Text Classification and Named Entities for New Event Detection

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

ABSTRACT

New Event Detection is a challenging task that still offers scope for great improvement after years of effort. In this paper we show how performance on New Event Detection (NED) can be improved by the use of text classification techniques as well as by using named entities in a new way. We explore modifications to the document representation in a vector space-based NED system. We also show that addressing named entities preferentially is useful only in certain situations. A combination of all the above results in a multi-stage NED system that performs much better than baseline single-stage NED systems.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support*

General Terms

Algorithms,Performance,Experimentation

Keywords

Topic Detection and Tracking, New Event Detection, Named Entities, Text Classification

1. INTRODUCTION

New Event Detection (NED) is one of the tasks in the Topic Detection and Tracking (TDT) program. The TDT program seeks to develop technologies that search, organize and structure multilingual news-oriented textual materials from a variety of broadcast news media. NED is concerned with developing systems that can detect the first story on a *topic* of interest, where a *topic* is defined as "a seminal event or activity, along with directly related events and activities"[4]. An example of a topic could be the sinking of an oil tanker. The first story on the topic would be the article that first reports the sinking of the tanker itself. Other stories on the same topic would be those discussing the environmental damage, the salvaging efforts, the commercial impact and so on. A good NED system would be one that correctly identifies the article that first reports the sinking as the first story.

NED has practical applications in domains like financial markets, news analyses, intelligence gathering etc. where useful information is usually buried in a mass of data that grows rapidly with time. A NED system could also be used as part of a larger system that gathers and organizes information.

A natural way to go about detecting new stories is to compare the story on hand with all the stories that have been seen in the past. This is done by measuring the degree of overlap between stories, usually in the form of the cosine similarity metric. While a number of attempts have been made to use alternate techniques that range from language modeling to machine learning, the vector space model has achieved the best results to date. However, as more ambitious performance goals are targeted in the future, the limitations of the vector space approach have become apparent.

Failure analysis of a simple vector space system approach suggests the development of not only new and better similarity metrics, but also better document representations. An example of the push in the direction of the former is the Hellinger similarity metric[8], and of the latter is the composite document representation[13]. Our approach is to instead stay within the existing cosine similarity metric and vector space model, and modify the way they are used.

We start off this paper by summarizing the previous work in NED in Section 2. We then briefly describe the evaluation methodology for NED in Section 3. After touching upon the basic vector space model in Section 4 we move on to explaining the modifications we made to the basic model that provided us with a better tool for NED in Section 5. Section 6 describes our experimental setup. We finally wrap up with the experimental results in Section 7 and the conclusions and future work in Section 8 .

2. PREVIOUS RESEARCH

On-line NED [11] was the focus of a paper by Papka et al.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.

When a new document was encountered, it was processed immediately to extract features and build up a query representation of the document's content. 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. The threshold model developed for the task incorporated time information, the intuition being that documents that are widely spaced apart in time are more likely to deal with new (different) events. Performance-wise, it was found that increasing the number of features used to build the queries results in improved performance, with an unacceptable increase in running time of the system. At low feature dimensionality, misses were attributed to the inability of the feature extraction process to weight event-level features more heavily than more general topic-level features. Even at higher feature dimensionalities misses occurred, which were finally ascribed to the poor weight assignment strategy for query features.

A paper by Stokes et al. [13] presented an approach to NED that utilized a combination of evidence derived from two distinct representations of a document's content. 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 - again as a vector of terms. This method automatically disambiguated terms. The two vectors were combined in a linear fashion, and the usual cluster-document similarity-threshold approach was followed. It was concluded that a marginal increase in effectiveness could be achieved when lexical chain representations are used in conjunction with the free text representation, i.e. the data fusion model was marginally better.

Allan et al.[6] argued that NED approaches that relied on exploiting existing news tracking technology would invariably exhibit poor performance. Systems that used tracking technology for NED followed the mantra - every time a new topic was found and tracked by a topic tracking system, it was equivalent to finding a new event. Thus, the NED system was only as good as the tracking system it was built on. 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.

A summer workshop[5] on topic-based novelty detection held at Johns Hopkins University extensively studied the NED problem. Similarity metrics, effect of named entities, pre-precessing of data, and language and hidden markov models were explored. Combinations of NED systems were also discussed.

In the topic-conditioned novelty detection[14] approach, documents were classified into broad topics and NED was performed within these categories. Additionally, named entities were re-weighted relative to the normal words for each topic, and a stop list was created for each topic. However the experiments were done on a corpus different from the TDT corpus.

The most recent published work on NED[8] extended a basic incremental TF-IDF model to include source-specific models, similarity score normalization techniques, and segmentation of documents. Good improvements on TDT benchmarks were shown.

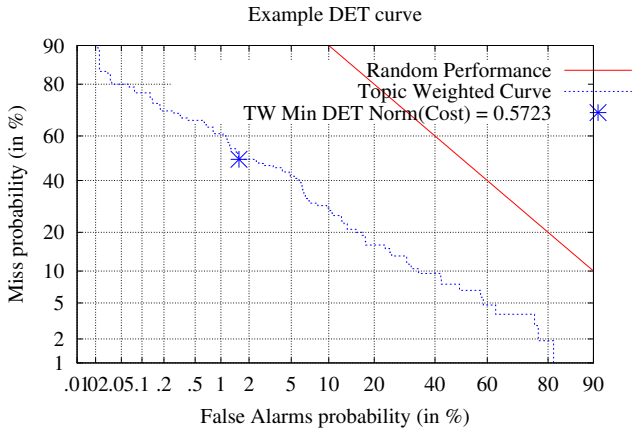

Figure 1: An example DET curve. Each point on the curve is the due to the misses and false alarms at a particular threshold. A sweep across all the possible thresholds from 0 to 1 generates the points in the DET curve.

3. NED EVALUATION

Every story upon arrival is assigned a confidence score between 0 and 1 by the NED algorithm. This assignment of scores is done either immediately upon arrival or after a fixed look-ahead window of stories. A (cosine similarity) score of 0 translates to complete confidence that the story is new, and a score of 1 implies the greatest confidence that the story is old. To evaluate performance, the scores are sorted, and a threshold sweep is performed. All stories with scores above the threshold are declared old, while those below it as considered new. At each threshold value, the misses and false alarms are identified, and a cost is calculated as a linear function of their number. The threshold that results in the least cost is selected as the optimum one. Different NED systems are compared based on their minimum cost. The Detection Error Tradeoff (DET) curve is a convenient way to represent the miss and false alarm values at each threshold, and to compare the performance of NED algorithms at different regions of the graph, i.e. at different threshold values.

4. BASIC MODEL

We used the cosine similarity (Equation 1) metric to judge the similarity of a document with those seen in the past.

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

where

$$weight(w, d) = tf * idf$$

$$tf = \log(termfrequecy + 1.0)$$

$$idf = \log((docCount + 1)/(documentfreq + 0.5))$$

4.1 Preprocessing

We used version 1.9 of the open source Lemur system[1] 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 [9], and the K-stem stemming algorithm [10] implementation provided as part of Lemur.

Incremental TF-IDF weighting[8] was used, and document similarity normalization [8] was performed before a final score was assigned to a story.

5. MODIFIED MODEL

While the cosine similarity metric has proved to be the most successful metric to date for NED, it has its own shortcomings. This is primarily because it is used as a substitute for a human notion or *newness* and *oldness*, something that is hard to capture and is based on individual perception, as well as what level of a hierarchy of events is of interest. Attempts to tailor the cosine metric to perform as desired have focused on re-weighting terms and building different document models - all without much success. Our approach to improve NED was different. We acknowledged the fact that the basic cosine similarity metric can make mistakes, and decided to buttress our confidence in the score by looking into other parameters like the category (*finance*,*accidents*, etc.), the overlap of named entities, the overlap of non-named entities, and so on. We developed simple rules that reflect the questions that a human being would ask before deciding if a story were *new* or *old*. The following sections describe our observations and conclusions.

5.1 Modification to document model

5.1.1 Motivation

A look at the contribution of individual terms to the overall similarity scores between documents revealed that the assignment of weights to them left much to be desired. For example, while comparing two stories on different topics in health care, terms like *drugs, cost, coverage, plan, prescription* etc. contributed most to the overall similarity score.

While in a way this was a good thing to happen, it only helped in identifying that the two stories were on a similar issue. The two stories actually involved completely different locations and individuals. While tf-idf weighting should have suppressed the contributions of the most similar terms and revealed the difference in the stories in the form of an overall low similarity score, it apparently failed to do so to the degree required.

We believed that the problem of weight assignment to terms could be resolved by first placing stories into broad categories, and then computing term weights using the statistics within those categories. The next step was to determine what those broad categories would be. Our goal was to improve the performance on the benchmark corpus provided by the Linguistic Data Consortium[1]. Since we were in essence trying to capture the LDC's methodology for determining topics and events (and hence new/old stories), a natural choice for the categories were the thirteen topic types specified by the LDC. These topic types come along with special rules of interpretation (ROI)[2] for each one of them. The ROI are used to help annotators achieve consistency in their judgements of which topic a story belongs to. Some of the LDC categories are listed below.

1. Elections
2. Scandals/Hearings
3. Legal/Criminal Cases
4. Natural Disasters
5. Accidents
6. Acts of Violence or War

5.1.2 Classification according to LDC topics

To classify stories, we used BoosTexter[12]. BoosTexter is a general purpose machine-learning program, based on boosting, for building a classifier from text and/or attribute-value data. Given training data, BoosTexter creates a series of simple rules that are used to build the classifier. In our experiments, the terms in a document, weighted by their frequency of occurrence in it, were used as features.

For our classification experiments, we trained on TDT-2 judged documents and tested on TDT-3 documents. For our classification of TDT-4 we trained on judged documents from both TDT-2 and TDT-3.

5.1.3 Using the classifications

Using the classifications to group together stories belonging to the same LDC category and computing the term-weights according to the statistics of each group as in Yang et al. [14] was a failure. The reason for this is that it is possible for a story to be on multiple topics at the same time. Mistakes made by the classifier can also throw such a system off-track. We believe that such topic-conditioned weighting and comparison would work well only in situations where the granularity of the categories was much finer. Hence constraining the stories into water-tight categories proved detrimental in our case. To address this problem, and at the same time down-weight frequent terms, we developed stop lists for each category from training stories in TDT-2. The stop lists contained the top 300 terms that occurred most frequently in each LDC category in the training set. Thus the old document vector representation for each story was replaced by a new vector that had these stop words removed. We then relaxed the constraint on comparison only within a category and determined the score of a story after comparing with all the stories in the past irrespective of the category they belonged to.

5.2 Modification to Similarity Metric

5.2.1 Motivation

False alarms are caused when an old story is assigned a low score. Misses, which are more costly than false alarms, are caused when a new story is assigned a high score. An in-depth look at misses revealed that it was important to isolate the named entities and treat them preferentially. This is nothing new - Allan et al. in [5] boosted the weights of the named entities so that they contributed substantially to the overall similarity score. Yang et al. in [14] also utilized a

[1] `http://www.cs.cmu.edu/~lemur`

statistical basis to re-weight named entities. However, both these approaches had limited utility, as TF-IDF weighting already provided a subtle differential in the scores of named entities and non-named entities. This difference in weights is not of the degree and uniformity required for NED, and is not increased by both the methods mentioned to have a significant impact on NED.

To understand the utility of named entities we present two examples. Consider the two stories from the *science* category given below.

Story 1: New Story

*In Argentina, scientists have discovered a huge **dinosaur** nesting site. Their findings, outlined in the journal "nature," describe thousands of six-inch fossil **eggs** along with embryo remains. Scientists say the **eggs** were laid 70 to 90 million years ago. They believe the large number of **eggs** suggests **dinosaurs** converged at the site to lay them. Researchers believe a catastrophe, perhaps a flood, kept the **eggs** from hatching.*

Story 2 : Closest Match

*A research team in Pasadena, California, is getting help from a local hospital in its investigation into prehistoric life on earth. The team is using the hospital's ct scanner to X-Ray fossilized **dinosaur eggs** found in China. The researchers were surprised to find the undisturbed fetus of a baby **dinosaur** in one of the **eggs**. Visually, when you cut open an **egg**, you don't see any of this. Maybe a little bit of foetal bone representated. And you see minerals that are almost similar to the minerals outside of the **egg**, which tells you that to the unaided eye you've got nothing.*

While both these stories talk about dinosaurs and their eggs, they obviously are about different events. However, the presence of high IDF terms like *dinosaur* and *eggs* led to **Story 1** being assigned a high score. This could have been avoided if the *location* named entities had been accorded greater attention.

Story 3 : New Story

*Thousands of democracy demonstrators rallied in a Tehran park Sunday to protest the rejection of scores of candidates for the election of a powerful assembly. The Council of Guardians, which is dominated by political hard-liners, has rejected 214 candidates out of 400 who wanted to run in Friday's election to the 86-member Assembly of Experts. Many of the disqualified candidates were supporters of **President** Mohammad **Khatami**, a political moderate. Twelve candidates withdrew their names in protest against the screening. The move by the 12-member Council of Guardians is seen as part of the struggle between conservatives and reformists in **Iran**. **Khatami** went on national television late Saturday to criticize the selection of the candidates for the assembly, which has the power to appoint or dismiss **Iran's** supreme leader.*

Story 4 : Closest Match

***Iranian** hardliners are moving forward with efforts to silence publications supportive of moderate **president** kasra naji reports. The authorities in **Iran** this week closed down the biggest circulation newspaper in the country, tous on charges of insulting the late Ayatollah Khomeini. This is part of a new clamp-down on the press by hard-line opponents of moderate **President Khatami**, who's been advocating greater civil liberties. Only a couple of months ago, the newspaper, jam'eh, had changed its title to tous to escape an earlier closure order. With its daring and often libertarian attitude, the new newspaper had become the most popular in a matter of months, with a circulation of nearly 500,000. But the newspaper had powerful critics, too.*

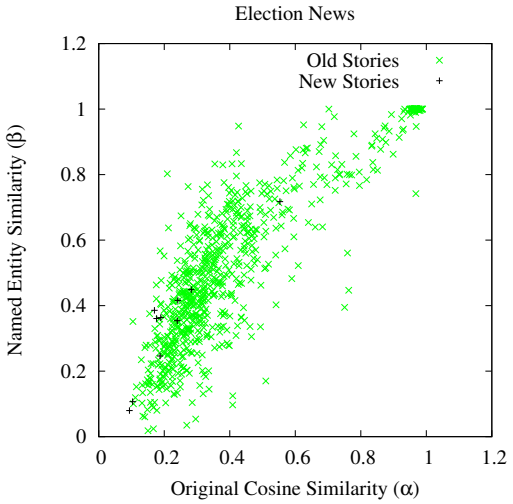

Figure 2: Elections (α versus β).

The two stories are about the *location* Iran and mention the country's President. Though they are on different topics, the named entities caused a high similarity score for **Story 3**. In this situation, it would have been apt to ignore the named entities and instead just compare the distributions of the non-named entity terms.

Thus we can observe that named entities are a double-edged sword, and deciding when to use them can be tricky.

5.2.2 Multiple document representations

Our solution to this problem is to first create three vector representations α, β, and γ for each document . The first representation α consists of all terms in the document (with stop words removed), the second β consists of only the named entities, and the third γ comprises of only the non-named entity terms. Named entities were identified using BBN Identifinder [7]. We considered only the *Event*, *GPE*, *Language*, *Location*, *Nationality*, *Organization*, *Person*, *Cardinal*, *Ordinal*, *Date*, and *Time* named entities to create β (and hence γ). When two stories were compared, each document representation for a document was compared only with the corresponding representation for the other document.

Once the most similar document to the story under consideration based on α similarity was identified, the similarity using the β and γ representations were also calculated.

5.2.3 Using α, β, and γ

To gain insight into the relationship between the α, β, and γ scores, graphs of α versus β and α versus γ were plotted for the all the stories in each category. Some of them are shown as Figures 2 to 7. It should be kept in mind that some of the stories might be misclassified, and thus our discussions and conclusions are particular to the categories and text classifier we have selected.

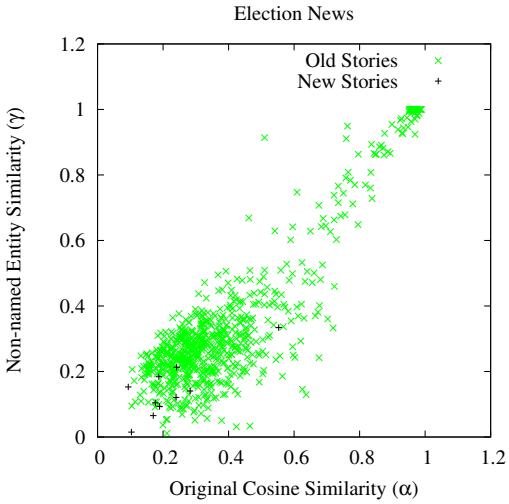

Figure 3: Elections (α versus γ).

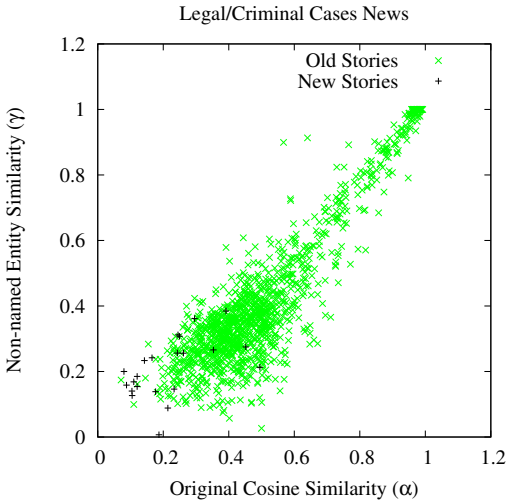

Figure 5: Legal/Criminal Cases (α versus γ).

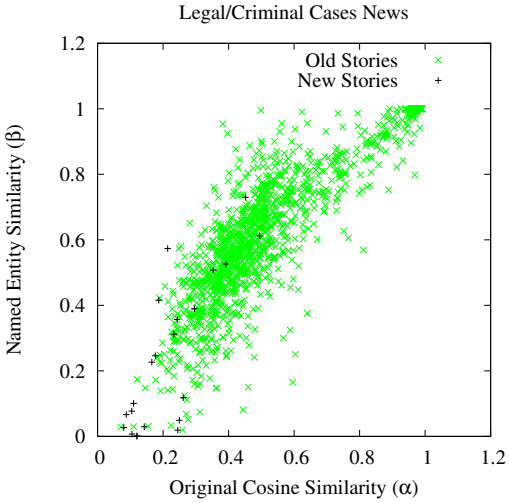

Figure 4: Legal/Criminal Cases (α versus β).

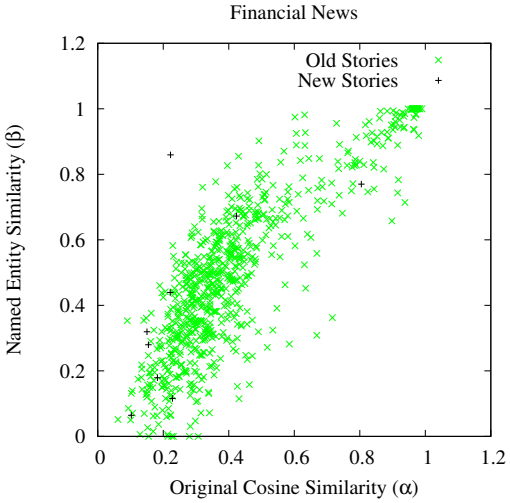

Figure 6: Financial News (α versus β).

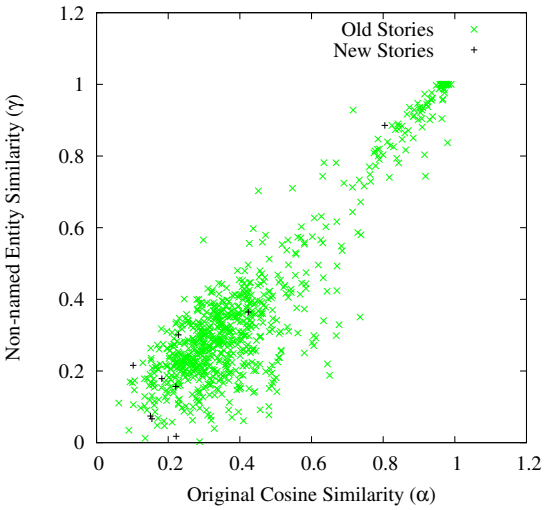

Figure 7: Financial News (α versus γ).

Consider Figure 2 and Figure 3 for the category *Elections*. We see that a majority of the new stories (shown as black '+') have a γ score less than 0.2, while the β scores are spread out among the slew of old stories (shown as light colored 'X'). This leads to the intuition that we could use the γ score as a way to confirm the status of the story (old/new) as suggested by the α score. Thus, we see that *on an average* it is not named entities that matter more in finally detecting new *Election* stories, but the rest of the terms.

Figures 4 and 5 for the category *Legal/Criminal Cases* tell a different story. Here we observe that a β and γ score less than 0.4 characterize most new stories. However there are more old stories with γ score less than 0.4 than there are with a similar β score. Hence it is more useful to use the β score as an additional metric than the γ score, i.e. considering named entities is a win over ignoring them.

Unfortunately, making such clear cut decisions for all categories is not possible. This is illustrated by Figures 6 and 7 for the *Financial News* category. We see that it is impossible to select between the β and γ scores. For such categories, we currently proceed with the α scores as the only metric.

Table 4 summarizes our findings.

6. EXPERIMENTAL SETUP

We used the TDT3 and TDT4 datasets for our experiments. TDT3 contains news stories from October to December 1998. It contains around 35,000 stories from sources like CNN, New York Times, ABC, Voice of America etc. Only the English stories in the collection were considered. TDT4 consists of approximately 28,500 stories from the period October 2000 to January 2001, and from the same sources. TDT3 contains 115 topics (and hence a similar number of new events), while TDT4 contains 80 topics.

The judged stories in TDT3 were used as a training set to determine the simple thresholding rules involving the α, β, and γ scores for each category. Topic judgements for TDT4 were unavailable at the time of writing this paper.

7. EXPERIMENTAL RESULTS

We ran our system both on TDT3 as well as TDT4 data sets.

Table 1 provides a guide to the three NED systems that we used, while Table 2 compares the performance of our systems with those reported by other systems on TDT3([3],[8]). We can observe that the category-specific stop lists and selective use of named entities succeeds in lowering the min-cost to a very competitive value. Figure 8 shows the DET curves for SYSTEM 1, SYSTEM 2, and SYSTEM 3.

Though the precise details of topics and stories in the TDT4 corpus are still unavailable, the fact that all the systems that were run on it (as part of this year's official TDT evaluations) performed badly lends credence to the belief that the TDT4 corpus is more challenging than the TDT3 corpus. A cursory look at Table 3 seems to reveal that our extensions to the basic model actually hurt performance. In this situation, it is more apt to look at the DET curves in Figure 9 instead. We see that at the regions of high accuracy[2] there is a clear and consistent drop in the number of misses. At the lower regions of the graph, the curves are virtually indistinguishable.

From the TDT4 DET curves, it also appears that our second modification involving named entities hardly had any impact. However, as we had discussed in Section 5.2.2, the utility and impact of named entities depends on the topics present in the corpus. Without any information about the TDT4 corpus at this time, we can best hazard a guess that most of the stories were on topics that are not conducive to named entity-based modifications.

Table 1: NED systems fielded. The features are additive.

System	Features
SYSTEM 1 (baseline)	Incremental TF-IDF weighting, stopping, stemming, remove short stories, similarity normalization [8]
SYSTEM 2	Category stop lists
SYSTEM 3	α, β, and γ scores processing

Table 2: Comparison of NED results on the TDT3 corpus.

System	Topic-weighted Minimum Cost
Other systems performance range	0.5783 - 0.9785
SYSTEM 1	0.5723
SYSTEM 2	0.5578
SYSTEM 3	0.5229

8. CONCLUSIONS AND FUTURE WORK

[2]A system that reports all stories as *new* will obviously get all the new stories right, but that defeats the purpose of new event detection. We need systems that have low false alarm rates too.

Table 4: This table provides a guide to deciding which score to use, β or γ, depending on the category

	Use β scores	Use γ scores	Undecided
News Category	Scandals/Hearings	Elections	Financial News
	Legal/Criminal Cases	Accidents	
	Natural Disasters	Violence and War	
	Science/Discovery News	New Laws	
	Celebrity/Human Interest News	Sports News	
	Miscellaneous News	Political and Diplomatic Meetings	

Table 3: Comparison of NED results on the TDT4 corpus.

System	Topic-weighted Minimum Cost
Other systems performance range	0.5312 to 0.6346
SYSTEM 1	0.5144
SYSTEM 2	0.5293
SYSTEM 3	0.5510

We have presented a new multi-stage system for performing NED. Extensions to the baseline vector space system were made in the form of story categorization and better use of named entities. Both these extensions contributed significantly to improvement in performance on the TDT3 and TDT4 corpora, with the latter accounting for a major fraction of the improvement. The improvements in performance are consistent across all the regions of the DET curve for TDT3 and in the high accuracy regions for TDT4. Our extensions to the basic model did not result in increased computational time. We have shown a way to harness the named entities in documents, and also illustrated their utility in different situations.

While we believe that our extensions are a promising direction for NED technology to proceed in, the simple rules we came up with for utilizing the named entities could do with more in-depth study and development. Also, attempts need to be made to accommodate stories that are classified into multiple categories, and develop stop lists for such situations. The creation of the stop lists too was ad-hoc - metrics like information gain could lead to better stop lists. We also believe that temporal information should be factored into any NED system.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval 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.

9. REFERENCES

[1] The linguistic data consortium, http://www.ldc.upenn.edu/.

[2] Rules of interpretation, http://www.ldc.upenn.edu/projects/tdt4/annotation/.

[3] Tdt 2001 evaluations, http://www.nist.gov/speech/tests/tdt/tdt2001/index.htm.

[4] In *Topic Detection and Tracking. Event-based Information Organization*. Kluwer Academic Publishers, 2002.

[5] J. Allan, H. Jin, M. Rajman, C. Wayne, G. D., L. V., R. Hoberman, and D. Caputo. Summer workshop final report. In *Center for Language and Speech Processing*, 1999.

[6] J. Allan, V. Lavrenko, and H. Jin. First story detection in tdt is hard. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 374–381, 2000.

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

[8] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *Proceedings of ACM SIGIR2003*, pages 330–337, 2003.

[9] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications*, pages 78–83, 1992.

[10] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of ACM SIGIR93*, pages 61–81, 1998.

[11] R. Papka and J. Allan. On-line new event detection using single pass clustering TITLE2:. Technical Report UM-CS-1998-021, , 1998.

[12] R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. In *Machine Learning 39(2/3):1*, pages 35–168. Kluwer Academic Publishers, 2000.

[13] N. Stokes and J. Carthy. First story detection using a composite document representation. In *Proceedings of Human Language Technology Conference*, 2001.

[14] Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *Proceedings of ACM SIGKDD03*.

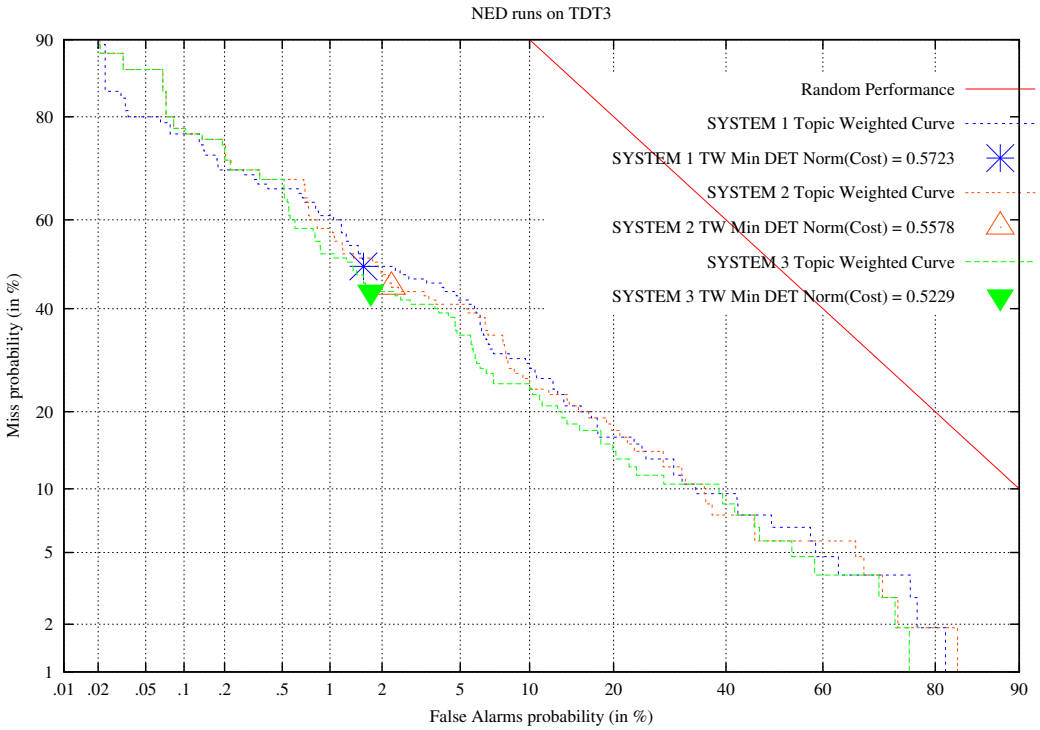

Figure 8: DET curves for SYSTEMs 1,2, and 3 on the TDT3 corpus

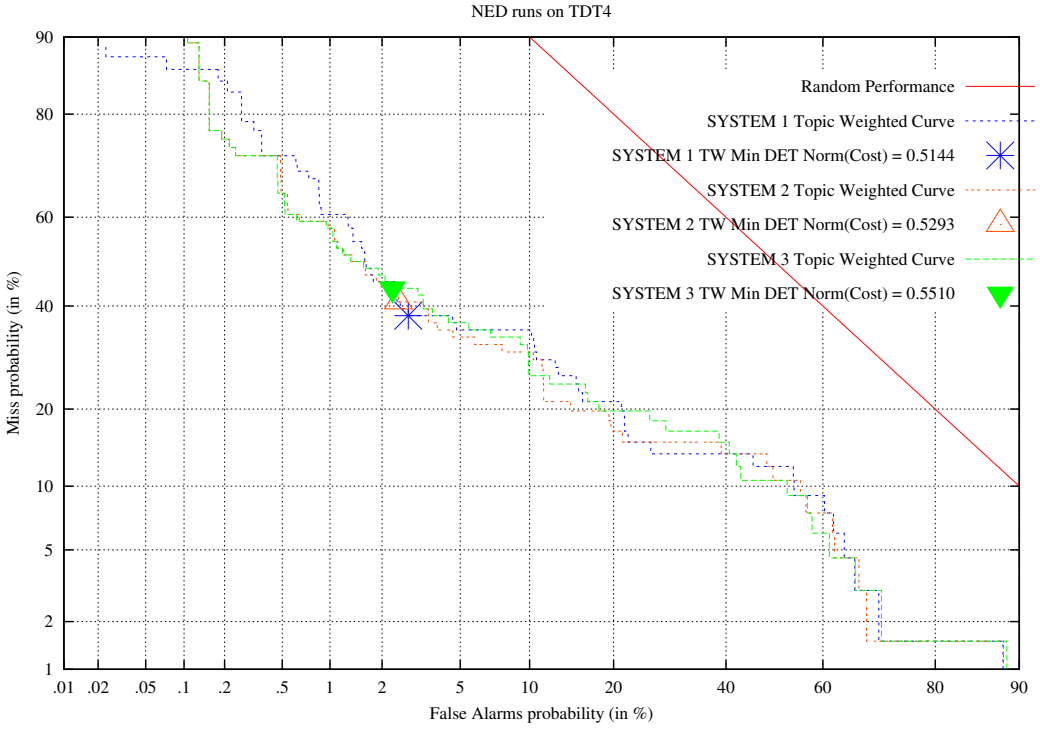

Figure 9: DET curves for SYSTEMs 1,2, and 3 on the TDT4 corpus