A Text Annotation Method Based on Semantic Sequences

JunPeng Bao† j.bao@herts.ac.uk

Caroline Lyon‡ c.m.lyon@herts.ac.uk

Peter C. R. Lane‡ peter.lane@bcs.org.uk

November 25, 2006

†Dept. of Computer Science & Technology, Xi'an Jiaotong University, Xi'an 710049, China
‡School of Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK

Abstract

This paper presents a text annotation method based on semantic sequences to label a document and a cluster of documents. The basic idea underlying the semantic sequence approach is to find locally frequent meanings to act as the labels of a document, using an ontology such as WordNet. The ontology is also used to measure the semantic similarity of labels that indicate similarity between documents. Further, a text clustering method based upon four natural rules is introduced to cluster documents and label each cluster. This method does not need any pre-defined number of clusters, which is necessary for the partitioning clustering method, and avoids the need to set appropriate levels as in the hierarachical clustering method.

1 Introduction

It is a great challenge to find information from a large collection of documents that satisfies a particular user's need. Many research areas, such as Information Retrieval/Extraction, Text Classification/Clustering, Text Mining, and Semantic Analysis have to address this issue from their various perspectives. Amid all these areas, a basic question is confronted: what is a document mainly about? Or, what topics is a document addressing? In this paper, a text annotation method based on semantic sequences is presented in order to give a document an appropriate label that consists of a combination of several words or phrases and indicates the document's topic categories.

Nowadays, many ontologies, such as WordNet, SUMO [4], and MILO [3] are used to describe the relationships among words and meanings in a format that can be understood by both human and computer. Ontology-based methods find and match concepts from a document, and then derive topic labels, or similarity measures between documents, based on concepts.

We find in practice that a document's content may cover various topics, and each topic is supported by a number of concepts. Thus concepts involved in the topic are often repeated in a document. If these frequent concepts are found then the main topics of the document can be derived. As a result, these frequent concepts are appropriate labels for the document.

2 Semantic Sequence Based Annotation

2.1 Semantic Similarity

In previous work, Bao [1] presented a semantic sequence method to extract locally frequent word strings. In this prior work, a semantic sequence was taken as a sequence of words that were repeated in a local section. In the current work, *meaning* repetition is considered, rather than *word* repetition. WordNet is used to find the meaning of each word, and then locally frequent meanings are extracted to construct a semantic sequence. Additionally, a consecutive meaning sequence gives us more context and reduces ambiguity so that it carries more weight than a single isolated meaning in our model. The consecutive meaning sequences and the most frequent isolated meanings are extracted from the semantic sequences in a document and taken as the labels of the document. For example, the following paragraph is extracted from the file 6203newsML in the Reuters Corpus Volume 1 [5]. Three labels are extracted from the paragraph: "budget", "small business", and "small business confidence".

> Australian Prime Minister John Howard said after a fall in confidence ahead of his government's first budget, [that] the reduction of the deficit and the prospect of lower rates should help rebuild small business confidence. "I hope the budget will have a very beneficial effect," Howard said in an interview on television, responding to a question on when small business could expect to see consumers start spending again after a pre-budget slump.

Semantic similarity between documents can be measured based on the labels of a document. WordNet is used again to find the similarity between meanings. Let $|L_A|$ denote the number of meanings in the labels of document A. $|L_A \cap L_B|$ denotes the number of similar meanings between labels of two documents A and B. The semantic similarity is $S(A,B)$.

$$S(A,B) = \frac{1}{2}\left(\frac{|L_A \cap L_B|}{|L_A|} + \frac{|L_A \cap L_B|}{|L_B|}\right) \tag{1}$$

2.2 Semantic Text Clustering

Popular text clustering approaches include the hierarchical method, partitioning method and density method [2]. The hierarchical clustering method can be applied in either a top-down or a bottom-up manner, and then an appropriate intermediate level has to be determined as required. With the partitioning method, a pre-defined number of clusters k is needed, which is quite arbitrary in most cases. The semantic similarity is a kind of distance measure, so that it can not be applied to density-based text clustering.

The text clustering algorithm described here is based on four rules and does not need to pre-define cluster numbers. (a) *Expanding a cluster*: If a file is similar enough to most members of a cluster, then the file belongs to the cluster – a file may belong to more than one cluster; (b) *Creating a cluster*: If a file does not belong to any cluster, then it becomes a new cluster itself; (c) *Merging clusters*: If most members of a cluster are contained in another, then that cluster is merged with the other; and (d) *Splitting clusters*: If two clusters have a considerable overlap, then the overlapped part is extracted from the two clusters and becomes a new cluster. *Similar enough* is defined as $S(A, B) > t_1$, *most members* is defined as more than p percent of the total members, *considerable overlap* is defined as $t_2 < O_{Ci,j} < t_3$ where $O_{Ci,j}$ means the proportion of overlapped members to the whole cluster C_i and C_j. In current experiments we take thresholds $t_1 = 0.8$, $t_2 = 0.4$, and $t_3 = 0.8$ and $p = 80\%$. However, these parameters can be adapted to different applications.

After clustering, the cluster labels are derived from the most common meanings in the cluster.

3 Discussion

WordNet can find words of similar meanings, but it often gives a number of meanings for a word, some of which are very specialised and used infrequently, while some are described as rare which we would consider common. The number of meanings decrease processing efficiency and make it hard to find the right context. As WordNet is a general-purpose ontology it lacks coverage of the meanings and relationships of some domain-specific terms.

Many terms that consist of two or more words can express clear concepts. We should like to be able to take these as single terms rather than as several words. A domain-specific ontology could contain many more multi-word terms than WordNet, and it is important for an application to be supported by appropriate ontologies.

The semantic sequence method ignores infrequent words or meanings so useful information can be lost, especially in short texts. Words that occur only once make up a significant part of most texts. However, by extending the method from words to meanings we have a better chance of capturing the underlying concepts. The method works better with longer texts because topic meanings usually occur sufficiently frequently.

4 Conclusions

We propose a method for labelling a single document and also for labelling a cluster of documents. This can be used to annotate document categories effectively. We have extended the original semantic sequence method from word repetition to meaning repetition, using WordNet. Four rules are presented for clustering a collection of texts and labelling each cluster. This text clustering method is based on semantic similarity, which is a kind of distance measure. It avoids the need to find appropriate levels in hierarchical clustering, and the need to pre-define the number of clusters in the partitioning method. However, the disadvantages of our method are that it is not fit for very short texts and that relevant information is sometimes omitted.

Acknowledgments

Dr. JunPeng Bao carried out this work at the University of Hertfordshire, UK, sponsored by the Royal Society as a Visiting International Fellow.

References

[1] J. P. Bao, J. Y. Shen, X. D. Liu, H. Y. Liu, and X. D. Zhang. Semantic sequence kin: A method of document copy detection. In *Proceedings of the Advances in Knowledge Discovery and Data Mining*, volume 3056, pages 529–538. Lecture Notes in Computer Science, 2004.

[2] M. Hearst. Clustering versus Faceted Categories for Information Exploration. In *Communications of the ACM*, 49, 2006.

[3] I. Niles, and A. Terry. The MILO: A general purpose mid-level ontology. In *Proceedings of the International Conference on Information and Knowledge Engineering*, 2004.

[4] A. Pease, I. Niles, and J. Li. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, 2002.

[5] T. Rose, M. Stevenson, and M. Whitehead. The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, 2002