for the minimum DET curve costs). From the DET curves in Figure 10, we can see that the systems performed similarly.

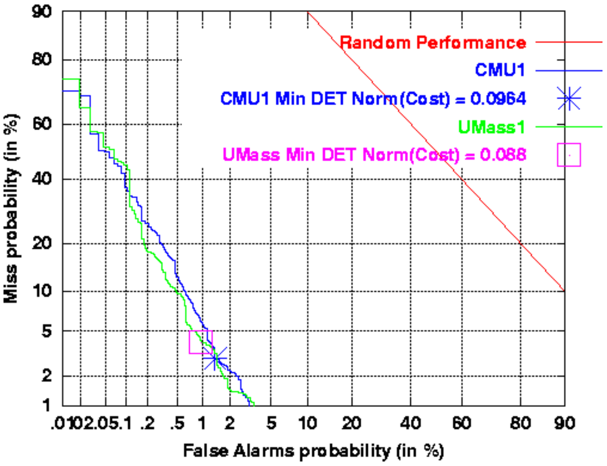

Figure 10 1999 TDT Primary Link Detection Systems

[3] Offsets are measured by time since English words and Mandarin characters do not represent the same units.

Humans are better at this task even though system performance is good. Human annotators achieved a link detection cost of 0.06, which is 38% better than the automatic systems.

3. TDT 2000 AND BEYOND

The 2000 TDT Evaluation will occur during September of 2000. Results will be discussed after the TREC conference in November of 2000. After the 2000 evaluation, the project will be extended again in light of the results. The number of languages is expected to grow and the range of sources is expected to increase for the languages other than English. Parties interested in future evaluations should contact NIST[4].

4. CONCLUSIONS

The 1999 Topic Detection and Tracking evaluation project involved eleven research groups, LDC and NIST. The project incorporated Mandarin source data for the first time, and two new evaluation tasks were added. The project supported five evaluation tasks: namely Topic Tracking, Topic Detection, Story Segmentation, First Story Detection and Link Detection. The best performance by task, (i.e., lowest detection costs), were 0.09, 0.26, 0.32, 0.74 and 0.10 respectively.

Through conditioned analysis of the results, several generalizations were drawn from the evaluation:

- cross-lingual topic tracking performance degraded by 44% compared to monolingual tracking,
- multilingual topic detection performance degraded by 37% compared to monolingual topic detection,
- English and Mandarin story segmentation performance is similar, and
- English story segmentation performance did not degrade on new, unseen broadcast sources.

5. DISCLAIMER

The views expressed in this paper are those of the authors. The test results are for local, system-developer-implemented tests. NIST's role was one that involved working with the community to define the evaluation task definitions, develop and implement scoring software, and score and tabulate the results. The views of the authors and these results are not to be construed or represented as endorsements of any systems or as official findings on the part of NIST or the U. S. Government.

5. References

1. "1998 TDT-2 Evaluation Specification Version 3.7" http://www.nist.gov/speech/tests/tdt/tdt99

2. TDT Web Site http://www.nist.gov/TDT

3. Fiscus, J., Doddington, G., Garofolo, J., and Martin, A., "NIST's 1998 Topic Detection and Tracking Evaluation (TDT)", Fifth European Conf. On Speech Comm. and Tech., Vol. 4, pp. 247-250

4. Cieri, C., Graff, D., Libermann, M., Martey, N., Strassel, S., "Large, Multilingual, Broadcast News Corpora for Cooperative Research in Topic Detection and Tracking: The TDT-2 and TDT-3 Corpus Efforts", Second International Conference on Language Resources and Evaluation, 31 May - 2 June, 2000, pp. 925-930.

5. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., "The DET Curve in Assessment of Detection Task Performance", 1997 Fifth European Conf. On Speech Comm. and Tech., Vol. 4, pp. 1895-1898

6. Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R., and Caputo, D., "Topic-based novelty detection" 1999 summer workshop at CLSP, final report. http://www.clsp.jhu.edu/ws99/projects/tdt

7. DARPA Translingual Information Detection and Summarization Program http://www.darpa.mil/ito/research/tides/index.html

8. Wayne, C., "Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation", Second International Conference on Language Resources and Evaluation, 31 May - 2 June, 2000, pp. 1487-1493.

[4] Jonathan.fiscus@nist.gov