

# Representing Documents with Named Entities for Story Link Detection (SLD)

Chirag Shah, W. Bruce Croft, David Jensen  
Department of Computer Science  
University of Massachusetts  
Amherst, MA 01003  
{chirag,croft,jensen}@cs.umass.edu

## ABSTRACT

Several information organization, access, and filtering systems can benefit from different kind of document representations than those used in traditional Information Retrieval (IR). Topic Detection and Tracking (TDT) is an example of such an application. In this paper we demonstrate that named entities serve as better choices of units for document representation over *all* words. In order to test this hypothesis we study the effect of words-based and entity-based representations on Story Link Detection (SLD) - a core task in TDT research. The experiments on TDT corpora show that entity-based representations give significant improvements for SLD. We also propose a mechanism to expand the set of named entities used for document representation, which enhances the performance in some cases. We then take a step further and analyze the limitations of using only named entities for the document representation. Our studies and experiments indicate that adding additional topical terms can help in addressing such limitations.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Information filtering*; H.3.4 [Information Storage and Retrieval]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

## General Terms

Measurement, Performance, Experimentation

## Keywords

Topic Detection and Tracking, Story Link Detection, Document representation, Named Entities

## 1. INTRODUCTION

Document representation is one of the most common and crucial stages of an information organization and access system. Several methods and models of document representation have been proposed based on the target application. Examples include word-based [7], language models [6], and graph-based [3]. Some of them are general enough to be applicable to almost any IR-based application. However, some tasks demand a different approach to document representation. Topic Detection and Tracking (TDT) [1] is one such domain.

Copyright is held by the author/owner(s).  
CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.  
ACM 1-59593-433-2/06/0011.

Loosely speaking, since TDT deals with information that is tightly related to certain event happening at certain places and/or with certain people or organization, it makes sense to explicitly consider such terms, also called named entities, while representing documents. In this paper we argue that using named entities - the names of people, places, organizations, etc. - is more suitable than using all words for document representation in TDT. In order to demonstrate this, we use some examples and extensive experimentation on TDT corpora with the Story Link Detection (SLD) task as the focus. SLD deals with finding if two news stories are on the same topic or not, which is at the core of almost all the tasks in TDT. Our experiments and analysis demonstrate that a named entity based representation brings down the detection cost by a significant amount, and thus improves the performance of a SLD system.

## 2. EXPERIMENTAL SETUP

Our experiments reported in this paper were carried out on TDT3 and TDT4 corpora available from NIST.<sup>1</sup> One set of experiments were performed splitting the TDT3 corpus in two almost equal parts for training and testing, whereas for another set of experiments we used the entire TDT3 corpus for training and the TDT4 corpus for testing. The evaluation of the SLD system is done using the measure given by NIST:

$$C_{Det} = (C_{Miss} * P_{Miss} * P_{Target} + C_{Fa} * P_{Fa} * (1 - P_{Target})) \quad (1)$$

where  $P_{Miss}$  = No. of missed detection/No. of targets,  $P_{Fa}$  = No. of false alarms/No. of non-targets,  $C_{miss}$  and  $C_{Fa}$  are the costs of a missed detection and a false alarm respectively, and are pre-specified,  $P_{Target}$  is the *a priori* probability of finding a target.<sup>2</sup> This detection cost is then normalized as given below.

$$(C_{Det})_{Norm} = \frac{C_{Det}}{MIN(C_{Miss} * P_{Target}, C_{Fa} * (1 - P_{Target}))} \quad (2)$$

## 3. PROPOSED METHODS

This section describes our baseline as well as a set of methods that make use of named entities in various ways.

<sup>1</sup><http://www.nist.gov>

<sup>2</sup>TDT evaluation script predefines  $P_{Target}$  to be 0.02, however, we calculate its value from the corpus that we are experimenting on. Due to this change, the scores that we obtain would be different than the ones obtained by the scripts that TDT provides. However, they reflect merely the change of scale and not the relative performances of various systems.

### 3.1 TFIDF on all the words - Baseline

TFIDF based representation of documents is widely used in many IR applications [7]. We construct vectors for each document using TFIDF weighting and then find the cosine between two vectors. This score becomes the similarity measurement for the given documents. If this score is above the threshold, then the given pair of stories are said to be on the same topic. Later we shall see how to derive this threshold from the training data.

### 3.2 TFIDF on named entities

As we discussed in the previous section, making use of named entities in case of a task such as SLD makes more sense than using all the words. One simple approach of putting this in effect is to use traditional TFIDF technique on named entities. We used BBN's Identifinder [2] to extract the named entities and threw away the rest of the words from the documents. Thus, the extracted named entities became the representation of the documents. We executed the similar process on them as a regular vector space model with TFIDF term weighting and found the corresponding vectors. These vectors were compared with the cosine similarity measure to find the similarity between the documents.

### 3.3 Unweighted expansion

We realized that in the above technique we were throwing away a lot of information. What remained was indeed the information with high content, but in some cases it may not be sufficient to match two documents. In such cases when the original named entities are not enough, we may want to introduce some more information. We do this by augmenting the original documents, represented as a collection of named entities, with *related* named entities. In order to do this, we created a graph  $G$  with named entities from the training corpus as the nodes. We connected two nodes with an edge if the named entities that they represent occur together at least in one document. Now, while comparing two documents, we extend the original set of named entities by adding their immediate neighbors from  $G$ . After adding such *related* named entities, we use the same procedure as the above method to perform matching.

The basic idea of this technique is to expand the original set of named entities that we extracted with some related entities. This approach may sound similar to various query expansion techniques in IR like Relevance Model [5], however, it is to be noted that we are not performing any retrieval here. The expansion is done over the named entities and based on past knowledge about their inter-connections.

### 3.4 Weighted expansion

In the above proposed method, we added the related named entities with full confidence giving them equal importance as the original set of named entities. We also did not distinguish among the newly added named entities for their relative importance to the documents in which they were added. We address this issue by finding the strength of the connections among the named entities. In other words, instead of considering an unweighted graph, we now try to find weights on each edge. Among many possibilities of quantifying the strength of the connection between two entities, we find mutual information more appropriate for our task. The mutual information between two entities  $e_i$  and  $e_j$  was calculated using

$$MI(e_i, e_j) = \log_2 \left( \frac{p(e_i, e_j)}{p(e_i) \cdot p(e_j)} \right) \quad (3)$$

where  $p(e_i)$  and  $p(e_j)$  respectively are the probabilities of entities  $e_i$  and  $e_j$  occurring in the corpus and  $p(e_i, e_j)$  is the probability of entities  $e_i$  and  $e_j$  occurring together in the same document. When

we add an entity  $e_j$  to the existing set of entities because it is related to entity  $e_i$  (already present in the document), we add it with the confidence  $MI(e_i, e_j)$ . The rest of the process of matching two documents is the same as that of the above method.

## 4. EXPERIMENTS AND RESULTS

The results from our experiments on the two different data-sets are summarized in Table 1. The results obtained by techniques 2 and 4 were found statistically significant with respect to the baseline using two-tailed paired  $t$ -test as well as McNemar's statistical significance test [4] (a non-parametric test, which is more appropriate for this task). Since lower cost is better, we can say that techniques 2 and 4 demonstrated significant improvements over the baseline.

**Table 1: Normalized detection cost for two sets of experiments: TDT3 split for training and testing, and entire TDT3 for training and TDT4 for testing. Lower cost is better.**

#	Technique	TDT3-TDT3	TDT3-TDT4
1	TFIDF on words (Baseline)	0.7964	0.8764
2	TFIDF on entities	0.7374	0.4860
3	Unweighted expansion	0.8723	0.8854
4	Weighted expansion	0.4937	0.7759

## 5. CONCLUSION AND FURTHER WORK

In the work reported here, we identified the uniqueness of Topic Detection and Tracking (TDT) tasks. We showed that using named entities for representation in case of TDT is a better idea than using a word-based technique. In order to demonstrate this, we worked with Story Link Detection (SLD), which is at the core of the TDT. Our experiments on various TDT corpora revealed that our proposed representation is indeed very effective. In some additional experiments not reported here, we also tried to test the limits of named entities. Our experiments and analysis showed that named entities lack enough information in some cases. We discovered that augmenting the named entities with some additional *topical terms* helps in addressing this limitation.

## 6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0527159. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

## 7. REFERENCES

- [1] James Allan, editor. *Topic Detection and Tracking*. Kluwer Academic Publishers, 2002.
- [2] D. Bikel, R. Schwartz, and R. Weischedel. An algorithm that learns what's in a name. *Machine Learning - Special Issue on NL Learning*, 34(1):1-3, 1999.
- [3] K. Collins-Thompson and J. Callan. Query expansion using random walk models. In *CIKM*, 2005.
- [4] L. Gillick and S. Cox. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proceedings of IEEE's ICASSP 1989*, pages 532-535, 1989.
- [5] V. Lavrenko and W.B.Croft. Relevance-based language models. In *Proceedings of the 24th annual international ACM SIGIR conference*, pages 120-127, 2001.
- [6] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275-281, 1998.
- [7] Gerald Salton, editor. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, 1989.