Automated Word Sense Disambiguation for Web Information Retrieval

Christopher M. Stokoe
School of Computing and Technology
University of Sunderland, UK.
christopher.stokoe@sunderland.ac.uk

http://www.cet.sunderland.ac.uk/~cs0cst/download/thesis.pdf

A word in the English language is considered ambiguous if, regardless of context, it can have more than one possible interpretation or meaning. Many words exhibit lexical ambiguity suggesting that it has the potential to impact upon the performance of text retrieval systems. This may be particularly true in the case of web retrieval given the hypothesis that short queries may not provide sufficient context to adequately differentiate between opposing meanings of constituent words. Word sense disambiguation is an active field of study which seeks to create software which automatically resolves ambiguity through mapping word use to meaning. In this study the author examined the use of word sense disambiguation in order to resolve ambiguity within an IR collection. The motivation behind this work was to demonstrate the potential for increased retrieval effectiveness as a result of performing word sense disambiguation.

The experimental work consisted of the design, development, and evaluation of a supervised word sense disambiguator for use in information retrieval. An evaluation of the disambiguator's accuracy demonstrated that it had performance comparable with state-of-the-art disambiguation systems. The disambiguator was subsequently used to produce a sense based document representation from which to perform retrieval. The quality of this representation was evaluated comparatively against a document representation where ambiguity had not been directly resolved. Results showed increased retrieval effectiveness when performing retrieval from a sense based representation as opposed to the traditional term based model. Subsequent experiments highlighted features of both the disambiguation and the problem domain in order to explain why the results of this study run contrary to those previously reported in the literature. These features include the short average query size associated with web retrieval and the inherent frequency bias that exists in supervised disambiguation systems.