DISAMBIGUATION TECHNIQUE FOR POLYSEMOUS WORDS

Rekha Jain*

Neha Singh**

ABSTRACT

With the rapid growth of the Web, providing relevant pages of the highest quality to the users based on their queries becomes increasingly difficult and the level of difficulty increases with search for polysemous words. Usually users get easily fade out in the rich hyper text while searching over the web for the polysemous word. The Word Sense Disambiguation (WSD) technique is designed to identify which one of the multiple senses of a polysemous word that can be associated in a particular context around the word during the web search. In this paper an attempt is made to disambiguate polysemous word by selecting the most appropriate meaning or sense to a given ambiguous word which will result in more relevant and intelligent search from user's perspective.

***Keywords** : Word Sense Disambiguation, Polysemous Words.*

*Assistant Professor, Banasthali University, Jaipur.

**Banasthali University, Jaipur.

1 INTRODUCTION

With the rapid growth of the Web, providing relevant pages of the highest quality to the users based on their queries becomes increasingly difficult and the level of difficulty increases with search for polysemous words. The reasons are that some web pages are not self-descriptive and that some links exist purely for navigational purposes and sometimes inappropriate association of meaning of the word that results in indifferent search from users' perspective.

Homonymy and polysemy are two well-known semantic problems. A polysemy is a word or phrase with different, but related senses. *Bank* in *river bank* and *Bank of England* are homonymous i.e. they share no meaning and they function as two totally unrelated independent words. *River bed* and *hospital bed* seem to be somehow semantically linked; it shows the case of polysemy. These polysemous words provide a biggest challenge for Search Engines. In the field of computational linguistics, the problem is generally called word sense disambiguation (WSD), and is defined as the problem of computationally determining which "sense" of a word is activated by the use of the word in a particular context. WSD is essentially a task of classification: where word senses are treated as classes, the context provides the evidence, and each occurrence of a word is assigned to one or more of its possible classes based on the evidence. Word Sense Disambiguation (WSD) is one of the basic and typical tasks dealt in Natural Language Processing (NLP). The two properties of the English language that are considered in this technique are polysemy and homonymy. In other words WSD can be explained as the method that deals with issue of determining the correct meaning or sense of a word depending on the context with respect to which it is used. It has many applications and act as the intermediate task in computational linguistics, including Machine Translation, Information Retrieval, Text Mining or Speech Processing [1]. WSD is applied in a semantic analysis and understanding of text.

Due to many problems like lack of general world knowledge and dictionary insufficiencies for the treatment of polysemy, ranging from a variety of applied approaches for sense identification like Knowledge-based, Corpus based & Hybrid approach (Geeraerts, 1993; Cruse, 1986; Kilgarriff, 1997), the broader goal of this paper is to build a platform or an interface that will optimize the search for polysemous words. A more practical goal of the paper (and of the project) is to explore a method of solving such problems by classifying the search result characterized by the domain reflecting the users' perspective.

2 LITERATURE REVIEW

There are two main approaches in automated word sense discovery: Word Sense Disambiguation and Word Sense Discrimination. The objective of the former approach separates contexts of multiple meanings of a word into groups, usually by cluster analysis. The retrieved groups are assumed to represent different meanings. The aim of the latter is to automatically label a word with a sense tag taken from a predefined set of meanings, relying on tagged training corpora. The most relevant research areas of the proposed technique are word sense disambiguation. An overview of it and its previous work is presented in the following subsection.

In the middle of the 20th century, Word Sense Disambiguation in computational linguistics started emerging. The problem of WSD was first put forward in 1949 by Weaver who presented a mimeographed text discussing the need of WSD and elaborated a very important problem related to the context used for disambiguating words. Disambiguation of a certain word considers neighboring words. In the following decades researchers adopted many methods in an attempt to solve the problem of automatic word sense disambiguation, including: AI-based method, Knowledge based method and Corpus-based method [1]. But the problem arises in this WSD technique because of unavailability of standardized system for word sense disambiguation, difficulty in obtaining large sense-tagged data sets adequately and along with it potential for WSD varies by task [2].

The automatic disambiguation of word senses is a problem that has been studied for many years - Gale, Church and Yarowsky [7] cite work dating back to 1950. Earlier methods to devise disambiguators [8, 9, 10] relied on a combination of hand built lexicons and rules. They are working well for the examples they were programmed for, although researchers never succeeded in making the generalized disambiguators that work efficiently with large disambiguation problems. However in 1986 Lesk [11] built a disambiguator that do the semantic analysis by using the textual definitions of word senses in an on-line dictionary to provide sense evidence which in fact similar to the techniques used in IR. With this large reference work, Lesk's disambiguator had the potential to be applied to large scale problems. According to this disambiguator, the word W appearing in a certain context (for example, the 20 words surrounding W), the definitions of all the potential senses of W were looked up in the online dictionary and then the rank retrieval of the definitions is made and the sense defined by the top ranked definition was chosen as the sense of W.

Since Lesk's paper a bewildering range of disambiguators have been built: Cowie [12], Black [13], Wallis [14] and Demetriou [15] have made further use of dictionaries; Zernik [16] used a morphological analyser; Hearst [17] used learning based on human evidence; Dagan [18] used bilingual corpora; Church [19] made an attempt for aligned bilingual corpora; Voorhees [20] and Sussna [21] used the WordNet thesaurus; and Yarowsky [22] used a combination of Roget's thesaurus and Grollier's encyclopaedia to produce one of the better performing disambiguators to date.

3 PROBLEM STATEMENT

For resolving the ambiguity of a word, two essential ingredients are needed: some kind of knowledge related to the word and the context in which the word has been used. As for computers disambiguation accounts a big challenge however for humans it's just a mere task to disambiguate the words because they possess "general world knowledge". For example, if the word "table" appears in a text that also contains words like "furniture" or "wooden round", we will know that the word "table" refers to a piece of furniture and not to the arrangement of data. A word needs to be disambiguated only if it has multiple senses. There are four parts-of-speech that allow polysemy: nouns, verbs, adjectives and adverbs. This paper focuses on noun polysemous word to resolve ambiguation.

Since computers lack the "world knowledge" used by humans for disambiguation, they need to use some other resources and repositories for fulfilling this task. These resources can be split into three different categories:

- Dictionaries and other lexical resources used to define each possible sense of a word and hence act as the immense storage of possible meanings. Machine readable dictionaries have been in use for a long time in WSD. The most widely used dictionary is WordNet [11].
- Tagged corpora examples of text where each instance of a word is tagged with its corresponding sense. These can be used to learn and analyze the context in which each sense has a high probability to appear.
- Untagged corpora consist of sets of documents containing raw text. They are the base of sense-tagged corpora, but also used in WSD to derive useful statistics.

These resources are very important and even form the base for a WSD classification, but they are not sufficient enough to resolve the ambiguity properly.

Senses of the same word are seldom ambiguous in context, but the less specific the context, the greater the possibility of ambiguity; for example, if someone who is looking at a picture says "What big cranes!" it may not be immediately clear to anyone who cannot see the picture whether the comment refers to *birds* or *machines*. So, when the search for ambiguous polysemous word is made, due to inappropriate association of meaning with word, produces the irrelevant search result from user's perspective. In this paper effort is made to to develop an efficient disambiguation system that must be able to resolve word senses to a high degree of accuracy and can yield fruitful search and hence optimize the search result of polysemous word.

4 PROPOSED SYSTEM

The proposed system will make use of semantic knowledge in order to resolve ambiguity in entity extraction. The proposed technique identifies all possible meanings or senses of an entity and decides the most appropriate meaning of the entity inspired by the domain defined.

4.1 Framework

4.1.1 Polysemous Word:

A word that is categorized into the noun part-of-speech is defined as an entity. Polysemous are defined as "having or characterized by many meanings or could say the existence of several meanings for a single word or phrase". Here the polysemous word is provided by user for the search. Knowledge about the word will be helpful in associating the most possible sense of an ambiguous entity.

4.1.2 Knowledge Repository:

The repository will provide the all possible meanings of the word considered for search which will helpful in further processing of the system for optimizing the search result.

4.1.3 Extracted Sense List:

Here the extracted list of all meanings of the word is stored which would be supplied further for the classification.

4.1.4 Sense Matching and Classification :

The predefined domain would be compared with the extracted meanings of the incoming input and the relevant and most appropriate meaning of the word is differentiated that would be provided to the search engine for further processing.

4.1.5 Optimized Search Result:

The result would be the list of relevant data searched by the user.

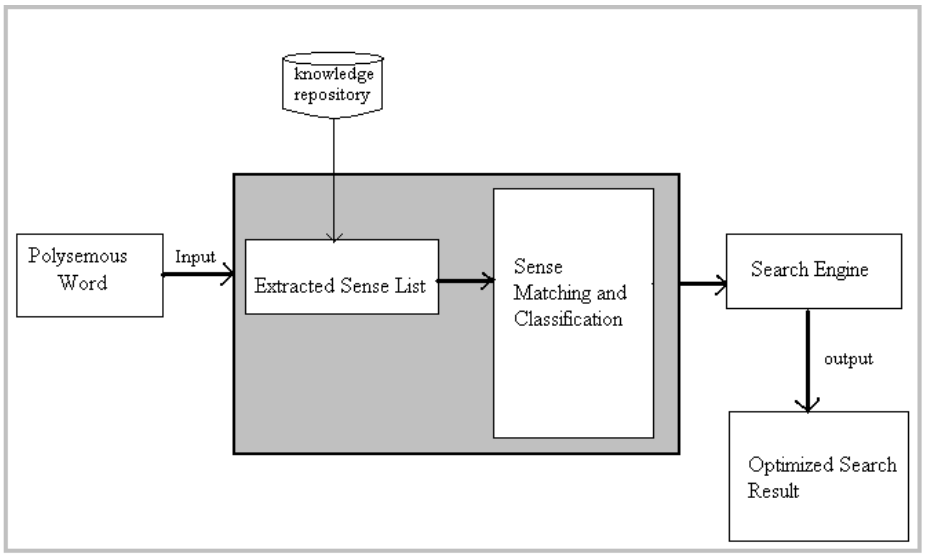

Figure 1: The Framework of the proposed technique

4.2 Ambiguity Resolution:

The word disambiguation is based on the predefined domain. This domain acts as the benchmark for evaluating and disambiguating the polysemous word. For optimizing the search result of an ambiguous word, effort is made to associate an appropriate meaning to the word, which will be further guiding the search. This will act as catalyst for producing optimized and relevant search result from user's perspective.

5 RESULTS

When the search made by other search engines like Google for the ambiguous word say "table", by carpenter/interior decorator or Software Professional following set of results is produced ordered according to decreasing pagerank by Google.

Table 1: Result shown by Google

Search Result
HTML table tag
Tables in HTML documents
Table (database) - Wikipedia, the free encyclopedia
Images for table
India Coffee Table, India Coffee Table ... - India - Alibaba.com

When the search made by a carpenter/interior decorator through the proposed system for the same word "table", following search result would be shown ordered according to decreasing Google's pagerank.

Table 2: Result shown by proposed system to carpenter/interior decorator

Search Result
Table (furniture) - Wikipedia, the free encyclopedia
Office Chair Office Desk Conference Table Furniture
China Seating, Table & Furniture, Seating, Table ... - Made in China
Table Furniture | Decorative Furniture| Living Room Furniture
Accent Table, Accent Furniture

When the search made by a Software Professional through the proposed system for the same word "table", following search result would be shown ordered according to decreasing Google's page-rank.

Table 3: Result shown by proposed system to Software Professional

Search Result
HTML table tag
HTML Tables
Table (database) - Wikipedia, the free encyclopedia
Ascii Table - ASCII character codes and html, octal, hex and decimal
S.O.S. Math - Mathematical Tables and Formulas

Thus, according to the user's interest, search gets molded and showing only relevant results skipping irrelevant results and thus ensuring better search compared to other search engines.

6 CONCLUSIONS

The proposed technique would be able to produce a more optimized search result as compared to other search engines. The proposed technique would be able to handle the noun polysemous words efficiently and can easily disambiguate the word on the basis of domain and hence resulting in an intelligent search.

REFERENCES

1. Nancy Ide and Jean V´eronis. Introduction to the special issue on word sense disambiguation: the state of the art. Comput. Linguist. 24(1):2–40, 1998.

2. Philip Resnik, David Yarowsky, A Perspective on Word Sense Disambiguation Methods and their Evaluation,http://www.cs.jhu.edu/~yarowsky/pubs.html.
3. George A. Miller. Wordnet: a lexical database for english. Commun. ACM, 38(11):39–41, 1995.
4. Christopher D. Manning and Hinrich Sch¨utze. Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA, 1999.
5. Daniel Jurafsky and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.
6. David Yarowsky. One sense per collocation. In HLT '93: Proceedings of the workshop on Human Language Technology, pages 266–271, Morristown, NJ, USA, 1993. Association for Computational Linguistics.
7. Gale W, Church KW, Yarowsky D. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. Proceedings of the ACL, 1992; 30:249-256.
8. Weiss SF. Learning to disambiguate. Information Storage and Retrieval, 1973; 9:33-41.
9. Kelly E, Stone P. Computer recognition of English word senses. North-Holland Publishing Co., Amsterdam,1975.
10. Small S, Rieger C. Parsing and comprehending with word experts (a theory and its realisation). In: Strategies for Natural Language Processing, Lehnert WG, Ringle MH (Eds), LEA, 1992, pp 89-148.
11. Lesk M. Automatic sense disambiguation: how to tell a pine cone from an ice cream cone. Proceedings of the SIGDOC Conference 1986; 24-26.
12. Cowie J , Guthrie J, Guthrie L. Lexical disambiguation using simulated annealing. Proceedings of COLING Conference, 1992;359-365.
13. Black E. An experiment in computational discrimination of English word senses. IBM Journal, 1988; 32:185-194.
14. Wallis P. Information retrieval based on paraphrase. Proceedings of PACLING Conference, 1993.
15. Demetriou GC. Lexical disambiguation using constraint handling in Prolog (CHIP). Proceedings of the European Chapter of the ACL, 1993; 6:431-436.

16. Zernik U. TRAIN1 vs. TRAIN2: Tagging word senses in corpus. Proceedings of RIAO 91, Intelligent Text and Image Handling, 1991; 567-585.
17. Hearst MA. Noun homograph disambiguation using local context in large text corpora. Proceedings of the 7th conference, UW Centre for the New OED & Text Research Using Corpora, 1991; 7.
18. Dagan I, Itai A, Schwall U. Two languages are more informative than one. Proceedings of the ACL, 1991:29:130-137.
19. Church KW. Using bilingual materials to develop word sense disambiguation methods. Proceedings of ACM SIGIR Conference, 1992; 15: 350.
20. Voorhees EM. Using WordNet to disambiguate word sense for text retrieval. Proceedings of ACM SIGIR Conference, 1993; 16:171-180.
21. Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network. Proceedings of CIKM, 1993.
22. Yarowsky D. Word sense disambiguation using statistical models of Roget's categories trained on large corpora.