Mining Domain Specific Words from Web Documents

Jing-Shin Chang
Department of Computer Science & Information Engineering
National Chi-Nan University
1, Univ. Road, Puli, Nantou 545, Taiwan, ROC.
jshin@csie.ncnu.edu.tw

Abstract

Web pages provide not only plain text materials for training language models but also tag information for semantics annotation. The tags could be found either explicitly in the HTML documents or implicitly through the directory hierarchy of the documents, since the directory hierarchy can be regarded as a kind of classification tree for web documents, which assigns an implicit hidden tag to each document and hence the embedded words. For instance, the domain-specific words for documents under the “sport” hierarchy are likely to be tagged with a “sport” tag. These tags, in turn, can be used in various word sense disambiguation (WSD) tasks and other hot applications like anti-spamming mail filters. Such rich annotation provides a useful knowledge source for mining various semantic links among words. This presentation proposes a statistical method for finding domain-specific words in particular domains, and thus their associations, by taking advantages of the hierarchical structure of the web pages. With the statistical model, the document tree can virtually be converted into a large semantically annotated lexicon tree. Some preliminary results show that the current approach has its strength in finding domain-specific words.

Keywords: Word Association, Implicit Sense Labels (Hidden Tags), Cross-Domain Entropy, Word Sense Disambiguation, Lexicon Acquisition, WordNet