http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 7 (1): 231-233, 2008
ISSN 1812-5638

Effective Information Retrieval Using Supervised Learning Approach

[1]M. Sundara Rajan and [2]S.P. Rajagopalan
[1]Department of Computer Science, S.R.M. Arts and Science College, 61-2nd Main Road,
Baby Nagar, Velachery, Chennai, Pin-600042, TN, India
[2]School of Computer Science, Engineering and Applications, M.G.R. University,
Maduravoil, Chennai 600095, TN, India

Abstract: It has often been thought that word sense ambiguity is a cause of poor performance in Information Retrieval (IR) systems. The belief is that if ambiguous words can be correctly disambiguated, IR performance will increase. However, recent research into the application of a word sense disambiguator to an IR system failed to show any performance increase. From these results it has become clear that more basic research is needed to investigate the relationship between sense ambiguity, disambiguation and IR. Using a technique that introduces additional sense ambiguity into a collection, this study presents research that goes beyond previous work in this field to reveal the influence that ambiguity and disambiguation have on a probabilistic IR system. We conclude that word sense ambiguity is only problematic to an IR system when it is retrieving from very short queries. In addition we argue that if a word sense disambiguator is to be of any use to an IR system, the disambiguator must be able to resolve word senses to a high degree of accuracy.

Key words: WSD, IR, disambiguator, word senses, synsets, selectors

INTRODUCTION

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.

So far, there has been conflicting information on the effect of WSD on IR. While a sense disambiguator was described that improves the precision of an IR system by 4% (Schutze and Pedersen, 1995), results which show that disambiguation (unless at least 90% accurate) makes IR performance worse were also presented (Mark Sanderson, 1994). There has been a lot of work dealing with combination of word sense with information retrieval. The first work where a word sense based algorithm was used with an IR system was done in 1973 (Weiss, 1973). He tested the algorithm on five handpicked words in the Aviation Data Internet (ADI) collection. Resolving ambiguities before information retrieval, it was shown that it improved performance of an IR system by 1% for the five test words. More interesting results were established, giving information about the relation between Word Sense Disambiguation (WSD) and IR (Krovetz and Croft, 1992). It was concluded that WSD did not have a very important impact on IR, but that disambiguation could be beneficial to IR when there were a few words in common between the query and the document. A method using WordNet to disambiguate word sense for text retrieval was presented in 1993 (Voorhees, 1993). In 2003, a study was made (Stokoe and Tait, 2003), exploring the using of word sense disambiguation in retrieval effectiveness and subsequent evaluation of a statistical word sense disambiguation system which demonstrates increased precision from a sense based vector space retrieval model over traditional tf-idf techniques.

Over the past ten years a number of researchers have worked on trying to integrate Word Sense Disambiguation (WSD) techniques into text based (IR) systems in an attempt to eliminate ambiguity and increase effectiveness. However it is important to note that often work in this field has been difficult to assess due to a failure to effectively evaluate the accuracy of the

Corresponding Author: M. Sundara Rajan, Department of Computer Science, S.R.M. Arts and Science College, 61-2nd Main Road, Baby Nagar, Velachery, Cheunai, Pin-600042, TN, India Tel: 98402 40094

disambiguation used. To this end we identified the need to re-examine the possible effects of automated word sense disambiguation in text retrieval systems using more rigorous performance measures.

Given that studies have identified short queries may benefit most from a disambiguated collection we set out to evaluate the performance of automated word sense disambiguation within a web search system. The aim of this experimental work was to assess the relative benefits of searching from a collection with reduced ambiguity in an attempt to identify whether the introduction of automated word sense disambiguation can produce more effective results. In this study we examine the use of word sense disambiguation in order to resolve ambiguity within an IR collection.

The motivation behind this study 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. 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.

Despite the increasing importance of IR systems as data retrieval tools, the performance of most of these systems has not yet reached a satisfactory level. Word sense ambiguity is one of the reasons for their poor performance. Overcoming this problem may improve IR performance. Documents related to an IR query sometimes contain only the synonyms of the query words instead of the query words themselves. A simple IR system with no knowledge of synonyms fails to recognize the relevance of these documents to the query. So, we can improve recall of IR systems by considering the synonyms of the query words as a part of the IR query. However, only relevant synonyms of the query words in the given context contribute useful information to the query. We can identify these relevant synonyms with the help of a disambiguation algorithm.

MATERIALS AND METHODS

We use the local context of a word to identify its sense. Due to this definition of context, words used in the same context (called selectors) most of the time have similar or related meanings. That is, an occurrence of a word and its synonym often belong to the same sense if they have similar local contexts.

We use WordNet and selectors extracted from Associated Press Articles to find the appropriate synset of a word in its context. The figure below shows some of the selectors of the word charge in two different sentences. The examples show that contexts play an important role in finding selectors which enable us to identify the correct sense of an ambiguous word.

Local context is the ordered list of words from the closest context word on each side up to the target word expressed as a placeholder. For example, in the jury had been charged to investigate reports of irregularities in the primary, the right-side local context of charged is to investigate.

Disambiguation of charged in two different contexts: Let us consider the first context. The sentence under scrutiny is The company was charged for towing the car. The disambiguation process searches the corpus for the sentences with company or towing. In either case sentences like company was founded, company was fined, charged for towing, billed for towing might be found. So the disambiguator identifies words like founded, fined, charged and billed as selectors. The frequency count for the selectors is incremented for each of the words found. The frequencies are then listed and the word sense with the highest frequency is selected as the sense for that particular word. In this case the word identified is billed.

Similarly, in the next context the sentence under ambiguity is the jury had been charged to investigate reports of irregularities. The corpus is searched for sentences containing the words jury and investigate. The corpus has sentences like jury had been appointed, jury had been selected, commissioned to investigate, assigned to investigate. Selectors are then identified as appointed, selected, commissioned and assigned. The frequency counters are incremented and the sense with the highest frequency appointed is identified.

The WordNet senses of the input word charged are given in the Table 1.

Comparing the selectors of the input word against the WordNet Synset matches input sense 1 to WordNet sense 3 and input sense 2 to WordNet sense 4. The process has selected the most appropriate WordNet senses.

Table 1: Different Senses for the word charged in WordNet

Sense 1	Sense 2	Sense 3	Sense 4	Sense 5
Charged, bear down	Charged, accused	Charged, billed	Appointed, charged	Assigned, charged

Difficulty: Correctly disambiguating words is a difficult problem. When restricted to available on-line dictionaries like WordNet, it is sometimes impossible even for human beings to pick the right sense for words. Expecting a machine to resolve such ambiguities is not reasonable. But, a good online dictionary with example uses of words in each of their possible senses can allow a machine to disambiguate words accurately. Such dictionaries are not yet available.

RESULTS

The main aim of my work is to assess whether automated word sense disambiguation could be used to improve retrieval effectiveness. Although the use of automated disambiguation did lead to a small (0.0003%) increase, this is considered statistically insignificant and as such the overall results were disappointing. The disambiguation algorithm was tested on the Semcor corpus where each word is tagged with its correct part-of-speech and sense number from WordNet. On this corpus, the accuracy of our disambiguator was almost 60% excluding words which have only one sense. When incorporated into the IR system SMART, the disambiguation did not improve performance. Although in some cases the expansion of the query with synonyms helped, especially for short queries the disambiguation accuracy was low. Incorrect disambiguation not only excludes correct synonyms from the query but it also introduces incorrect information to it reducing retrieval performance. Although 60% accuracy is not insignificant for an unsupervised algorithm which tries to disambiguate any content word in a context, the performance of this disambiguator can be improved with the use of better online dictionaries with less fine-grained sense distinctions. Improving disambiguation performance can help IR.

FUTURE WORK

In the long term, the key idea of engineering a WSD system and information retrieval mechanism in a manner that seeks to reduce the negative impact of inaccurate disambiguation merits further study. We also plan on expanding the training data for our disambiguation system in an attempt to increase the WSD accuracy. Previous research suggests that using cross-lingustic information for disambiguation performs better than single language disambiguation. There is a lot of contextual information which is lost by trying to disambiguate an ambiguous word whose context is also ambiguous. Cross linguistic information can, to a certain extent, disambiguate the context of the ambiguous word and help the disambiguation of the word itself. Additionally, there is scope to explore the upper bounds for WSD performance within IR as disambiguation precision moves further beyond the baseline of sense frequency. Our future work will focus on development of such a system which, we expect, will significantly improve performance.

REFERENCES

Krovetz, R. and W.B. Croft, 1992. Lexical ambiguity and information retrieval in ACM. Transactions on Information Retrieval, Vol. 10.

Sanderson, M., 1994. Word-sense disambiguation and information retrieval. In: Proceedings of ACM-SIGIR.

Schutze, H. and J. Pedersen, 1995. Information Retrieval Based on Word Senses. In: Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, U. of Nevada at Las Vegas.

Stokoe, C.M. and J. Tait, 2003. Word Sense disambiguation in information retrieval revisited. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR2003), pp: 159-166.

Voorhees, E.M., 1993. Using wordnet to disambiguate word senses for text retrieval. In: SIGIR '93, Proceedings of the 16th Annual International ACM SIGIR Conference of Research and Development in Information Retrieval, pp: 171-180.

Voorhees and M. Ellen, 1998. Using Wordnet for Text Retrieval. In: WordNet, an Electronic Lexical Database, MIT Press, Cambridge MA, pp: 285-303.

Weiss, S.F., 1973. Learning to disambiguate. Information Storage and Retrieval.