Using Names and Topics for New Event Detection

Giridhar Kumaran and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003, USA
{giridhar,allan}@cs.umass.edu

Abstract

New Event Detection (NED) involves monitoring chronologically-ordered news streams to automatically detect the stories that report on new events. We compare two stories by finding three cosine similarities based on names, topics and the full text. These additional comparisons suggest treating the NED problem as a binary classification problem with the comparison scores serving as features. The classifier models we learned show statistically significant improvement over the baseline vector space model system on all the collections we tested, including the latest TDT5 collection.

The presence of automatic speech recognizer (ASR) output of broadcast news in news streams can reduce performance and render our named entity recognition based approaches ineffective. We provide a solution to this problem achieving statistically significant improvements.

1 Introduction

The instant and automatic detection of new events is very useful in situations where novel information needs to be detected from a real-time stream of rapidly growing data. These real-life situations occur in scenarios like financial markets, news analyses, and intelligence gathering. In this paper we focus on creating a system to immediately identify stories reporting new events in a stream of news - a daunting task for a human analyst given the enormous volume of data coming in from various sources.

The Topic Detection and Tracking (TDT) program, a DARPA funded initiative, seeks to develop technologies that search, organize and structure multilingual news-oriented textual materials from a variety of broadcast news media. One of the tasks in this program, New Event Detection (NED), involves constant monitoring of streams of news stories to identify the first story reporting topics of interest. A *topic* is defined as "a seminal event or activity, along with directly related events and activities" (Allan, 2002). An earthquake at a particular place is an example of a topic. The first story on this topic is the story that first carries the report on the earthquake's occurrence. The other stories that make up the topic are those discussing the death toll, the rescue efforts, the reactions from different parts of the world, scientific discussions, the commercial impact and so on. A good NED system would be one that correctly identifies the article that reports the earthquake's occurrence as the first story.

NED is a hard problem. For example, to distinguish stories about earthquakes in two different places, a vector space model system would rely on a tf-idf weighting scheme that will bring out the difference by weighting the locations higher. More often then not, this doesn't happen as the differences are buried in the mass of terms in common between stories describing earthquakes and their aftermath. In this paper we reduce the dependence on tf-idf weighting by showing the utility of creating

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 121–128, Vancouver, October 2005.

three distinct representations of each story based on named entities. This allows us to view NED as a binary classification problem - i.e., each story has to be classified into one of two categories - *old* or *new*, based on features extracted using the three different representations.

The paper starts by summarizing the previous work on NED in Section 2. In Section 3, we explain the rationale behind our intuition. Section 4 describes the experimental setup, data pre-processing, and our baseline NED system. We then briefly describe the evaluation methodology for NED in Section 5. Model creation and the results of applying these models to test data are detailed in Section 6. In the same section, we describe the effect on performance if the manually transcribed version of broadcast news is replaced with ASR output. Since its hard to recognize named entities from ASR data, performance expectedly deteriorates. We follow a novel approach to work around the problem resulting in statistically significant improvement in performance. The results are analyzed in Section 7. We wrap up with conclusions and future work in Section 8.

2 Previous Research

Previous approaches to NED have concentrated on developing similarity metrics or better document representations or both. A summer workshop on topic-based novelty detection held at Johns Hopkins University extensively studied the NED problem. Similarity metrics, effect of named entities, pre-processing of data, and language and Hidden Markov Models were explored (Allan et al., 1999). Combinations of NED systems were also discussed. In the context of this paper, selective re-weighting of named entities didn't bring about expected improvement.

Improving NED by better comparison of stories was the focus of following papers. In an approach to solve on-line NED, when a new document was encountered it was processed immediately to extract features and build up a query representation of the document's content (Papka and Allan, 1998). The document's initial threshold was determined by evaluating it with the query. If the document did not trigger any previous query by exceeding this particular threshold, it was marked as a new event. Unlike the previous paper, good improvements on TDT benchmarks were shown by extending a basic incremental TF-IDF model to include source-specific models, similarity score normalization techniques, and segmentation of documents (Brants et al., 2003).

Other researchers have attempted to build better document models. A combination of evidence derived from two distinct representations of a document's content was used to create a new representation for each story (Stokes and Carthy, 2001). While one of the representations was the usual free text vector, the other made use of lexical chains (created using WordNet) to obtain the most prevalent topics discussed in the document. The two vectors were combined in a linear fashion and a marginal increase in effectiveness was observed.

NED approaches that rely on exploiting existing news tracking technology were proved to inevitably exhibit poor performance (Allan et al., 2000). Given tracking error rates, the lower and upper bounds on NED error rates were derived mathematically. These values were found to be good approximations of the true NED system error rates. Since tracking and filtering using full-text similarity comparison approaches were not likely to make the sort of improvements that are necessary for high-quality NED results, the paper concluded that an alternate approach to NED was required. This led to a series of research efforts that concentrated on building multi-stage NED algorithms and new ways to combine evidence from different sources.

In the topic-conditioned novelty detection approach, documents were classified into broad topics and NED was performed within these categories (Yang et al., 2002). Additionally, named entities were re-weighted relative to the normal words for each topic, and a stop list was created for each topic. The experiments were done on a corpus different from the TDT corpus and, apparently didn't scale well to the TDT setting.

The DOREMI research group treated named entities like *people* and *locations* preferentially and developed a new similarity measure that utilized the semantics classes they came up with (Makkonen et al., 2002). They explored various definitions of the NED task and tested their system accordingly. More recently, they utilized a perceptron to learn a weight

function on the similarities between different semantic classes to obtain a final confidence score for each story (Makkonen et al., 2004).

The TDT group at UMass introduced multiple document models for each news story and modified similarity metrics by splitting up stories into only named entities and only terms other than named entities (Kumaran and Allan, 2004). They observed that certain categories of news were better tackled using only named entities, while using only topic terms for the others helped.

In approaches similar to named entity tagging, part-of-speech tagging (Farahat et al., 2003) has also been successfully used to improve NED.

Papers in the TDT2003 and TDT2004 workshops validated the hypothesis that ensemble single-feature classifiers based on majority voting exhibited better performance than single classifiers working with a number of features on the NED task (Braun and Kaneshiro, 2003; Braun and Kaneshiro, 2004). Examples of features they used are cosine similarity, text tiling output and temporally-weighted tf-idf.

Probabilistic models for online clustering of documents, with a mechanism for handling creation of new clusters have been developed. Each cluster was assumed to correspond to a topic. Experimental results did not show any improvement over baseline systems (Zhang et al., 2005).

3 Features for NED

Pinning down the character of new stories is a tough process. New events don't follow any periodic cycle, can occur at any instant, can involve only one particular type of named entity (people, places, organizations etc.) or a combination, can be reported in any language, and can be reported as a story of any length by any source[1]. Apart from the source, date, and time of publication or broadcast of each news story, the TDT corpora do not contain any other clues like placement in the webpage, the number of sources reporting the same news and so on. Given all these factors, we decided that the best features to use would be those that were not particular to the story in question only, but those that measure differences between the story and those it is compared with.

Category-specific rules that modified the baseline confidence score assigned to each story have been developed (Kumaran and Allan, 2004). The modification was based on additional evidence in the form of overlap of named entities and topic terms (terms in the document not identified as named entities) with the closest story reported by a baseline system. We decided to use these three scores: namely the baseline confidence score, named entity overlap, and topic-term overlap as features. The named entities considered were *Event*, *Geopolitical Entity*, *Language*, *Location*, *Nationality*, *Organization*, *Person*, *Cardinal*, *Ordinal*, *Date*, and *Time*. These named entities were detected in stories using BBN IdentiFinder^{TM}(Bikel et al., 1999). Irrespective of their type, all named entities were pooled together to form a single named entity vector.

The intuition behind using these features is that we believe every event is characterized by a set of people, places, organizations, etc. (named entities), and a set of terms that describe the event. While the former can be described as the *who*, *where*, and *when* aspects of an event, the latter relates to the *what* aspect. If two stories were on the same topic, they would share both named entities as well as topic terms. If they were on different, but similar, topics, then either named entities or topic terms will match but not both.

We illustrate the above intuition with examples. Terms in **bold face** are named entities common to both stories, while those in *italics* are topic terms in common. We start with an example showing that for old stories both the named entities as well as topic terms overlap with a story on the same topic.

Story 1. : Story on a topic already reported

While in **Croatia** today, **Pope John Paul II** called on the *international community* to *help* end the fighting in the Yugoslavia's **Kosovo** province.

Story 2. : Story on the same topic

Pope John Paul II is urging the *international community* to quickly *help* the ethnic Albanians in **Kosovo**. He spoke in the coastal city of Split, where he ended a three-day visit to **Croatia**.

[1]It could be argued that articles from a source, say *NYTimes*, are much longer than news stories from *CNN*, and hence the length of stories is a good candidate for use as a feature. However, when there is no pattern that indicates that either of the two sources reports new stories preferentially, the use of length as a feature is moot.

Story 1 is an old story about Pope John Paul II's visit to Yugoslavia. Story 2 was the first story on the topic and it shares both named entities likes **Pope John Paul II** and **Croatia** and also topic terms like *international community* and *help*.

Our next example shows that for new stories, either the named entities or topic terms match with an earlier story.

Story 3. : Topic not seen before
Turkey has sent 10,000 troops to its southern border with Syria amid growing tensions between the two neighbors, newspapers reported Thursday. Defense Minister **Ismet Sezgin** denied any troop movement along the border, but said **Turkey's** patience was running out. **Turkey** accuses Syria of harboring Turkish Kurdish rebels fighting for autonomy in **Turkey's** southeast; it says rebel leader Abdullah Ocalan lives in Damascus.
Story 4. : Closest Story due to Named Entities
A senior **Turkish** government official called Monday for closer military cooperation with neighboring Bulgaria. After talks with President Petar Stoyanov at the end of his four-day visit, Turkish Deputy Premier and National Defense Minister **Ismet Sezgin** expressed satisfaction with the progress of bilateral relations and the hope that Bulgarian-**Turkish** military cooperation will be promoted.

Story 3 is a new story about the rising tensions between Turkey and Syria. The closest story as reported by a (baseline) basic vector space model NED system using cosine similarity is Story 4, a story about Turkish-Bulgarian relations. The named entities **Turkey** and **Ismet Sezgin** caused this match. We see that none of the topic terms match. However, the system reported with a high confidence score that Story 3 is old. This is because of the matching of high IDF-valued named entities. Determining that the topic terms didn't match would have helped the system avoid this mistake.

4 Experimental Setup and Baseline

We used the TDT2, TDT3, TDT4, and TDT5 corpora for our experiments. They contain a mix of broadcast news (*bn*) and newswire (*nwt*) stories. Only the English stories in the multi-lingual collections were considered for the NED task. The broadcast news material is provided in the form of an audio sampled data signal, a manual transcription of the audio signal (*bn-man*), and a transcription created using an automatic speech recognizer (*bn-asr*).

We used version 3.0 of the open source Lemur system[2] to tokenize the data, remove stop words, stem and create document vectors. We used the 418 stopwords included in the stop list used by InQuery (Callan et al., 1992), and the Krovetz-stemmer algorithm implementation provided as part of Lemur.

Documents were represented as term vectors with incremental TF-IDF weighting (Brants et al., 2003; Yang et al., 1998). We used the cosine similarity metric to judge the similarity of a story *S* with those seen in the past.

$$Sim(S,X) = \frac{\sum_w weight(w,S) * weight(w,X)}{\sqrt{\sum_w weight(w,S)^2}\sqrt{\sum_w weight(w,X)^2}} \quad (1)$$

where

$$\begin{aligned} weight(w,d) &= tf * idf \\ tf &= \log(termfrequency + 1.0) \\ idf &= \frac{\log((docCount+1)}{(documentfreq+0.5)} \end{aligned}$$

The maximum similarity of the story *S* with stories seen in the past was taken as the confidence score that *S* was old. This constituted our baseline system.

We extracted three features for each incoming story *S*. The first was the confidence score reported by the baseline system. The second and third features were the cosine similarity between only the named entities in *S* and *X* and the cosine similarity between only the topic terms in *S* and *X*. We trained a Support Vector Machine (SVM) (Burges, 1998) classifier on these features. We chose to use SVMs as they are considered state-of-the-art for text classification purposes (Mladeni et al., 2004), and provide us with options to consider both linear and non-linear decision boundaries. To develop SVM models we used SVM^{Light}(Joachims, 1999), which is an implementation of SVMs in C. SVM^{Light} is an implementation of Vapnik's Support Vector Machine (Vapnik, 1995).

For training, we used the TDT3 and TDT4 corpora. There were 115 and 70 topics respectively giving us a total of 185 positive examples (new stories)

[2] `http://www.lemurproject.org`

and 7800 negative examples (old stories). We balanced the number of positive and negative examples by oversampling the minority class until there were equal number of positive and negative training instances. Testing was done on the TDT2 and TDT5 corpora (96 and 126 topics resp.).

5 NED Evaluation

The official TDT evaluation requires a NED system to assign a confidence score between 0 (new) and 1 (old) to every story upon its arrival in the time-ordered news stream. This assignment of scores is done either immediately upon arrival or after a fixed look-ahead window of stories. To evaluate performance, the stories are sorted according to their scores, and a threshold sweep is performed. All stories with scores above the threshold are declared *old*, while those below it are considered *new*. At each threshold value, the misses and false alarms are identified, and a cost C_{det} is calculated as follows.

$$C_{det} = C_{miss} * P_{miss} * P_{target} + C_{FA} * P_{FA} * P_{non-target}$$

where C_{Miss} and C_{FA} are the costs of a Miss and a False Alarm, respectively, P_{Miss} and P_{FA} are the conditional probabilities of a Miss and a False Alarm, respectively, and P_{target} and $P_{non-target}$ are the a priori target probabilities ($P_{non-target}$ = 1 - P_{target}).

The threshold that results in the least cost is selected as the optimum one. Different NED systems are compared based on their minimum cost. In other words, the **lower** the C_{det} score reported by a system on test data, the **better** the system.

6 Results

Our first set of experiments were performed on data consisting of newswire text and manual transcription of broadcast news (*nwt+bn-man*). We used the features mentioned in Section 3 to build SVM models in the classification mode. We experimented with linear, polynomial, and RBF kernels. The output from the SVM classifiers was normalized to fall within the range zero and one.

We found that using certain kernels improved performance over the baseline system significantly. The results for both corpora, TDT2 and TDT5, were consistently and significantly improved by using the classification models. The 2004 NED evaluations conducted by the National Institute of Standards and Technology was on the TDT5 collection. The large size of the collection and existence of a large number of topics with a single story made the task very challenging. The best system fielded by the participating sites was the baseline system used here. Table 1 summarizes the results we obtained.

Kernel Type	TDT2 (*nwt+bn-man*)	TDT5 (*nwt*)
Baseline System	0.585	0.701
Linear Kernel	0.548	0.696
Poly. of deg. 1	0.548	0.696
Poly. of deg. 2	0.543	0.688
Poly. of deg. 3	0.545	0.684
Poly. of deg. 4	0.535	0.694
Poly. of deg. 5	0.533	0.688
Poly. of deg. 6	0.534	0.693
RBF with γ = 1	**0.540**	**0.661**
RBF with γ = 5	0.530	0.699

Table 1: Summary of the results of using SVM classifier models for NED on the TDT2 and TDT5 collections. The numbers are the minimum cost (C_{det}) values (lower is better). The sign test, with α = 0.05, was performed to compare the baseline system with only a classifier using RBF kernels with γ = 1. For both collections, the improvements were found to be statistically significant (shown in bold). While there are better performing kernels for TDT2, we chose to perform significance tests for only one kernel to show that significant improvement over the baseline can be obtained using a single kernel across different test collections.

All statistical significance testing was done using the sign test. We counted the number of topics for which using the SVM classifier improved over the baseline (in terms of detecting more previously undetected new and old stories), and also the number of topics for which using the SVM classifier actually converted originally correct decisions into wrong ones. These were used as input for the sign test. The test were used to check whether improvement in performance using the classifier-based system was spread across a significant number of topics, and not confined to a few. Table 2 gives some examples of topics and the associated improvements in detecting them.

Topic ID	Number of old stories	Num. detected by baseline system	Num. detected by SVM classifier	Improvement (Higher the better)
55105	420	407	403	-4
55010	21	21	20	-1
55023	5	5	4	-1
55089	226	226	225	-1
55125	120	114	120	6
55107	331	327	331	7
55106	808	787	795	8
55200	196	185	193	8

Table 2: Examples of improvements due to using the SVM classifier on a per-topic basis. Shown here are the four topics each in which the greatest degradation and improvements in performance were seen. The topics vary in size. The SVM classifier resulted in overall (statistically significant, refer Table 1) improvement as it corrected more errors than introduced them.

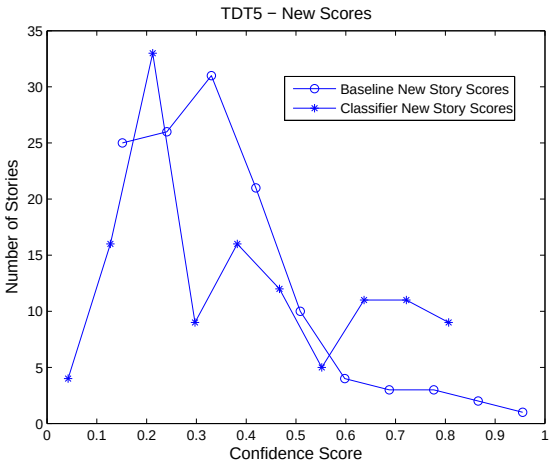

Figure 1: Distribution of new story scores for the baseline and SVM model systems.

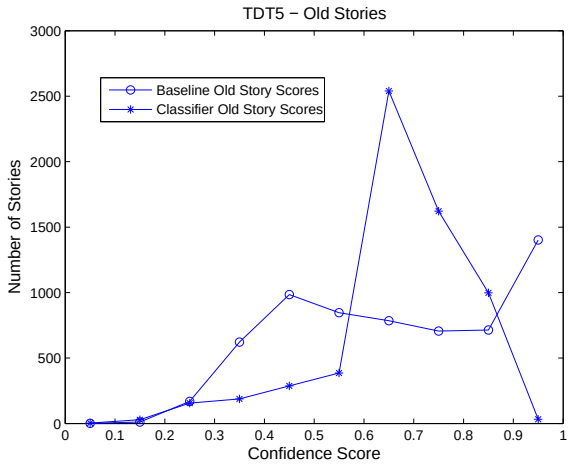

Figure 2: Distribution of old story scores for the baseline and SVM model systems.

7 Analysis

The main goal of our effort was to come up with a way to correctly identify new stories based on features we thought characterized them. To understand what we had actually achieved by using these models, we studied the distribution of the confidence scores assigned to new and old stories for the baseline and a classifier-based NED system for the TDT5 collection (Figures 1 and 2 respectively).

We observe that the scores for a small fraction of new stories that were initially missed (between scores 0.8 and 1) are decreased by the model-based NED system while a larger fraction (between scores 0.1 and 0.4) is also decreased by a small amount. However, the major impact of using SVM model-based NED systems appears to be in detecting old stories. We observe that the scores of a significant number of old stories (between scores 0.2 and 0.55) have been increased to be closer to one. This had the effect of increasing the score difference between old and new stories, and hence improved classification accuracy as measured by the minimum cost.

We investigated the relative importance of the three features by looking that the linear kernel SVM model. While the original cosine similarity metric *CS* remained the prominent feature, the contribution

of the third feature *non-NE-CS* was slightly more than if not equal to the contribution of named entities *NE-CS* (Table 3). This explains why simple re-weighting of named entities alone (Allan et al., 1999) doesn't suffice to improved performance.

Feature	*CS*	*NE-CS*	*non-NE-CS*
Weight	5.4	1.58	1.83

Table 3: Weights assigned to features by the linear kernel SVM.

If this method of harnessing named entities and topic terms were indeed so effective, then we should have been able to detect every old story in every topic. However, analysis reveals that this approach makes an assumption about the way stories in a topic are related. Not all topics are *dense*, with both named entities and topic terms threading the stories together. Examples of such topics are natural disaster topics. While the first story might report on the actual calamity and the region it affected, successive stories might report on individual survivor tales. These stories might be connected to the original story of the topic by as tenuous a link as only the name of the calamity, or the place. Such topic structures are very common in newswire. Hence our approach will fail in such topics with loosely connected stories. Much more advanced processing of story content is required in such cases. Mistakes made by the named entity recognizer also impede performance.

Given that its impractical to expect manual transcriptions of all broadcast news, we tested our baseline and classifier systems on a version of TDT2 with newswire stories and ASR output of the broadcast news (*nwt+bn-asr*). TDT5 was left out as it doesn't have any broadcast news. As shown in Table 4, the baseline system performed significantly worse when manual transcription was replaced with ASR output. The classifier systems did even worse than the *nwt +bn-asr* baseline result. An analysis of the named entities extracted revealed that the accuracy was very poor - worse than extraction from *bn-man* documents. This was primarily because the version of IdentiFinder (IdentiFinder-*man*) we used was by default trained on *nwt*.

To alleviate this problem we re-trained IdentiFinder using a simulated ASR corpus with named entities identified correctly. Since the amount of training data required was huge, we obtained the training data from the *bn-man* version of TDT3. We ran IdentiFinder-*man* on the *bn-man* version of TDT3 and tagged the named entities. We then removed punctuation and converted all the text to uppercase to simulate ASR to a limited degree. We re-trained IdentiFinder on this simulated ASR corpus and used it to tag named entities in only the *bn-asr* stories in TDT2. We retained the use of IdentiFinder-*man* for the *nwt* stories. The same three features were then extracted and we re-ran the classifiers. The results are shown in Table 4 in the column titled IdentiFinder-*asr*.

Kernel Type	TDT2 (*nwt+bn-asr*)	
Baseline System	0.640	
	IdentiFinder-*man*	IdentiFinder-*asr*
Linear Kernel	0.653	**0.608**
Poly. of deg. 1	0.654	0.608
Poly. of deg. 2	0.658	0.619
Poly. of deg. 3	0.659	0.616
Poly. of deg. 4	0.671	0.632
Poly. of deg. 5	0.676	0.640
Poly. of deg. 6	0.682	0.652
RBF with $\gamma = 1$	0.649	0.636
RBF with $\gamma = 5$	0.668	0.679

Table 4: The baseline system was the same used for the *nwt+bn-man* collection. We find that using a linear kernel for the procedure using IdentiFinder-*asr* to tag named entities results in statistically significant improvement.

8 Conclusions and Future Work

We have shown the applicability of machine learning classification techniques to solve the NED problem. Significant improvements were made over the baseline systems on all the corpora tested on. The features we engineered made extensive use of named entities, and reinforced the importance and need to effectively harness their utility to solve problems in TDT. NED requires not only detection and reporting of new events, but also suppression of stories that report old events. From the study of the distributions of scores assigned to stories by the baseline

and SVM model systems, we can see that we now do a better job of detecting old stories (reducing false alarms). Thus we believe that attacking the problem as "*old story detection*" might be a better and more fruitful approach. We have shown the effects of ASR output in the news stream, and demonstrated a procedure to alleviate the problem.

A classifier with RBF kernel with γ set to one exhibited the best performance. The reason for this superior performance over other kernels needs to be investigated. Engineering of better features is also a definite priority. In the future NED can also be extended to other interesting domains like scientific literature to detect the emerge of new topics and interests.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #IIS-9907018, and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

J. Allan, Hubert Jin, Martin Rajman, Charles Wayne, Daniel Gildea, Victor Lavrenko, Rose Hoberman, and David Caputo. 1999. Topic-based novelty detection. Technical report, Center for Language and Speech Processing, Johns Hopkins University. Summer Workshop Final Report.

J. Allan, Victor Lavrenko, and Hubert Jin. 2000. First story detection in tdt is hard. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 374–381. ACM Press.

J. Allan. 2002. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers.

Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231.

Thorsten Brants, Francine Chen, and Ayman Farahat. 2003. A system for new event detection. In *Proceedings of the 26th Annual International ACM SIGIR Conference*, pages 330–337, New York, NY, USA. ACM Press.

Ronald K. Braun and Ryan Kaneshiro. 2003. Exploiting topic pragmatics for new event detection in tdt-2004. Technical report, National Institute of Standards and Technology. Topic Detection and Tracking Workshop.

Ronald K. Braun and Ryan Kaneshiro. 2004. Exploiting topic pragmatics for new event detection in tdt-2004. Technical report, National Institute of Standards and Technology. Topic Detection and Tracking Workshop.

Christopher J. C. Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

James P. Callan, W. Bruce Croft, and Stephen M. Harding. 1992. The INQUERY retrieval system. In *Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications*, pages 78–83.

Ayman Farahat, Francine Chen, and Thorsten Brants. 2003. Optimizing story link detection is not equivalent to optimizing new event detection. In *ACL*, pages 232–239.

Thorsten Joachims. 1999. *Making large-scale support vector machine learning practical*. MIT Press, Cambridge, MA, USA.

Giridhar Kumaran and J. Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the 27th Annual International ACM SIGIR Conference*, pages 297–304, New York, NY, USA. ACM Press.

Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. 2002. Applying semantic classes in event detection and tracking. In *Proceedings of International Conference on Natural Language Processing (ICON 2002)*, pages 175–183.

Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. 2004. Simple semantics in topic detection and tracking. *Information Retrieval*, 7(3–4):347–368.

Dunja Mladeni, Janez Brank, Marko Grobelnik, and Natasa Milic-Frayling. 2004. Feature selection using linear classifier weights: interaction with classification models. In *Proceedings of the 27th Annual International ACM SIGIR Conference*, pages 234–241, New York, NY, USA. ACM Press.

R. Papka and J. Allan. 1998. On-line new event detection using single pass clustering. Technical Report UM-CS-1998-021.

Nicola Stokes and Joe Carthy. 2001. Combining semantic and syntactic document classifiers to improve first story detection. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, pages 424–425, New York, NY, USA. ACM Press.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc.

Yiming Yang, Tom Pierce, and Jaime Carbonell. 1998. A study of retrospective and on-line event detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference*, pages 28–36, New York, NY, USA. ACM Press.

Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proceedings of the 8th ACM SIGKDD International Conference*, pages 688–693. ACM Press.

Jian Zhang, Zoubin Ghahramani, and Yiming Yang. 2005. A probabilistic model for online document clustering with application to novelty detection. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1617–1624. MIT Press, Cambridge, MA.