Lexical Semantic Relatedness and Online New Event Detection

Nicola Stokes, Paula Hatch, Joe Carthy

Department of Computer Science, University College Dublin,
Ireland.
{nicola.stokes, paula hatch, joe.carthy}@ucd.ie

1. Introduction

In recent years there has been an explosive increase in the volume of information available on the Internet. In particular, we are concerned with the huge increase in the availability of multiple news sources reporting essentially the same news. In general, users whether they are journalists or ordinary consumers, are not interested in the full spectrum of world events. Also, they don't want to read the same story repeated several times from different sources. Instead people want to be able to follow events in which they have a particular interest as the story unfolds. They may also wish to be notified of any breaking news stories as they occur.

2. Topic Detection and Tracking

The Topic Detection and Tracking (TDT) Initiative was started to address some of these problems. The aim of this project is to investigate techniques for finding and following stories in a stream of broadcast news. In this project we are concerned mainly with the task of online new event detection. Previous TDT research [1] has tended to focus on building better clustering techniques to improve detection accuracy. In this project we aim to develop new techniques for story and event representation and so improve system accuracy.

3. Description of Experiment

Our new approach to document representation is based on the idea of conceptual indexing using lexical chaining.

In order to evaluate the advantages of this approach we have built two benchmark systems, TRAD and SYN [2]. TRAD uses the traditional IR method of representing documents using a 'bag of words' approach where terms are weighted according to their frequency within a document. SYN attempts to improve on this representation by solving the problem of synonymy. Thus documents are represented by sets of WordNet sense identifiers, i.e. numbers which represent meanings. However, SYN provides no facility for sense disambiguation and hence actually performs worse than TRAD. In the following section we describe how lexical chaining can be used to solve the problems presented by TRAD and SYN.

4. Lexical Chaining

'A text or discourse is not just a set of sentences each on some random topic. Rather the sentences and phrases of any sensible text will each tend to be about the same things – that is, the text will have a quality of unity. This is the property of cohesion... it is a way of getting text to hang together as a whole', Morris and Hirst [3].

A lexical chain is a succession of semantically related words in a text that creates a context and contributes to the continuity of meaning. This method of representing lexical cohesion using lexical chains was first formulated by Hasan [4, 5], who used them to measure the coherence of stories made up by children. Morris and Hirst then designed an algorithm that automatically built these chains. The lexical chains in a text can be identified using any lexical resource that relates words by their meaning. We use the WordNet thesaurus software [6] as our lexical resource.

Lexical chains have three primary purposes within our research:

1. Lexical Chains are associated with the topics that occur in a text.
2. A side effect of lexical chain creation is that words within a text are automatically disambiguated when considered in the context of the topic or concept being described by the lexical chain containing the word.

For example, when the word 'bank' is added to the following chain *{money, bank_manager, interest_rates, bank}* it is apparent that the meaning of bank in this context relates to the 'financial' rather than the 'river' sense of bank.

3. By disambiguating words in a text, the problems of *synonymy* (many words referring to the same concept) and *polysemy* (many concepts having the same word) are both addressed. The impact of *synonymy* is that documents and clusters which use words that are synonyms of one another will not be considered related or at best will be considered to be less related than they actually are. *Polysemy* will have the opposite effect, causing documents and clusters that use the same word in a different sense to be considered related.

Current topic detection systems base their clustering mechanism on the idea that if a document and a cluster share enough terms, then that document refers to the same topic represented by the cluster. Our aim is to incorporate both syntactic (term repetition) and semantic (sense repetition) similarity into our clustering/topic detection algorithm by representing documents by an ordinary 'term index' and a 'conceptual index' which contains the lexical chains present in a particular document.

So for every candidate term of a document we first look for its unique set of sense identifiers stored in the WordNet noun file where each sense identifiers refers to a different sense of the word. If the word is not found then it is added to the term index otherwise the word is added to an existing lexical chain based on some semantic relationship between it and a word in that chain. If no such related chain is found the word is used to seed the beginning of a new chain for the document. When all candidate terms have been considered for a particular document the lexical chains created are then written to the 'concept index'.

Once a document has been represented as above, this document representation is compared to all topics /clusters centroids that have been detected by our topic detection system. If an acceptable level of similarity exists between the document and a particular cluster then the document representation is added to the cluster centroid representation, otherwise the document representation becomes the cluster centroid of a new cluster and we say that a new topic has been detected.

5. Summary

In this paper we propose a novel use for lexical chains as document representations within the Topic Detection domain.

6. Acknowledgments

This project is funded by the Enterprise Ireland basic research grant scheme [SC/1999/083].

7. References

[1] James Allan et al., *Topic Detection and Tracking Pilot Study Final Report*, In the proceedings of the DARPA Broadcasting News Transcript and Understanding Workshop 1998, pp. 194-218.

[2] Paula Hatch, Nicola Stokes, Joe Carthy, *Topic Detection, a New Application for Lexical Chaining?*, In the proceedings of BCS-IRSG 2000, Cambridge, pp. 94-103.

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

[4] R Hasan, *Coherence and Cohesive Harmony*, in J.Flood(ed), Understanding Reading Comprehension, IRA: Newark, Delaware, 1984.

[5] M Halliday, R Hasan, *Cohesion in English*, Longman: 1976.

[6] George Miller, Special Issue, *WordNet: An On-line Lexical Database*, International Journal of Lexicography, 3(4), 1990.