# Combining Semantic and Syntactic Document Classifiers to improve First Story Detection.

Nicola Stokes, Joe Carthy,
Department of Computer Science,
University College Dublin,
Ireland.
{nicola.stokes,joe.carthy}@ucd.ie

## ABSTRACT

In this paper we describe a type of data fusion involving the combination of evidence derived from multiple document representations. Our aim is to investigate if a composite representation can improve the online detection of novel events in a stream of broadcast news stories. This classification process otherwise known as *first story detection* FSD (or in the Topic Detection and Tracking pilot study as online new event detection [1]), is one of three main classification tasks defined by the TDT initiative. Our composite document representation consists of a semantic representation (based on the *lexical chains* derived from a text) and a syntactic representation (using *proper nouns*). Using the TDT1 evaluation methodology, we evaluate a number of document representation combinations using these document classifiers.

## 1. A DOCUMENT REPRESENTATION STRATEGY USING LEXICAL CHAINS

The purpose of any document representation strategy in IR is two fold. Firstly the efficiency of the IR process (i.e. filtering, clustering etc.) improves greatly when a smaller dimensionality than full dimensionality is imposed on the word space of a document. Secondly, the effectiveness of the IR task may also improve as more pertinent features are retained to describe document content and 'noisy' features are removed. Using some sort of feature selection method is often an essential part of any IR process, where the most common approach used is based on the gathering of corpus statistics i.e. analyzing document content in terms of word distributions within the corpus. However in this paper we explore an alternative non-statistical approach to feature selection based on the identification of *lexical chains* using an online lexical taxonomy called WordNet [3].

When reading any text it is obvious that it is not merely made up of a set of unrelated sentences, but that these sentences are in fact connected to each other in one of two ways: *cohesion* and *coherence*. Morris and Hirst [2] describe textual cohesion as the way in which text 'tends to hang together'. They found that the cohesive structure of a text can be explored and represented by creating sequences of semantically related words called lexical chains. For example in a document concerning 'airplanes' a typical chain might consist of the following words {*plane, airplane, pilot, cockpit, airhostess, wing, engine*}, where each word in the chain is directly or indirectly related to another word by a semantic relationship such as *holonymy*, *hyponymy, meronymy* and *hypernymy*. Our feature selection criterion is based on the identification of these topics or chains in the text. We assume that words, which fail to be chained, do not take part in the overall cohesive structure of the text and hence are not essential topic descriptors. Another advantages of using chain words as document classifiers is that they address two linguistic problems associated with traditional syntactic representations i.e. *synonymy* and *polysemy*. Firstly, WordNet allows us to represent synonymous words like {*car, automobile, motorcar*} in terms of a single unique identifier called a 'synset number'. Secondly, when a polysemous word i.e. *bank* is added to a lexical chain its correct sense within the context of the document is discovered. In other words during the chain formation process words are implicitly disambiguated as follows. Firstly each term contained in a particular document is dealt with in order of occurrence. Then each word is added to an existing lexical chain if a semantic path (of predefined maximum length) between the two words exists in WordNet, otherwise this word becomes the seed of a new chain. A stronger criterion than simple semantic association is imposed on the addition of a term to a chain, where terms must be added to the most recently updated (semantically related) chain. This favors the creation of lexical chains containing words that are in close proximity within the text, prompting the correct disambiguation of a word based on the context in which it was used.

So the first element of our combined document representation is a chain word representation made up of WordNet synset numbers which map all syntactic forms of a concept to a single number. The necessity of the second element of our combined document representation becomes apparent when we consider the descriptive power of proper nouns when representing events in a news story domain, and the absence of proper nouns in our chain word classifier. This deficiency is due to the failure of WordNet to make semantic associations between proper nouns and other word types in its taxonomy. Hence to address this deficiency in our chain word representation we identify proper nouns using a simple heuristic based on capitalization and use these words as our syntactic representation of our combined document representation.

## 2. DETECTION USING TWO CLASSIFIERS

Online Detection or First Story Detection (FSD) is in essence a classification problem where documents arriving in chronological order on the input stream are tagged with a 'YES' flag if they discuss a previously unseen news event, or a 'NO' flag when they discuss an old news topic.

*SIGIR '01*, September 9-12, 2001, New Orleans, Louisiana, USA.

However unlike detection in a retrospective environment a story must be identified as novel before subsequent stories can be considered. The single-pass clustering algorithm bases its clustering methodology on the same assumption, the general structure of which is summarised as follows. A more detailed analysis of the comparison and thresholding strategies defined in this algorithm is given in [3].

1. Convert the current document into a weighted chain word vector and a weighted proper noun vector.
2. The first document on the input stream will become the first cluster.
3. All subsequent incoming documents are compared with all previously created clusters up to the current point in time. A comparison strategy is used here to determine the extent of the similarity between a document and a cluster. In our IR model we use sub-vectors to describe our two distinct document representations. This involves calculating the closeness or similarity between the chain word vectors and proper noun vectors for each document/cluster comparison using the standard cosine similarity measure (used in this variation of the vector space model to compute the cosine of the angle between two weighted vectors). The data fusion element of this experiment involves the combination of similarity measures from two distinct representations of document content in a single cluster run i.e. *k* equals 2 in equation (1). So the overall similarity between a document *D* and a cluster *C* is a linear combination of the similarities for each sub-vector formally defined as:

$$Sim\ (D, C) = \sum_{j=1}^{k} w_j \cdot Sim\ (D_j, C_j) \qquad (1)$$

   where *Sim(X, Y)* is the cosine similarity measure for two vectors *X* and *Y*, and *w* is a coefficient that biases the weight of evidence each document representation *j*, contributes to the similarity measure.
4. When the most similar cluster is found the thresholding strategy is used to discover if this similarity measure is high enough to warrant the addition of that document to the cluster and the classification of the current document as an old event. If this document does not satisfy the similarity condition set out by the thresholding methodology then the document is declared to discuss a new event, and this document will form the seed of a new cluster.
5. This clustering process will continue until all documents in the input stream have been classified.

## 3. THE DATA FUSION EXPERIMENT

Using the TDT1 corpus as input, the objective of this experiment was to determine if the use of a combined representation (a lexical chain and proper noun representation) would lead to improved FSD performance compared with a singular document representation using either proper nouns or chain words. Figure 1 is a Detection Error Tradeoff (DET) graph showing the impact of our combined representation on detection. A DET graph illustrates the tradeoff between misses and false alarms, where points closer to the origin indicate better overall performance. As can be seen the graph with the closest point to the origin is the LexDetect system, leading to the conclusion that a composite document representation using chain words and proper nouns marginally outperforms a system (CHAIN and P_NOUN) containing only either one of these representations. Optimal results for the LexDetect system in this experiment were achieved when both chain and proper noun representations were considered as equal evidence of similarity between two documents i.e. $w_j = 1$ for both representations in equation (1).

## 4. CONCLUSIONS

A variety of techniques for data fusion have been proposed in IR literature [5]. Results from data fusion research have suggested that significant improvements in system effectiveness can be obtained by combining multiple index representations, query formulations and search strategies. In this paper we investigated if improved FSD performance could be achieved when a composite document representation was used in this TDT task. Our results showed that a marginal increase in system effectiveness is achieved when lexical chain (semantic) representations were used in conjunction with proper noun (syntactic) representations. In particular, we saw that the miss rate of our FSD system LexDetect decreased with little or no impact to the false alarm rate of the system.

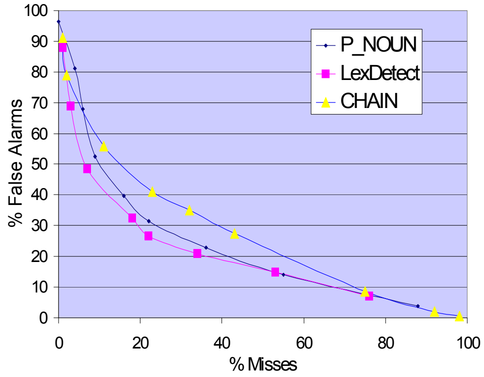

**Figure 1. The effect on performance when a combined document representation is used.**

## 5. ACKNOWLEDGMENTS

This project is funded by an Enterprise Ireland Grant, project number [SC/1999/083].

## 6. REFERENCES

[1] Ron Papka, James Allan, Topic Detection and Tracking: Event Clustering as a basis for first story detection, Kluwer Academic Publishers, 4:97-126, 2000.

[2] Jane Morris, Graeme Hirst, Lexical Cohesion by Thesaural Relations as an Indicator of the Structure of Text, Computational Linguistics 17(1), March 1991.

[3] Christiane Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[4] Nicola Stokes, Paula Hatch, Joe Carthy, Topic Detection, a new application for lexical chaining?, In the Proceedings of BCS IRSG Colloquium 2000, pp. 94-103, 2000.

[5] W. B. Croft, Combining Approaches to information retrieval, Advances in Information Retrieval, 1:1-36 Kluwer Academic Publishers, 2000.