International Journal of Soft Computing 2 (5): 592-594, 2007
ISSN: 1816-9503

Information Retrieval in Medical Domain Using Root Sense Tagging 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-600 042 TN, India
[2]School of Computer Science, Engineering and Applications, M.G.R. University, Maduravoil,
Chennai 600095 TN, India

Abstract: Medical NLP systems, generally designed to analyze medical texts for decision support or indexing purposes, have to deal with ambiguities in language. Information Retrieval (IR) in the Medical Domain often ignores word senses in document relevance calculations. This is largely due to the fact that word sense disambiguation is not an easy task. However, in IR some of the strictest requirements of traditional approach to detailed disambiguation can be relaxed. This study discusses a successfully implemented approach to using word senses in IR tasks, which can be combined with various IR methods for successful use in the Medical Domain.

Key words: IR (Information Retrieval), WSD (Word Sense Disambiguation), WN (Word Net), RST (Root Sense Tagging), MWN (Medical Word Net), Natural Language Processing (NLP)

INTRODUCTION

Information Retrieval (IR) systems are used to retrieve documents from document collections that are relevant to a posed question. The amount of information on the Web is growing rapidly as well as the number of new users inexperienced in the art of Web research. Although it is full of information, finding the information sometimes resembles picking needles from haystacks. Questions are often composed of one or more words such as Bush or George Bush (looking for documents of President George Bush). These requests do retrieve the relevant documents, but they retrieve also documents with the other senses of word bush, like bush as plant or vegetation. This problem is not rare, since word sense ambiguity is always present in text. Nonetheless, many IR systems do not use semantical information when retrieving documents. A comprehensive study of Web queries concluded that queries are normally very short-an average user query is only 2.3 words. In addition, the study also finds that for 85% of the queries, only the first result screen is viewed and 77% of the sessions only contain one query-that is the queries were not modified in these sessions. From this study, it can be concluded that most users submit short, abbreviated queries to search engines. As a result they receive a long list of "Hits". The retrieved documents share terms with a query but due to the polysemy in natural language, the documents may not be related to user's information needs. Due to the synonym property of the natural languages and the fact that a majority of users do not perform any query modifications, many relevant documents do not share terms with query, so those documents will not be retrieved. This may also affect the relevancy ranking of the search engines.

Another study found that when an experienced user performs interactive query expansion, it could significantly improve the search process. However, results also showed that inexperienced users did not make good term selections; therefore, interactive query expansion led to no improvement in the search process. However, disambiguation is not an easy task, which is one of the reasons why word senses are often ignored in IR. Much research on disambiguation area has been done and methods have improved, so it is ever closer to benefit from disambiguation in IR. This study will discuss different aspects of using word sense disambiguation in IR and of how to implement it effectively. The study will discuss different approaches to using Word Sense Disambiguation (WSD) in IR and the reasons why WSD has failed (Sanderson, 2000).

Corpus- and fact-based approaches to information retrieval: Patel *et al*. (2002) make clear that if a medical information system is to mediate between experts and non-experts, then it must rest on an understanding of

Corresponding Author: M. Sundara Rajan, Department of Computer Science, S.R.M Arts and Science College, 61-2nd Main Road, Baby Nagar, Velachery, Chennai, Pin-600 042 TN, India

both expert and non-expert medical vocabulary. But terms, or word forms, are not always associated with word meanings in a clear-cut and unambiguous fashion; and the problem of polysemy is compounded when different speaker populations are involved. A lexical database must represent all and only the meanings of each given term in such a way that these meanings can be clearly discriminated and mapped onto word occurrences in natural text and speech. Achieving these ends is one of the hardest challenges facing both theoretical and applied linguistic science today. It is generally agreed that the meanings of highly polysemous terms cannot be discriminated without consideration of their contexts (Pustejovsky, 1995). People manage polysemy without apparent difficulties; but modeling human speakers' capacity for lexical disambiguation in automatic language processing tasks is hard. The idea underlying the present proposal draws on currently emerging NLP methodologies that harness the ability of powerful and fast computers to store and manipulate both lexical databases and large collections of text collections or corpora. The strategy is to train automatic systems on large numbers of semantically annotated sentences that are naturally used and understood by human beings and to exploit standard pattern-recognition and statistical techniques for purposes of disambiguation. Words and the representation of their senses, stored in lexical databases, can be linked for this purpose to specific occurrences in corpora.

Word sense approaches in IR: The disambiguation task in IR has been approached mainly from three different perspectives (Sanderson, 2000):

- Dictionaries and synonym sets have been used to choose the correct interpretation for a word.
- Word senses are derived from corpora.
- Retrieval is done without identifying word senses.

The first approach assumes that a word has a predefined number of (dictionary) senses and the task of disambiguation is to decide, which of these senses is the right one. The decision can be done in various ways. For example one can use the dictionary descriptions of words, compare them with the word context in the text and choose the (correct) sense. The chosen sense has the largest number of common words with the context. As another example, if one has a dictionary where words are organized into a hierarchical hyponym tree, such as WordNet (WN), one can calculate the semantic distance between any two words. Then given an ambiguous word appearing in text, all the synonym sets (senses) containing that word are looked up in WN and the senses are scored based on the semantic distances between the context of the word and the sense. The highest scored sense is chosen. The problem of choosing the right sense for the word is that the other words appearing in the context can also be ambiguous. Thus the problem accelerates as the sense combinations increase exponentially. The precision of disambiguators needs to be high, if they are to be of benefit for IR systems, otherwise those documents that are assigned faulty word senses reduce the precision and recall of the result.

Precision is often lower the more fine grained the senses are. For IR purposes a more coarse grained, but more accurate method might be useful. The second approach assigns senses to words using their contexts taken from the document collection. The contexts can be clustered and the clusters then represent word senses. This approach leaves open the question of how the users could mark up their query with wordsenses, since the senses are not real senses, but are defined by the clusters of surrounding context words. Also, the problem of clustering is that it often requires heavy computation. The third approach does not use the word senses for disambiguation when choosing the right documents, but instead uses the information of multiple senses in other ways. For example, query words can be assigned weights of query relevance based on ambiguousness: words with many senses are probably less useful in retrieval than words with only one sense, since they are more likely to bring irrelevant documents to the result set. The drawback is that, not all senses are covered by dictionaries and also some queries have junk words that are irrelevant to the query despite their non-ambiguousness (e.g., I want articles that).

From wordnet to medical wordnet: A consumer health information system must be able to comprehend both expert and non-expert medical vocabulary and to map between the two. WordNet is the principal lexical database used in Natural Language Processing (NLP) research and applications (Miller, 1995; Fellbaum, 1998). While WordNet's current version (2.0) has broad medical coverage, it manifests a number of defects and also the fact that WordNet was not built for domain-specific applications. So a Medical WordNet (MWN) (Smith *et al*., 2004) a free-standing lexical database was designed specifically for the needs of natural-language processing in the medical domain. MWN's initial focus is on English single-word expressions as used and understood by non-experts. It systematically reviewed WordNet's existing medical coverage by assembling a validated corpus of sentences involving specific medically relevant vocabulary. A major stumbling block for existing NLP applications is automatic sense disambiguation. An automatic system can detect with high reliability that a given occurrence of a word like feel

or dead is a verb or adjective. But it cannot easily determine which of a variety of alternative meanings such polysemous words have in any given context. WordNet's architecture, designed for representing and distinguishing word senses, has made an important contribution towards a solution of the automatic word sense disambiguation problem.

Root Sense Tagging (RST) approach: This approach presents an IR method using word senses that has promising results. The study discusses the Root Sense Tagging (RST) method (Mirva Salminen, 2004), which is designed to overcome some of the problems faced in the previous IR systems using word senses.

The RST approach is based on three principles, which all aim to solve some problems of disambiguation in IR:

The RST approach does not use fine grained senses but instead uses coarse grained disambiguation that is assigns words only to their root senses. For example, .actor. has two fine grained senses actor as a doer and actor as a role player, but only one coarse grained sense person.

Accurate disambiguation is complicated, because the context of words in different documents, even the context of the words with same fine grained sense differs some times and also different senses can appear in similar context. The RST approach thus relies on consistent disambiguation instead of accurate disambiguation. Consistent disambiguation assumes that it is more important to assign senses in a consistent manner even though the consistent manner is sometimes faulty.

Word sense disambiguation does not yet reach high enough accuracy for IR needs. This problem is passed by using flexible disambiguation instead of strict one, which allows for multiple sense assignment for a word.

RST approach for the medical IR: RST was basically designed to work with WordNet. But it can also be applied to work with MedicalWordNet as its construction is similar to WN. If a word is ambiguous in the medical domain, it has multiple root senses in the MWN and is classified based on the context words in the document. Consider the following sentences that include the word cold taken from three different MEDLINE® abstracts:

(1) A greater proportion of mesophil microorganisms were to be found during the cold months than in warmer months.
(2) In a controlled randomised trial we analysed whether the use of the term "smoker's lung" instead of chronic bronchitis when talking to patients with Chronic Obstructive Lung Disease (COLD) changed their smoking habits.
(3) The overall infection rate was 83% and of those infected, 88% felt that they had a cold.

The sense of the word cold is different in each sentence. Cold in sentence (1) is an indication of the temperature, in sentence (2) the acronym of chronic obstructive lung disease and in sentence (3) cold is a disease.

The disambiguation is done using MI based RST using co-occurrence data. Our methodology is designed (1) to document natural language sentential contexts for each relevant word sense in such a way that the expressed information can be (2) validated by medical experts and (3) accessed automatically by NLP applications such as information retrieval, machine translation, question-answer systems and text summarization.

CONCLUSION

The root sense tagging approach has a lot of both advantages and disadvantages. The advantages are really worth taking into account as the root sense tagging method seems to be a useful approach to applying word senses in IR, since it mostly improves results. The method takes the word senses into account, but does not require accurate and detailed disambiguation. The method does not require complex computation. The major disadvantage with this approach is this method does not use verb and adjective senses of WordNet (WN) and this method does not solve the problem of synonyms.

REFERENCES

Fellbaum, C., 1998. WordNet: An electronic lexical database. MIT Press, Cambridge, MA.

Miller, G.A., 1995. WordNet: A lexical database for English. Comm ACM 38, 11, pp: 39-41.

Mirva Salminen, 2004. Research Seminar on Intelligent Systems, Helsinki.

Patel, V.L., J.F. Arocha, J.F. and A. Kushniruk, 2002. Patients' and physicians' understanding of health and biomedical concepts: Relationship to the design of EMR systems. J. Biomed. Informatics, 35: 8-16.

Pustejovsky, J., 1995. The Generative Lexicon, MIT Press, Cambridge.

Sanderson, M., 2000. Retrieving with root sense. Inform. Retrieval, 2: 49-69.

Smith, B. and F. Christiane, 2004. Medical Word Net: A New Methodology for the Construction and Validation of Information Resources for Consumer Health. In: Proceedings of Coling: The 20th International Conference on Computational Linguistics, Geneva.