International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5 Issue 1, January 2016

Word Sense Disambiguation in Hindi Language : A Survey

Riya Dhopavkar, Shreya Kakade, Shruti Kakade, Shashank Singh Yaduvanshi, Prof.Seema Ghondhalekar

Student,Computer Department,KKWIEER, Nashik, Maharashtra, India

***Abstract*—Word Sense Disambiguation (WSD) is an open problem of Natural Language Processing (NLP), which is the task of selecting meaning of the ambiguous word based on the context in which word occurs. In this paper, we made survey on WSD. There are three approaches available for WSD - supervised, unsupervised and knowledge based and also the combination of all the three approaches. Research in WSD has been conducted for different Indian languages and finally we made a survey on Hindi language, which is the National language of India. Comparison of various algorithms for WSD in Hindi language has been done. In this paper, Modified Lesk Approach is surveyed.**

Keywords—Ambiguity, Hindi language, Natural Language Processing, Word Sense Disambiguation.

I. INTRODUCTION

A single word can have different meanings in Hindi Language. Humans have the knowledge to judge the correct sense of a word by writing or reading the other words in the context. A computer program has no basis for knowing which one is appropriate, even if is obvious to human. Word Sense Disambiguation is a challenging problem in Natural Language Processing. It is a process of automatically giving appropriate meaning to ambiguous word in context.

In Table 1, the word '**पास**' is common in both the sentences but have different meaning. For human it is very easy to determine the correct sense of the word '**पास**' in both the sentences. In the first sentence the word '**पास**' refers to nearness. In the second sentence, the word '**पास**' refers to belongings. Word Sense Disambiguation will help to find correct sense of the ambiguous word.

Word Sense Disambiguation is considered as AI complete problem, that is task whose solution is at least as hard as themost difficult problem in artificial intelligence. WSD is the ability to identify the meaning of the words in context in a computational manner. The main approaches of WSD are as follows:

Context 1:	राजकाघरपासहै।
Context 2:	मेरेपासएकगायहै।

Table 1: Example of ambiguity

Supervised WSD, unsupervised WSD and knowledge based. The application where Word Sense Disambiguation (WSD) is used are machine translation, information retrieval, text processing, speech processing.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool, it is necessary to determine the time factor, economy and company strength.

Singh et al. [1] describes the effects of context window size, stop word elimination and stemming on WSD. The work uses Hindi word net to find the correct sense of Hindi ambiguous words. The basic approach to WSD is to overlap the context meaning of the target word with the dictionary meaning of the target word. Two methods of WSD are explained viz. knowledge based and corpus based.

Navigli et al. [2] provides a survey on WSD; it helps in solving the ambiguity of the words and provides a description of the task of finding the correct meaning of the word. Three different approaches i.e. supervised, unsupervised, knowledge based, and its applications are explained. The comparison of these three approaches has also been presented.

Tondon et al. [3]proposed an algorithm for WSD by counting the number of words shared between the sense definition and the context. The sense definition was obtained from a dictionary, so it only consisted of synonyms. The context was obtained by considering the words in the same sentence as the word to be disambiguated. Most approaches today that use a Knowledge Base have as root, the Lesk's Algorithm, with changes made to the semantic relations used to obtain the sense definitions and the use of efficient metrics to compare the sense definitions and the contexts.

Yadav et al. [4] proposes that Hindi Word Sense Disambiguation" that was the first attempt for an Indian language at automatic WSD. The use of the Wordnet for Hindi developed at IIT Bombay, which is a highly important lexical knowledge base for Hindi. The main idea is to compare the

ISSN: 2278 – 1323

context of the word in a sentence with the contexts constructed from the Wordnet and chooses the winner. The output of the system is a particular synset number designating the sense of the word.

Banerjee et al. [5] It provides an adaptation of Lesk's dictionary based word sense disambiguation algorithm. Lexical database is employed rather than is made using a standard dictionary as the source of glosses. Instead of using a standard dictionary as the source of glosses, the lexical database word net is employed. The Lesk algorithm is the prototypical approach and is based on detecting shared vocabulary between the definitions of words.

Patel et al. [6] gives survey on WSD in different international and Indian languages.The research work in those languages has been proceeded according to the availability of different resources like corpus, tagged data set, Word Net Comparison of supervised, unsupervised and knowledge based approach is given.

RadhikeSawhney et al. [7] gives idea about dynamic context window and gives brief explanation about Modified Lesk Approach.

III. WSD APPROACHES

A. Unsupervised WSD:

In this approach, no supervision is provided. It is divided into two type, type-based and token-based approach. The type-based approach disambiguates by clustering instances of a target word while token-based approach disambiguates by clustering context of an ambiguous word. Main disadvantage of this approach is that senses are not well defined.This technique uses un-annotated corpus. Performance of unsupervised WSD has been lower than other methods.

B. Supervised WSD:

In this approach, there are large number of algorithms for WSD and it uses machine learning techniques ford disambiguation. It makes use of sense-annotated corpus. Disadvantage of this approach is that it requires large sense-annotated data. It is not suitable for resource scarce language. It is better than unsupervised and knowledge based approach.

C. Knowledged based:

This approach is based on machine-readable dictionaries or sense inventories. It can also be used with corpus-based methods. Word-net is used for knowledge-based approach. It assumes that words used together in text are related to each other and that the relation among them can be observed in the definitions of that words and their senses. Two or more words are disambiguated by finding the pair of dictionary senses with the greatest word overlap in their dictionary definitions.

Approach	Supervised Approach	Unsupervised Approach	Knowledge-based Approach
Advantages	Training is provided to a classifier from labelled training sets.	It focusses on different knowledge sources to build their models.	It uses the available information in large lexical database such as Word-net.
Disadvantages	Due to large Training sets, it is quite expensive in terms of space and time.	The performance is not as good as the other approaches and the algorithms are difficult to implement.	The performance depends on the dictionary and it is overlap based.

Table 2: Comparison of WSD approaches

IV. MODIFIED LESK APPROACH

The Lesk algorithm is used to remove the disambiguity from Hindi words. The original Lesk approach used fixed size context window. The context window is nothing but the number of left and right words surrounding the target ambiguous word.The basic idea of the algorithm is to find overlap among the senses of the word, which are to be disambiguated. The assumption of Lesk algorithm is that the target word which has same meaning must have common topic in its neighborhood. The senses of target ambiguous word can be detected more accurately if the target word contains more left and right words and hence dynamic context window is used instead of fixed sized window in Modified Lesk approach.

The Modified Lesk approach first finds all the ambiguous word in the given file and stores them in an array. Then it includes removal of special symbols like ',' and '|' along with all the special tokens. Retrieve all the possible number of senses of the ambiguous word from a lexical database called as Hindi Word Net developed by Prof. Pushpak Bhattacharya, III Bombay. Now assign sense count to each defined sense. Perform overlapping of the context meaning and the dictionary meaning, if they overlap increment the instance count for each sense. The number of sense of given target word describes instance output. The correct sense of the target word is the sense with the maximum score.

ISSN: 2278 – 1323

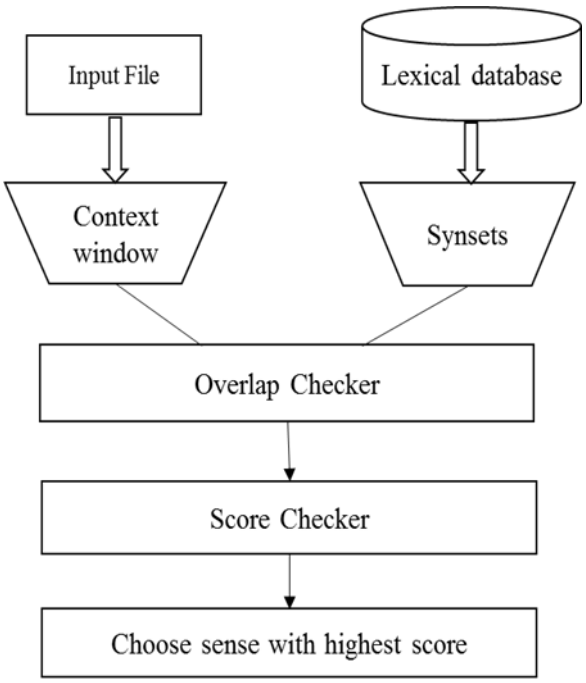

Figure 1: Lesk Approach

The proposed system will use a customized dictionary as a lexical resource. It will be trained to identify all the ambiguous words of Hindi Language and return possible senses of each word. If a word is not found, then it will assist the Hindi Word Net and store it with senses. It assigns a score to each sense by performing overlapping between context window and the given sense.

V. CONCLUSION

A modified Lesk approach is used in this paper to identify the correct sense of the ambiguous word. The comparison of the different approaches of WSD that are supervised approach, unsupervised approach and Knowledge-based approach is being done.We found the knowledge based approach as best approach and from that Modified Lesk Approach is used to carry out disambiguation of Hindi Language. If a word is not found, then it will assist the Hindi Word-Net and store the word with its senses. To judge the efficiency of the target word the context window size is taken dynamically. If the knowledge sharing between Hindi and other language will help to cross the language barrier among regional people. Lesk algorithm can be used for disambiguation and its application will encourage and enable knowledge sharing and translation.

VI. REFERENCES

[1] Satyendr Singh and Tanveer J. Siddiqui,"Evaluating Effect of Context Window Size, Stemming and Stop Word Removal on Hindi word sense Disambiguation," in Proc. IEEE, 2012.

[2] Roberto Navigli, "Word Sense Disambiguation:A survey," in ACM Comput. Surv. 41, 2, Article 10, pages 69, DOI =10.1145/1459352.1459355http://doi.acm.org/10.1145/1459532.1459355, February 2009

[3] Word Sense Disambiguation using Hindi Word Net, Rashish Tandon (Y6377) Advisor: Dr. Amitabha Mukherjee, Dept. Of Computer Science and Engineering, IT Kanpur, February 18, 2009.

[4] International Journal For Research In Applied Science And Engineering Technology (IJRASET), Study of Hindi Word Sense Disambiguation Based on Hindi WorldNet, Preeti Yadav, Mohd. Shahid Husain ,Department of Computer Science, Lucknow, India

[5] Satanjeev Banerjee and Ted Pederson, "AnAdapted Lesk Algorithm for Word Sense Disambiguation using Word Net, "in University of Minnesota, Duluth MN55812 USA, 2002.

[6] Nirali Patel1, Bhargesh Patel, Rajvi Parikh, Brijesh Bhatt, "A Survey: Word Sense Disambiguation", International Journal of Advance Foundation and Research in Computer (IJAFRC),January 2015.

[7] RadhikeSawhney and Arvinder Kaur, "A Modified Technique for Word Sense Disambiguation Using Lesk Algorithm in Hindi Language", International Conference on Advances in Computing,Communications and Informatics (ICACCI) IEEE, 2014.

Riya DhopavkarBE Computer K.K.Wagh College of Engineering Nashik
Shreya KakadeBE Computer K.K.Wagh College of Engineering Nashik
Shruti KakadeBE Computer K.K.Wagh College of Engineering Nashik
Shashank Singh YaduvanshiBE Computer K.K.Wagh College of Engineering Nashik

ISSN: 2278 – 1323