Optimizing Story Link Detection is not Equivalent to Optimizing New Event Detection

Ayman Farahat
PARC
3333 Coyote Hill Rd
Palo Alto, CA 94304
farahat@parc.com

Francine Chen
PARC
3333 Coyote Hill Rd
Palo Alto, CA 94304
fchen@parc.com

Thorsten Brants
PARC
3333 Coyote Hill Rd
Palo Alto, CA 94304
thorsten@brants.net

Abstract

Link detection has been regarded as a core technology for the Topic Detection and Tracking tasks of new event detection. In this paper we formulate story link detection and new event detection as information retrieval task and hypothesize on the impact of precision and recall on both systems. Motivated by these arguments, we introduce a number of new performance enhancing techniques including part of speech tagging, new similarity measures and expanded stop lists. Experimental results validate our hypothesis.

1 Introduction

Topic Detection and Tracking (TDT) research is sponsored by the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program. The research has five tasks related to organizing streams of data such as newswire and broadcast news (Wayne, 2000): story segmentation, topic tracking, topic detection, new event detection (NED), and link detection (LNK). A link detection system detects whether two stories are "linked", or discuss the same event. A story about a plane crash and another story about the funeral of the crash victims are considered to be linked. In contrast, a story about hurricane Andrew and a story about hurricane Agnes are not linked because they are two different events. A new event detection system detects when a story discusses a previously unseen or "not linked" event. Link detection is considered to be a core technology for new event detection and the other tasks.

Several groups are performing research in the TDT tasks of link detection and new event detection. Based on their findings, we incorporated a number of their ideas into our baseline system. CMU (Yang et al., 1998) and UMass (Allan et al., 2000a) found that for new event detection it was better to compare a new story against all previously seen stories than to cluster previously seen stories and compare a new story against the clusters. CMU (Carbonell et al., 2001) found that NED results could be improved by developing separate models for different news sources to that could capture idiosyncrasies of different sources, which we also extended to link detection. UMass reported on adapting a tracking system for NED detection (Allan et al., 2000b). Allan *et. al* , (Allan et al., 2000b) developed a NED system based upon a tracking technology and showed that to achieve high-quality first story detection, tracking effectiveness must improve to a degree that experience suggests is unlikely. In this paper, while we reach a similar conclusion as (Allan et al., 2000b) for LNK and NED systems , we give specific directions for improving each system separately. We compare the link detection and new event detection tasks and discuss ways in which we have observed that techniques developed for one task do not always perform similarly for the other task.

2 Common Processing and Models

This section describes those parts of the processing steps and the models that are the same for New Event Detection and for Link Detection.

2.1 Pre-Processing

For pre-processing, we tokenize the data, recognize abbreviations, normalize abbreviations, remove stop-words, replace spelled-out numbers by digits, add part-of-speech tags, replace the tokens by their stems, and then generate term-frequency vectors.

2.2 Incremental TF-IDF Model

Our similarity calculations of documents are based on an incremental TF-IDF model. In a TF-IDF model, the frequency of a term in a document (TF) is weighted by the inverse document frequency (IDF). In the incremental model, document frequencies $df(w)$ are not static but change in time steps t. At time t, a new set of test documents C_t is added to the model by updating the frequencies

$$df_t(w) = df_{t-1}(w) + df_{Ct}(w) \tag{1}$$

where df_{Ct} denote the document frequencies in the newly added set of documents C_t. The initial document frequencies $df_0(w)$ are generated from a (possibly emtpy) training set. In a static TF-IDF model, new words (i.e., those words, that did not occur in the training set) are ignored in further computations. An incremental TF-IDF model uses the new vocabulary in similarity calculations. This is an advantage because new events often contain new vocabulary.

Very low frequency terms w tend to be uninformative. We therefore set a threshold θ_d. Only terms with $df_t(w) \geq \theta_d$ are used at time t. We use $\theta_d = 2$.

2.3 Term Weighting

The document frequencies as described in the previous section are used to calculate weights for the terms w in the documents d. At time t, we use

$$weight_t(d,w) = \frac{1}{Z_t(d)} f(d,w) \cdot \log \frac{N_t}{df_t(w)} \tag{2}$$

where N_t is the total number of documents at time t. $Z_t(d)$ is a normalization value such that either the weights sum to 1 (if we use Hellinger distance, KL-divergence, or Clarity-based distance), or their squares sum to 1 (if we use cosine distance).

2.4 Similarity Calculation

The vectors consisting of normalized term weights $weight_t$ are used to calculate the similarity between two documents d and q. In our current implementation, we use the the Clarity metric which was introduced by (Croft et al., 2001; Lavrenko et al., 2002) and gets its name from the distance to general English, which is called *Clarity*. We used a symmetric version that is computed as:

$$\begin{aligned} sim(d,q) &= -KL(d||q) + KL(d||GE) \\ &\quad -KL(q||d) + KL(q||GE) \end{aligned} \tag{3}$$

$$KL(d,q) = \sum_w weight_t(d,w) \cdot log(\frac{weight_t(d,w)}{weight_t(q,w)}. \tag{4}$$

where "KL" is the Kullback-Leibler divergence, GE is the probability distribution of words for "general English" as derived from the training corpus. The idea behind this metric is that we want to give credit to similar pairs of documents that are very different from general English, and we want to discount similar pairs of documents that are close to general English (which can be interpreted as being the noise). The motivation for using the clarity metric will given in section 6.1.

Another metric is Hellinger distance

$$sim_t(d,q) = \sum_w \sqrt{weight_t(d,w) \cdot weight_t(q,w)}. \tag{5}$$

Other possible similarity metrics are the cosine distance, the Kullback-Leibler divergence, or the symmetric form of it, Jensen-Shannon distance.

2.5 Source-Specific TF-IDF Model

Documents in the stream of news stories may stem from different sources, e.g., there are 20 different sources in the data for TDT 2002 (ABC News, Associated Press, New York Times, etc). Each source might use the vocabulary differently. For example, the names of the sources, names of shows, or names of news anchors are much more frequent in their own source than in the other ones. In order to reflect the source-specific differences we do not build one incremental TF-IDF model, but as many as we have different sources and use frequencies

$$df_{s,t}(w) \tag{6}$$

for source s at time t. The frequencies are updated according to equation (1), but only using those documents in C_t that are from the same source s. As a consequence, a term like “CNN” receives a high document frequency (thus low weight) in the model for the source CNN and a low document frequency (thus high weight) in the New York Times model.

Instead of the overall document frequencies $df_t(w)$, we now use the source specific $df_{s,t}(w)$ when calculating the term weights in equation (2).

Sources s for which no training data is available (i.e., no data to generate $df_{s,0}(w)$ is available) might be initialized in two different ways:

1. Use an empty model: $df_{s,0}(w) = 0$ for all w;
2. Identify one or more other but similar sources s' for which training data is available and use

$$df_{s,0}(w) = \sum_{s'} df_{s',0}(w). \tag{7}$$

2.6 Source-Pair-Specific Normalization

Due to stylistic differences between various sources, e.g., news paper vs. broadcast news, translation errors, and automatic speech recognition errors (Allan et al., 1999), the similarity measures for both on-topic and off-topic pairs will in general depend on the source pair. Errors due to these differences can be reduced by using thresholds conditioned on the sources (Carbonell et al., 2001), or, as we do, by normalizing the similarity values based on similarities for the source pairs found in the story history.

3 New Event Detection

In order to decide whether a new document q that is added to the collection at time t describes a new event, it is individually compared to all previous documents d using the steps described in section 2. We identify the document d^* with highest similarity:

$$d^* = \operatorname*{argmax}_d sim_t(q, d). \tag{8}$$

The value $score(q) = 1 - sim_t(q, d^*)$ is used to determine whether a document q is about a new event and at the same time is an indication of the confidence in our decision. If the score exceeds a threshold θ_s, then there is no sufficiently similar previous document, thus q describes a new event (decision YES). If the score is smaller than θ_s, then d^* is sufficiently similar, thus q describes an old event (decision NO). The threshold θ_s can be determined by using labeled training data and calculating similarity scores for document pairs on the same event and on different events.

4 Link Detection

In order to decide whether a pair of stories d and q are linked, we identify a set of similarity metrics $\mathbf{S}$ that capture the similarity between the two documents using Clarity and Hellinger metrics:

$$\mathbf{S}(d, q) = \langle sim_c(d, q), sim_h(d, q) \rangle. \tag{9}$$

The value $\mathbf{S}(d, q)$ is used to determine whether stories “q” and “d” are linked. If the similarity exceeds a threshold θ_{lnk} we the two stories are sufficiently similar (decision YES). If the similarity is smaller than θ_{lnk} we the two stories are sufficiently different (decision NO). The Threshold θ_{lnk} can be determined using labeled training data.

5 Evaluation

All TDT systems are evaluated by calculating a *Detection Cost*:

$$C_{Det} = C_{miss} \cdot P_{miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{nontarget}. \tag{10}$$

where C_{miss} and C_{FA} are the costs of a miss and a false alarm. They are set to 1 and 0.1, respectively, for all tasks. P_{miss} and P_{FA} are the conditional probabilities of a miss and a false alarm in the system output. P_{target} and $P_{nontarget}$ a the a priori target and non-target probabilities. They are set to 0.02 and 0.98 for LNK and NED. The detection cost is normalized such that a perfect system scores 0, and a random baseline scores 1:

$$(C_{Det})_{Norm} = \frac{C_{Det}}{\min(C_{miss} \cdot P_{target}, C_{FA} \cdot P_{nontarget})} \tag{11}$$

TDT evaluates all systems with a topic-weighted method: error probabilities are accumulated separately for each topic and then averaged. This is motivated by the different sizes of the topics.

The evaluation yields two costs: the *detection cost* is the cost when using the actual decisions made by the system; the *minimum detection cost* is the cost when using the confidence scores that each system has to emit with each decision and selecting the optimal threshold based on the score.

In the TDT-2002 evaluation, our Link Detection system was the best of three systems, yielding $(C_{lnk})_{Norm} = 0.1947$ and $min(C_{lnk})_{Norm} = 0.1926$. Our New Event Detection system was ranked second of four with costs of $(C_{ned})_{Norm} = 0.5691$ and $min(C_{ned})_{Norm} = 0.5303$.

6 Differences between LNK and NED

In this section, we draw on Information retrieval tools to analyze LNK and NED tasks. Motivated by the results of this analysis, we compare a number of techniques in the LNK and NED tasks in particular we compare the utility of two similarity measures, part-of-speech tagging, stop wording, and normalizing abbreviations and numerals. The comparisons were performed on corpora developed for TDT, including TDT2 and TDT3.

6.1 Information Retrieval and TDT

The conditions for false alarms and misses are reversed for LNK and NED tasks. In the LNK task, incorrectly flagging two stories as being on the same event is considered a false alarm. In contrast in the NED task, incorrectly flagging two stories as being on the same event will cause the true first story to be missed. Conversely, in LNK incorrectly labeling two stories that are on the same event as not linked is a miss, but in the NED task, incorrectly labeling two stories on the same event as not linked can result in a false alarm where a story is incorrectly identified as a new event.

The detection cost in Eqn.10 which assigns a higher cost to false alarm $C_{miss} \cdot P_{target} = 0.02, C_{FA} \cdot P_{nontarget} = 0.098$. A LNK system wants to minimize false alarms and to do this it should identify stories as being linked only if they are linked, which translates to high precision. In contrast a NED system, will minimize false alarms by identifying all stories that are linked which translates to high recall. Motivated by this discussion, we investigated the use of number of precision and recall enhancing techniques with the LNK and NED system. We investigated the use of the Clarity metric (Lavrenko et al., 2002) which was shown to correlate positively with precision. We investigated the use of part-of-speech tagging which was shown by Allan and Raghavan (Allan and Raghavan, 2002) to improve query clarity. In section 6.2.1 we will show how POS helps recall. We also investigated the use of expanded stop-list which improves precision. We also investigated normalizing abbreviations and transforming spelled out numbers into numbers. On the one hand the enhanced processing list includes most of the term in the ASR stop-list and removing these terms will improve precision. On the other hand normalizing these terms will have the same effect as stemming a recall enhancing device (Xu and Croft, 1998) , (Kraaij and Pohlmann, 1996). In addition to these techniques, we also investigated the use of different similarity measures.

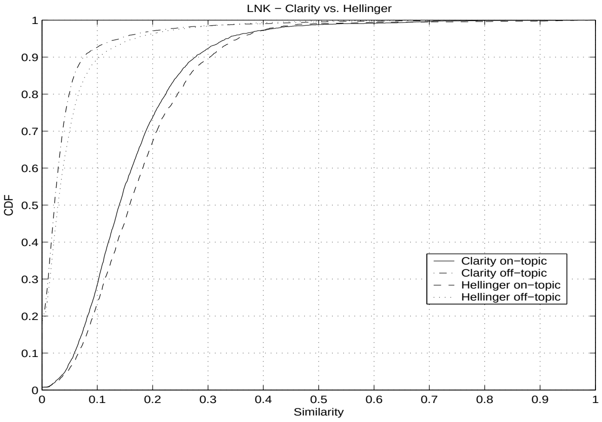

Figure 1: CDF for Clarity and Hellinger similarity on the LNK task for on-topic and off-topic pairs.

6.2 Similarity Measures

The systems developed for TDT primarily use cosine similarity as the similarity measure. We have developed systems based on cosine similarity (Chen et al., 2003). In work on text segmentation, (Brants et al., 2002) observed that the system performance was much better when the Hellinger measure was used instead. In this work, we decided to use the clarity metric, a precision enhancing device (Croft et al., 2001). For both our LNK and NED systems, we compared the performance of the systems using each of the similarity measures separately. Table 1 shows that for LNK, the system based on Clarity similarity performed better the system based on Hellinger similarity; in contrast, for NED, the system based on

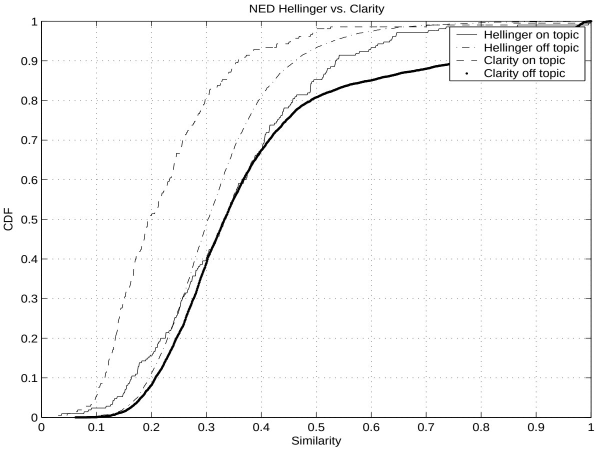

Figure 2: CDF for Clarity and Hellinger similarity on the NED task for on-topic and off-topic pairs.

Table 1: Effect of different similarity measures on topic-weighted minimum normalized detection costs for LNK and NED on the TDT 2002 dry run data.

System	Clarity	Hellinger	Δ	% Chg
LNK	0.3054	0.3777	-0.0597	-19.2
NED	0.8419	0.5873	+0.2546	+30.24

Hellinger similarity performed better.

Figure 1 shows the cumulative density function for the Hellinger and Clarity similarities for on-topic (about the same event) and off-topic (about different events) pairs for the LNK task. While there are a number of statistics to measure the overall difference between tow cumulative distribution functions, we used the Kolmogorov-Smirnov distance (K-S distance; the largest difference between two cumulative distributions) for two reasons. First, the K-S distance is invariant under re-parametrization. Second, the significance of the K-S distance in case of the null hypothesis (data sets are drawn from same distribution) can be calculated (Press et al., 1993). The K-S distance between the on-topic and off-topic similarities is larger for Clarity similarity (cf. table 2), indicating that it is the better metric for LNK.

Figure 2 shows the cumulative distribution functions for Hellinger and Clarity similarities in the NED task. The plot is based on pairs that contain the current story and its most similar story in the story history. When the most similar story is on the same event (approx. 75% of the cases), its similarity is part of the on-topic distribution, otherwise (approx. 25% of the cases) it is plotted as off-topic. The K-S distance between the Hellinger on-topic and off-topic CDFs is larger than those for Clarity (cf. table 2). For both NED and LNK, we can reject the null hypothesis for both metrics with over 99.99 % confidence.

Table 2: K-S distance between on-topic and off-topic story pairs.

	Clarity	Hellinger	Change (%)
LNK	0.7680	0.7251	-0.0429 (-5.58%)
NED	0.5353	0.6055	$+0.0702$ ($+13.14\%$)

Table 3: Effect of using part-of-speech on minimum normalized detection costs for LNK and NED on the TDT 2002 dry run data.

System	$-$ PoS	$+$ PoS	Change (%)
LNK	0.3054	0.4224	-0.117 (-38.3%)
NED	0.6403	0.5873	+0.0530 ($+8.3\%$)

To get the high precision required for LNK system, we need to have a large separation between the on-topic and off-topic distributions. Examining Figure 1 and Table 2 , indicates that the Clarity metric has a larger separation than the Hellinger metric. At high recall required by NED system (low CDF values for on-topic), there is a greater separation with the Hellinger metric. For example, at 10% recall, the Hellinger metric has 71 % false alarm rate as compared to 75 % for the Clarity metric.

6.2.1 Part-of-Speech (PoS) Tagging

We explored the idea that noting the part-of-speech of the terms in a document may help to reduce confusion among some of the senses of a word. During pre-processing, we tagged the terms as one of five categories: adjective, noun, proper nouns, verb, or other. A “tagged term” was then created by combining the stem and part-of-speech. For example, ‘N_train’ represents the term ‘train’ when used as a noun, and ‘V_train’ represents the term ‘train’ when used as a verb. We then ran our NED and LNK systems using the tagged terms. The systems were tested in the Dry Run 2002 TDT data. A comparison of the performance of the systems when part-of-speech is used against a baseline sys-

Table 4: Comparison of using an “ASR stop-list” and “enhanced preprocessing” for handling ASR differences.

	No ASR stop	ASR stop
	Std Preproc	Std Preproc
LNK	0.3153	0.3054
NED	0.6062	0.6407

tem when part-of-speech is not used is shown in Table 3. For Story Link Detection, performance decreases by 38.3%, while for New Event Detection, performance improves by 8.3%. Since POS tagging helps differentiates between the different senses of the same root, it also reduces the number of matching terms between two documents. In the LNK task for example, the total number of matches drops from 177,550 to 151,132. This has the effect of placing a higher weight on terms that match, *i.e.* terms that have the same sense and for the TDT corpus will increase recall and decrease. Consider for example matching “food server to “food service” and “java server”. When using POS both terms will have the same similarity to the query and the use of POS will retrieve the relevant documents but will also retrieve other documents that share the same sense.

6.2.2 Stop Words

A large portion of the documents in the TDT collection has been automatically transcribed using Automatic Speech Recognition (ASR) systems which can achieve over 95% accuracies. However, some of the words not recognized by the ASR tend to be very informative words that can significantly impact the detection performance (Allan et al., 1999). Furthermore, there are systematic differences between ASR and manually transcribed text, *e.g.,* numbers are often spelled out thus “30” will be spelled out “thirty”. Another situation where ASR is different from transcribed text is abbreviations, *e.g.* ASR system will recognize ‘CNN” as three separate tokens “C”, “N”, and “N”.

In order to account for these differences, we identified the set of tokens that are problematic for ASR. Our approach was to identify a parallel corpus of manually and automatically transcribed documents, the TDT2 corpus, and then use a statistical approach (Dunning, 1993) to identify tokens with significantly different distributions in the two corpora. We compiled the problematic ASR terms into an “ASR stop-list”. This list was primarily composed of spelled-out numbers, numerals and a few other terms. Table 4 shows the topic-weighted minimum detection costs for LNK and NED on the TDT 2002 dry run data. The table shows results for standard preprocessing without an ASR stop-list and with and ASR stop-list. For Link Detection, the ASR stop-list improves results, while the same list decreases performance for New Event Detection.

Table 5: Impact of recall and precision enhancing devices.

Device	Impact	LNK	NED
ASR stop	precision	+3.1%	-5.5 %
POS	recall	-38.8 %	8.3 %
Clarity	precision	+19 %	-30 %

In (Chen et al., 2003) we investigated normalizing abbreviations and transforming spelled-out numbers into numerals, “enhanced preprocessing”, and then compared this approach with using an “ASR stop-list”.

6.2.3 Impact of Recall and Precision

The previous two sections examined the impact of four different techniques on the performance of LNK and NED systems. The Part-of-speech is a recall enhancing devices while the ASR stop-list is a precision enhancing device. The enhanced preprocessing improves precision and recall. The results which are summarized in Table 5 indicate that precision enhancing devices improved the performance of the LNK task while recall enhancing devices improved the NED task.

6.3 Final Remarks on Differences

In the extreme case, a perfect link detection system performs perfectly on the NED task. We gave empirical evidence that there is not necessarily such a correlation at lower accuracies. These findings are in accordance with the results reported in (Allan et al., 2000b) for topic tracking and first story detection.

To test the impact of the cost function on the performance of LNK and NED systems, we repeated the evaluation with C_{miss} and C_{fa} both set to 1, and we found that the difference between the two re-

Table 6: Topic-weighted minimum normalized detection cost for NED when using parameter settings that are best for NED (1) and those that are best for LNK (2). Columns (3) and (4) show the detection costs using uniform costs for misses and false alarms.

	(1)	(2)	(3)	(4)
Metric	Hel	Cla	Hel	Cla
POS	$+$	$-$	$+$	$-$
ASR stop	$-$	$+$	$-$	$+$
C_{fa}	0.1	0.1	1	1
min C^{NED}_{norm}	0.5873	0.8419	0.8268	0.9498
% change	–	+30.24%	–	+14.73%

sults decreases from 30.24% to 14.73%. The result indicates that the setting (Hel, $+$PoS, $-$ASRstop) is better at recall (identifying same-event stories), while (Clarity, $-$PoS, $+$ASRstop) is better at precision (identifying different-event stories).

In addition to the different costs assigned to misses and false alarms, there is a difference in the number of positives and negatives in the data set (the TDT cost function uses $p_{target} = 0.02$). This might explain part of the remaining difference of 14.73%.

Another view on the differences is that a NED system must perform very well on the higher penalized first stories when it does not have any training data for the new event, event though it may perform worse on follow-up stories. A LNK system, however, can afford to perform worse on the first story if it compensates by performing well on follow-up stories (because here not flagged follow-up stories are considered misses and thus higher penalized than in NED). This view explains the benefits of using part-of-speech information and the negative effect of the ASR stop-list on NED : different part-of-speech tags help discriminate new events from old events; removing words by using the ASR stoplist makes it harder to discriminate new events. We conjecture that the Hellinger metric helps improve recall, and in a study similar to (Allan et al., 2000b) we plan to further evaluate the impact of the Hellinger metric on a closed collection *e.g.* TREC.

7 Conclusions and Future Work

We have compared the effect of several techniques on the performance of a story link detection system and a new event detection system. Although many of the processing techniques used by our systems are the same, a number of core technologies affect the performance of the LNK and NED systems differently. The Clarity similarity measure was more effective for LNK, Hellinger similarity measure was more effective for NED, part-of-speech was more useful for NED, and stop-list adjustment was more useful for LNK. These differences may be due in part to a reversal in the tasks: a miss in LNK means the system does not flag two stories as being on the same event when they actually are, while a miss in NED means the system does flag two stories as being on the same event when actually they are not. In future work, we plan to evaluate the impact of the Hellinger metric on recall. In addition, we plan to use Anaphora resolution which was shown to improve recall (Pirkola and Jrvelin, 1996) to enhance the NED system.

References

James Allan and Hema Raghavan. 2002. Using part-of-speech patterns to reduce query ambiguity. In *ACM SIGIR2002*, Tampere, Finland.

James Allan, Hubert Jin, Martin Rajman, Charles Wayne, and et. al. 1999. Topic-based novelty detection. Summer workshop final report, Center for Language and Speech Processing, Johns Hopkins University.

J. Allan, V. Lavrenko, D. Malin, and R. Swan. 2000a. Detections, bounds, and timelines: Umass and tdt-3. In *Proceedings of Topic Detection and Tracking Workshop (TDT-3)*, Vienna, VA.

James Allan, Victor Lavrenko, and Hubert Jin. 2000b. First story detection in TDT is hard. In *CIKM*, pages 374–381.

Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *International Conference on Information and Knowledge Management (CIKM)*, McLean, VA.

Jaime Carbonell, Yiming Yang, Ralf Brown, Chun Jin, and Jian Zhang. 2001. Cmu tdt report. Slides at the TDT-2001 meeting, CMU.

Francine Chen, Ayman Farahat, and Thorsten Brants. 2003. Story link detection and new event detection are asymmetric. In *Proceedings of NAACL-HLT-2002*, Edmonton, AL.

W. Bruce Croft, Stephen Cronen-Townsend, and Victor Larvrenko. 2001. Relevance feedback and personalization: A language modeling perspective. In *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*.

Ted E. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Wessel Kraaij and Renee Pohlmann. 1996. Viewing stemming as recall enhancement. In *ACM SIGIR1996*.

Victor Lavrenko, James Allan, Edward DeGuzman, Daniel LaFlamme, Veera Pollard, and Stephen Thomas. 2002. Relevance models for topic detection and tracking. In *Proceedings of HLT-2002*, San Diego, CA.

A. Pirkola and K. Jrvelin. 1996. The effect of anaphora and ellipsis resolution on proximity searching in a text database. *Information Processing and Management*, 32(2):199–216.

William H. Press, Saul A. Teukolsky, William Vetterling, and Brian Flannery. 1993. *Numerical Recipes*. Cambridge Unv. Press.

Charles Wayne. 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Language Resources and Evaluation Conference (LREC)*, pages 1487–1494, Athens, Greece.

Jinxi Xu and W. Bruce Croft. 1998. Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems*, 16(1):61–81.

Yiming Yang, Tom Pierce, and Jaime Carbonell. 1998. A study on retrospective and on-line event detection. In *Proceedings of SIGIR-98*, Melbourne, Australia.