International Journal of Latest Trends in Engineering and Technology
IJLTET Special Issue- ICRACSC-2016 , pp.063-066
e-ISSN: 2278-621X

WORD SENSE DISAMBIGUATION USING LESK

B. Surekha[1], Dr. K. Vijaya kumar[2] and S. Siva skandha[3]

Abstract- Word Sense Disambiguation (WSD) is an open problem of natural language processing and ontology. WSD is identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings. Disambiguating a word needs two things: Dictionary having a list of senses of ambiguous word i.e. semantic relations of a polysemous word and Corpus (Real World Text) consisting real world knowledge. It is difficult for system or even to human being to identify the correct sense without a sense repository i.e. knowledge sources. There are two types of knowledge sources, one is corpus and another one is WordNet. This paper Lesk algorithm is used to resolving ambiguity in a sentence which is based on integrating through WordNet.
Keywords- Word Sense Disambiguation, Information extraction and retrieval, WordNet.

I. INTRODUCTION

Now a day's everyone is searching any kind of data into the internet to disambiguate. The search engines sometimes give data relevant to the context and sometimes give irrelevant data. Such things happen because, if the query may contain ambiguous words or word having several possible meanings. So that, the search engines may not give the relevant data. Word Sense Disambiguation (WSD) [1] is the process for identification of several meaning of ambiguous words based on distinct situations. For example the word "cold" contains different senses, one is the 'disease', another one is the 'temperature', 'climate' and so on. The process of identification to decide appropriate meaning of an ambiguous word in a particular context is known as Word Sense Disambiguation. Disambiguating a word needs two things: Dictionary having a list of senses of ambiguous word i.e. semantic relations of a polysemous word and Corpus (Real World Text) consisting real world knowledge. It is difficult for system or even to human being to identify the correct sense without a sense repository i.e. one or more type of knowledge sources. There are two types of knowledge sources, first is corpus which is tagged or untagged with the word sense, and former is dictionaries.
Word Sense Disambiguation has been taken several approaches, they are knowledge based/ dictionary based approach, supervised approach, unsupervised approach and semi supervised approach. This paper describes Lesk algorithm to disambiguate the word in a context through WordNet.

II. LITERATURE SURVEY

A. WORD SENSE DISAMBIGUATION:

Word sense disambiguation is the process of finding the correct sense of a word depending on its context. A word sense is a correct meaning of a word. Consider the following two sentences, One is "I went to the bank to deposit my money" and other is "The river bank was full of dead fishes". The word BANK is used in two senses. One is of type of 'financial institution' and other is related to 'sloping land'. Selection of appropriate word sense is one of the elements. As without knowledge, it is impossible both for human being and computer to identify correct meaning so for that Knowledge sources are used. This knowledge source is of two types, one is corpus which is either unlabelled or annotated with word senses, and other is dictionaries like machine readable dictionaries [2]. WSD approaches or methods are as follows.

Knowledge base or Dictionary based approach:

[1] *Department of Computer Science & Engineering, CMR College of Engineering and Technology, Kandlakoya, Medchal, Hyderabad, Telangana, India,*
[2] *Department of Computer Science & Engineering, CMR College of Engineering and Technology, Kandlakoya, Medchal, Hyderabad, Telangana, India,*
[3] *Department of Computer Science & Engineering, CMR College of Engineering and Technology, Kandlakoya, Medchal, Hyderabad, Telangana, India,*

Knowledge based approach, often refer as dictionary based approach uses lexical knowledge bases such as dictionaries like WordNet [3], thesauri, ontology etc and acquire information related to word from word definition and relations present the respective knowledge base[2,4].
Agirre, Eneko & German Rigau (1996) [5] proposed Word Sense Disambiguation with Conceptual Density method which uses lexical knowledge base. This method's basic idea is to select a sense based on the conceptual distance i.e. how the ambiguous word and its context words are related. First find the noun in context then its senses and relations majorly the hypernym. This result is later extended by the same researcher i.e Agirre, Eneko & David Martinez (2001) suggested to finding the correct sense use Selectional preferences method [6]. This method look for the probable associations between word categories, the simplest measure for this word to word relation is frequency count.
Overlap based approaches like Lesk, Extended Lesk are purely based on the matching of word and context words. This approach is suggested by Satanjeev Banerjee, Ted Pedersen, 2002 [7]. Basic problems with this approach is it is heavily depends on dictionaries, which is also having some restrictions over acquiring the common sense knowledge. Methods treat a dictionary as both the source of the sense inventory as well as a repository of information about words that can be exploited to distinguish their meanings in text. WordNet as lexical database or sense inventory to access meanings and other information related to words.

Supervised approach:

Supervised approach makes use of sense annotated corpora to train from. These approaches use machine learning techniques to learn a classifier from labeled training sets. Some of the common techniques used are decision lists, decision trees, naive bayes, neural networks, support vector machines (SVM).

Unsupervised approach:

Unsupervised approach makes use of only raw annotated corpora and do not exploit any sense-tagged corpus to provide a sense choice for a word in context. These methods are context clustering, word clustering and co-occurrence graphs.

Semi supervised approach:

Semi supervised approach makes use of secondary source of knowledge such as small annotated corpus as seed data in bootstrapping process. It actually overcomes the main problems associated with building a classifier: the lack of annotated data and the data sparsity problem.

B. WORDNET:

WordNet is the most popular as a resource to be used in knowledge-based approach to disambiguate the meanings of polysemy words. WordNet is a large lexical database of English language [3]. WordNet is a machine-readable dictionary developed by George Miller and his colleagues at the Cognitive Science Laboratory at Princeton University.

WordNet are arranged semantically instead of alphabetically. Synonymous words are grouped together into synonym sets, called synsets. Each such synset represents a single distinct sense or concept. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet is also freely and publicly available for download. After the development of English WordNet, many other WordNet on other languages such as Spanish WordNet, Italian WordNet, Hindi WordNet etc were built. These WordNet are used as one important resource to disambiguate the different meanings of a polysemy words in respective languages. To disambiguate the meaning of a polysemy word using the WordNet, the related words from synset, gloss and different levels of hypernym are collected from the WordNet database and these related words are compared to find the overlaps using WSD. In this proposed system all the ambiguous words synset are displayed and identify the correct sense definition in a context.

III. PROPOSED SYSTEM

The proposed system uses Lesk algorithm to disambiguate the word definition through WordNet. The proposed system architecture shows figure 1.

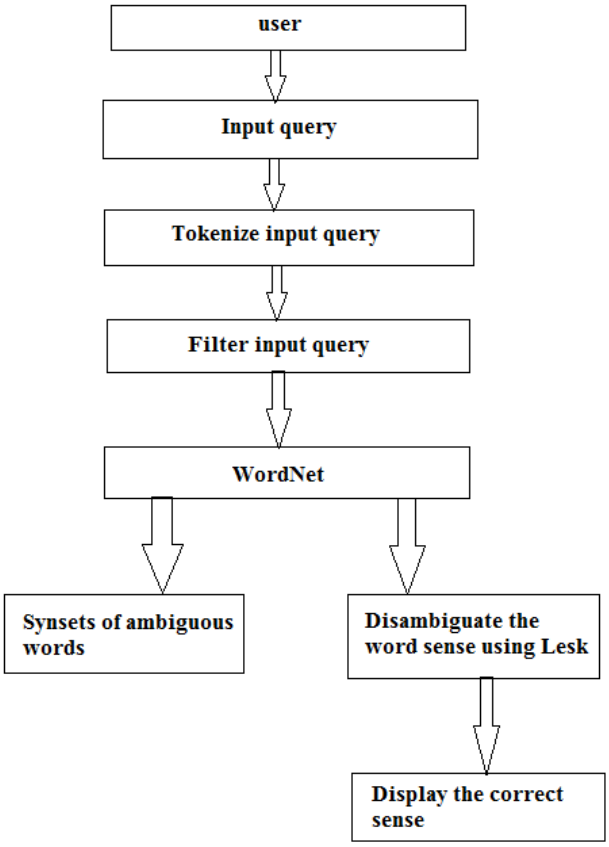

Figure 1: 'WSD Using Lesk' architecture

This proposed system takes input query from the user. Next input query is tokenized into individual words by using the tokenizer from nltk.corpus package. Next all the stop words are removed and display only ambiguous words from the input query. Next list of ambiguous words are used to retrieve synsets of those ambiguous words by using WordNet. Finally, Lesk algorithm in that simplified Lesk algorithm is used to disambiguate the word and display the correct definition of each ambiguous word from the context.

A. LESK ALGORITHM:

The Lesk algorithm is a classical algorithm for word sense disambiguation introduced by Michael E. Lesk in 1986. The Lesk algorithm is based on the assumption that words in a given "neighborhood" (section of text) will tend to share a common topic. A simplified version of the Lesk algorithm is to compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood. Versions have been adapted to use WordNet. An implementation might look like this:

- For every sense of the word being disambiguated one should count the amount of words that are in both neighborhood of that word and in the dictionary definition of that sense.
- The sense that is to be chosen is the sense which has the biggest number of this count.

To disambiguate a word, the gloss of each of its senses is compared to the glosses of every other word in a phrase. A word is assigned to the sense whose gloss shares the largest number of words in common with the glosses of the other words.

As an example consider the words pine & cone. The dictionary meanings are as follows:

PINE:

1. Kinds of evergreen tree with needle-shaped leaves
2. Waste away through sorrow or illness

CONE:

1. Solid body which narrows to a point
2. Something of this shape whether solid or hollow
3. Fruit of certain evergreen trees

As can be seen, the best intersection is Pine #1 ⋂ Cone #3 = 2. This value is calculated by the Lesk algorithm.

Two variations of the Lesk algorithm are present. The first counted the number of words in common between the instance in which the target word occurs and its gloss. Each word count was weighted by its inverse document frequency which was defined simply as the inverse of the number of times the word has occurred in the instance or the gloss in which the word occurs. The gloss with the highest number of words in common with the instance in which the target word occurs represents the sense assigned to the target word. A second approach proceeded identically, except that it added example texts that WordNet provides to the glosses.

B. SIMPLIFIED LESK ALGORITHM:

The most important and frequently used variant of the Lesk algorithm is the simplified Lesk algorithm with is implemented in the application. In Simplified Lesk algorithm, the correct meaning of each word in a given context is determined individually by locating the sense that overlaps the most between its dictionary definition and the given context. Rather than simultaneously determining the meanings of all words in a given context, this approach tackles each word individually, independent of the meaning of the other words occurring in the same context.

A comparative evaluation performed by Vasileseu et al. (2004)[8] has shown that the simplified Lesk algorithm can significantly outperform the original definition of the algorithm, both in terms of precision and efficiency.

The algorithmic description of simple Lesk algorithm may look like the following:

```
function SIMPLIFIED LESK (word, sentence) returns best sense of word
        best-sense <- most frequent sense for word
        max-overlap <- 0
        context <- set of words in sentence
        for each sense in senses of word do
                signature <- set of words in the gloss and examples of sense
                overlap <- COMPUTEOVERLAP (signature, context)
                if overlap > max-overlap then
                        max-overlap <- overlap
                        best-sense <- sense
end return (best-sense)
```

The COMPUTEOVERLAP function returns the number of words in common between two sets, ignoring function words or other words on a stop list. The original Lesk algorithm defines the context in a more complex way.

IV. CONCLUSION

This paper has described an algorithm to perform disambiguation of a word in given context using Lesk through WordNet. Lesk algorithm is very sensitive to the exact word definitions. So the absence of a certain word can radically change the results.

REFERENCES

[1] Word Sense Disambiguation.

[2] Navigli, roberto, “word sense disambiguation: a survey”, ACM computing surveys, 41(2), ACM press, pp. 1-69, 2009.

[3] Fellbaum. WordNet: An Electronic Lexical Database. MIT Press Cambridge, Massachusetts, 1998.

[4] Ping Chen and Chris Bowes, University of Houston-Downtown and Wei Ding and Max Choly, University of Massachusetts, Boston Word Sense Disambiguation with Automatically Acquired Knowledge, 2012 IEEE INTELLIGENT SYSTEMS published by the IEEE Computer Society.

[5] Agirre, Eneko & German Rigau. 1996. "Word sense disambiguation using conceptual density", in Proceedings of the 16th International Conference on Computational Linguistics(COLING), Copenhagen, Denmark, 1996.

[6] Agirre, Eneko & David Martínez. 2001. “Learning class-to-class selectional preferences” in Proceedings of the Conference on Natural language Learning, Toulouse, France, 15–22.

[7] Satanjeev Banerjee, Ted Pedersen, “An adaptive Lesk Algorithm for Word Sense Disambiguation Using WordNet”, Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, page no: 136-145, 2002.

[8] Florentina Vasilescu, Philippe Langlais, and Guy Lapalme. 2004. Evaluating Variants of the Lesk Approach for Disambiguating Words. LREC, Portugal.