Word Sense Disambiguation for Indian Languages: A brief Survey

Ms. Sonia¹ and Dr. Neetu Sharma²
¹M.Tech Scholar, GITAM Kablana, Jhajjar(Haryana)
²HOD, Deptt. Of Computer science and engineering, GITAM Kablana, Jhajjar(Haryana)

Abstract — In this paper, we made a survey on Word Sense Disambiguation (WSD) using some of the Indian languages. A research in WSD has been conducted upto different extents. In this paper, we have gone through a survey regarding the different approaches adopted in different research works, the State of the Art in the performance in this domain, recent works in different Indian languages such as Bengali, Hindi, Punjabi language. We have made a survey on different competitions in this field and built a comparison according to results, obtained from those competitions.

Keywords- NLP, WSD, AI, WORDNET

I. INTRODUCTION

Lexical Ambiguity Resolution or Word Sense Disambiguation (WSD) is the problem of assigning the appropriate meaning (sense) to a given word in a text or discourse where this meaning is distinguishable from other senses potentially attributable to that word. Thus, a WSD or Word Sense Tagging system must be able to assign the correct sense of a given word, for instance age, depending on the context in which the word occurs.

Many Natural languages like English, Hindi, Punjabi, French, Chinese, etc. are the languages which have some words whose meaning are different for same spelling in the different context. In English, Words likes Bark, Lie, book, etc. can be considered example of polysemous words. Human beings are blessed with the learning power. They can easily find out what is the correct meaning of a word in a context. But for computer it is a difficult task. So, we need to develop an automatic system which can perform like humans do i.e. the system which can find out the correct meaning of the word in particular context.

Research has progressed steadily to the point where WSD systems achieve consistent levels of accuracy on a variety of word types and ambiguities. A rich variety of techniques have been researched, from dictionary-based methods that use the knowledge encoded in lexical resources, to supervised machine learning methods in which a classifier is trained for each distinct word on a corpus of manually sense-annotated examples, to completely unsupervised methods that cluster occurrences of words, thereby inducing word senses. Among these, supervised learning approaches have been the most successful algorithms to date.

1.1 MOTIVATION

Human language is ambiguous, so that many words can be interpreted in multiple ways depending on the context in which they occur. For instance, consider the following sentences:
(a) I can hear bass sounds.
(b) They like grilled bass.

The occurrences of the word bass in the two sentences clearly denote different meanings: low-frequency tones and a type of fish, respectively.

Unfortunately, the identification of the specific meaning that a word assumes in context is only apparently simple. While most of the time humans do not even think about the ambiguities of language, machines need to process unstructured textual information and transform them into data structures.
which must be analyzed in order to determine the underlying meaning. The computational identification of meaning for words in context is called \textit{word sense disambiguation (WSD)}. For instance, as a result of disambiguation, sentence (b) above should be ideally sense-tagged as “They like/\textit{ENJOY} grilled/\textit{COOKED} bass/\textit{FISH}.”

WSD has been described as an AI-complete problem [Mallery 1988], that is, by analogy to NP-completeness in complexity theory, a problem whose difficulty is equivalent to solving central problems of \textit{artificial intelligence} (AI), for example, the Turing Test [Turing 1950]. Its acknowledged difficulty does not originate from a single cause, but rather from a variety of factors.

First, the task lends itself to different formalizations due to fundamental questions, like the approach to the representation of a word sense (ranging from the enumeration of a finite set of senses to rule-based generation of new senses), the granularity of sense inventories (from subtle distinctions to homonyms), the domain-oriented versus unrestricted nature of texts, the set of target words to disambiguate (one target word per sentence vs. “all-words” settings), etc.

Second, WSD heavily relies on knowledge. In fact, the skeletal procedure of any WSD system can be summarized as follows: given a set of words (e.g., a sentence or a bag of words), a technique is applied which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context. Knowledge sources can vary considerably from corpora (i.e., collections) of texts, either unlabeled or annotated with word senses, to more structured resources, such as machine readable dictionaries, semantic networks, etc. Without knowledge, it would be impossible for both humans and machines to identify the meaning, for example, of the above sentences.

\textbf{II. LITERATURE REVIEW}

Navigli and Crisafulli [29] suggest WSD methods for Bengali language. They have developed a word sense disambiguation(WSD) system for Bengali language and applied the system to get correct lexical choice in Bengali-Hindi machine translation. They are not aware of any existing system for Bengali WSD. Since there is no sense annotated Bengali corpus or sufficient amount of parallel corpus for Bengali-Hindi language pair, they had to use an unsupervised approach. They used a graph based method to find sense clusters in Bengali language. Following this they use a vector space based approach to map these sense clusters to Hindi translations of the target word and to predict translation of the target word in test instance. They used monolingual Bengali and Hindi corpora and the available Bengali and Hindi WORDNET and bilingual sense dictionary.

Agarwal et al. [32] describe the increased demand of Hindi language in today’s world. It is important to judge the correct sense of the polysemous words for the proper processing of the language. The author derives the target word from Hindi Word net which is developed at IIT Bombay. We are aware with one WSD system for Bengali language (Ayan Das and Sudeshna Sarkar), which is applied to the system to get correct lexical choice in Bengali-Hindi machine translation. Pushpak Bhattacharyya [28], The Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Hindi WordNet is inspired by the famous English

WordNet. Search Engines are the basic tool of fetching the information on the web. The IT revolution not only affected the technocrats, but the native users are also affected. This leads to the need of effective search engines to fulfill native user’s needs and provide them information in their native languages. The major population of India use Hindi as a first language. The Hindi language web information retrieval is not in a satisfactory condition. Besides the other technical setbacks, the Hindi language search engines face the problem of sense ambiguity. Their WSD method is based on Highest Sense Count (HSC). It works well with Google. The objective of the paper is comparative analysis of the WSD algorithm results on the three Hindi language search engines- Google, Raftaar and Guruji.
They have taken a test sample of 100 queries to check the performance level of the WSD algorithm on various search engines. The results show promising improvement in performance of Google search engine whereas the least performance improvement was there in Guruji search engine.

WSD also used for Punjabi language. The Punjabi language is morphologically rich. Rakesh and Ravinder [33] have given the WSD algorithm for removing ambiguity from the text document. WSD algorithm used by authors is Modified Lesk’s Algorithm. Their work deals with analyzing the correct meaning of the ambiguous word(s) in Punjabi language. Not much work has been done in this field which deals with the Punjabi language. A text with multiple senses in natural language is open problems of Natural Language Processing (NLP) which can be resolving using WSD. The Supervised learning methodology is used for this purpose which is the conventional approaches to WSD. The semantic lexicon for the various languages of India i.e., Indo WordNet has been used to obtain the sense definition of the Punjabi language.

For the first time, Haroon[31], R.P. (2010) has given the first attempt for an automatic WSD in Malayalam and Richard Singh and K. Ghosh [26] have given a proposed architecture for Manipuri Language in 2013. The system performs WSD in two phases: training phase and testing phase.

WSD is mainly used in Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT), Content Analysis, Word Processing, Lexicography and Semantic Web. Finally, WSD used for many Indian languages Hindi, Malayalam, Manipuri, Punjabi, Kannada and Bengali.

### III. WSD FOR INDIAN LANGUAGES

Various works on WSD are implemented in English and other European languages but less amount of work has been done in Indian languages due to large variety of morphological inflections and lack of different sense inventories, machine readable dictionaries, knowledge resources, which are required for WSD algorithms. The work in various Indian Languages are described below.

#### 3.1 WSD FOR BENGALI

Graph-based approaches have been quite successful in unsupervised word sense disambiguation, so we decided to work on graph-based WSD system for Bengali. We have studied the performance of two successful WSD methods suggested by Navigli and Crisafulli(2010) and Jurgens(2011) in Bengali. In this approach (Navigli and Crisafulli, 2010), the corpus is queried with a target word and a cooccurrence graph is constructed from the contexts after removal of the target word from the contexts. The main idea behind this approach is that edges in the cooccurrence graph participating in cycles are likely to connect vertices (i.e. words) belonging to the same meaning component. The work focuses on cycles of length 3 (triangle) and 4 (square). The edge weights are equal to the Dice-coefficient of cooccurrence of two words in the retrieved context set and edges with weight below a threshold are removed. Each of the remaining edges are assigned weight equal to the triangle and square scores and edges with score below a threshold value are removed. Finally, all the connected components of size greater than a threshold are identified and each such component is assumed to contain words that together indicate a distinct sense of the target word.

Jurgens (2011) proposed a community detection algorithm from a cooccurrence graph constructed from the nouns in the corpus that occur with frequency greater than a threshold. Initially, similarity between each edge pair is computed by a scoring function which equals zero if the edges do not share any vertex and is the ratio of the number of common neighbors and the total number of neighbors of the two vertices apart from the common vertex. Finally, the edges are agglomeratively clustered by single-link criteria. The construction of the dendrogram is stopped when the sum of the edge density in the clusters is highest.
3.2 WSD FOR HINDI

Pushpak Bhattacharyya, The Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Hindi WordNet is inspired by the famous English WordNet. In the Hindi WordNet the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set, or synset, in the Hindi WordNet, representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings. Synsets are the basic building blocks of WordNet. The Hindi WordNet deals with the content words, or open class category of words. Thus, the Hindi WordNet contains the following category of words: Noun, Verb, Adjective and Adverb.

3.3 WSD FOR MANIPURI

For the first time, Richard Singh and K. Ghosh [26] have given a proposed architecture for Manipuri Language in 2013. Manipuri is a Tibeto-Burman language, spoken in the valley of Manipur, a North-Eastern state of India. The Manipuri word sense disambiguation system contains the following steps: (i) preprocessing, (ii) feature selection and generation and (iii) training, (iv) testing and (v) performance evaluation.

Raw Data is processed in the order to get the features which can be used for training and testing data efficiently. In feature Selection, a total of 6 features are taken to build feature: (i) the focus word for which the sense is to be derived, (ii) the normalized position of the word in the sentence, (iii) the previous word, (iv) the previous-to-previous word, (v) the next word, (vi) the next to next word. A 5-gram window is formed using the pair of the focus word and its context words which forms the context information. A focus word, based on the context may have different senses. Hence, in order to disambiguate the sense of the focused word, the contextual information is very much necessary and helps in predicting the correct one.

3.4 WSD FOR MALAYALAM

It is one of the 22 official languages of India, and it is used by around 36 million people. It has given the first attempt for an automatic WSD in Malayalam.

Malayalam is a Dravidian language, mostly spoken at Kerala, a southern state of India. Haroon, R.P. (2010) has given the first attempt for an automatic WSD in Malayalam. The author has used the Lesk and Walker algorithm. In this algorithm, the collection of the contextual words is prepared for a target word. Next, different bags, containing few words of specific sense are generated from the Knowledge source. After that, the overlap between the contextual words and the bags are measured. A score of 1 is added to that sense, if any overlap is there. Highest score for a sense is selected as the winner.
3.5 WSD FOR PUNJABI

Punjabi language is world’s 12th most widely spoken language. Punjabi Language is used in both parts of Punjab, in India and also in Pakistan. Punjabi is syllabic in nature. It consists of 41 consonants called vianjans, 9 vowel symbols called laga or matras and 2 symbols for nasal sounds (. , °).

The Punjabi language is a morphologically rich language. Rakesh and Ravinder [27] have proposed a WSD algorithm for removing ambiguity from the text document. The authors used the Modified Lesk Algorithm for WSD. Two hypotheses have been considered in this approach. First, the cooccurring words in a sentence are be disambiguated by assigning the most closely related senses to them. The second hypothesis is considered as, the definitions of related senses have maximum overlap.

IV. RESULTS

Table 1: A comparative analysis of WSD in five Indian languages

<table>
<thead>
<tr>
<th>Type of algorithm</th>
<th>Author/s</th>
<th>Language</th>
<th>Performance</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic Algorithm</td>
<td>Sabnam Kumari Prof. Paramjit Singh</td>
<td>Hindi</td>
<td>91.6%</td>
<td>2013</td>
</tr>
<tr>
<td>Decision Tree based WSD System</td>
<td>Sivaji Bandyopadhyay and group</td>
<td>Manipuri</td>
<td>71.75%</td>
<td>2014</td>
</tr>
<tr>
<td>Modified Lesk’s Algorithm</td>
<td>Rakesh and Ravinder</td>
<td>Punjabi</td>
<td>75%</td>
<td>2011</td>
</tr>
<tr>
<td>Knowledge based Approach</td>
<td>Rosna P Haroon</td>
<td>Malayalam</td>
<td>81.3%</td>
<td>2010</td>
</tr>
<tr>
<td>Un-Supervised Graph-based Approach</td>
<td>Ayan Das, Sudeshna Sarkar</td>
<td>Bengali</td>
<td>60%</td>
<td>2013</td>
</tr>
</tbody>
</table>
V. CONCLUSION

In this paper, we made a survey on WSD in different international and Indian languages. The research work in those languages has been proceeded upto different extents according to the availability of different resources like corpus, tagged data set, WORDNET, thesauri etc..

In Asian languages, especially in Indian languages, due to large scale of morphological inflections, development of WORDNET, corpus and other resources are is under progress. Hindi language search engines are facilitating the users but only to some extent. The results after the sense disambiguation and expansion are compared for the three search engines Google, Raftaar and Guruji. The results show an overall increase of precision in all the three search engines. However if compared individually the highest improvement of precision is in Google, Raftaar shows an average increase in performance and Guruji shows the lowest increase in performance. The close observation identified that the performance of Guruji was also not very good with the original queries. In some cases it is found that the performance of Guruji deteriorates after disambiguation and query expansion. Hence we can conclude that the maximum benefit of the approach is in the case of Google and before disambiguation and query expansion its performance was better than the other two search engines.

REFERENCES


