Word Sense Disambiguation for Hindi by Genetic Algorithm

Anidhya Athaiya, Deepa Modi, Vipin Jain
Department of Computer Science and Engineering, Swami Keshvanand Institute of Technology, Management and Gramothan, Jaipur-302017 (INDIA)
Email- anidhya.athaiya1409@gmail.com, deepa.modi22@gmail.com, ervipin.skit@gmail.com
Received 05.12.2018 received in revised form 24.02.2019, accepted 26.02.2019

Abstract: Word Sense Disambiguation (WSD) is a method of selecting the correct sense or meaning for a word in a context. WSD is a primary task in computational linguistics for language understanding applications such as information retrieval, question answering, machine translation, text summarization and many more. There have been many attempts on word sense disambiguation for English, but the amount of attempts for Hindi is inadequate. This paper proposes ongoing efforts on establishing a word sense disambiguation algorithm using Hindi WordNet developed at IIT Bombay as a base and intend to determine the correct sense of the given ambiguous word in Hindi language. A dynamic context window is used. Dynamic context window is the number of lefts and right word of ambiguous word. The cardinal idea behind this approach is that the target word which has the same meaning must have a common topic in its neighborhood.

Keywords: Words Sense Disambiguation, Hindi Wordnet, Natural Language Processing, Genetic Algorithm, ambiguous words.

1. INTRODUCTION

Word Ambiguity [1] is not something that we come across in day to day life perhaps excluding the context of jokes or when an ambiguous word is addressed in a sentence; we can understand the correct sense of that word without taking into account its possible different senses. However, in applications where machines have to process a natural language, ambiguity is a problem. A word can has more than one meaning in natural language. This is a stumbling block in natural language processing which can be resolved with the help of Word Sense Disambiguation. Word Sense Disambiguation [1] is a process of automatically giving relevant or appropriate meaning to a polysemous word within a given context. For example, a Hindi word " हल "is taken, and meaning is differentiated in these two contexts as follows,

Context1: इलाहाबाद उच्च न्यायालय की लखनऊ पीठ से तीस सितम्बर को आए फैसले के बाद अंसारी ने बातचीत से मसले के हल के लिए एकदम से तेजी पकड़ ली थी । अदालत से बाहर मन्दिर-मस्जिद विवाद को हल करने के लिए की जा रही बातचीत का विरोध करने वाले सुन्नी सेन्ट्रल वक्फ बोर्ड के जफरयाब जिलानी को भी अंसारी ने नहीं बख्शा । अंसारी ने उन्हें भी अनाप-शनाप कहा था । अंसारी के पलटी मार देने से यह तय हो गया है कि मामले के हल का एकमात्र रास्ता अब उच्चतम न्यायालय ही है।

Context2: हल एक कृषि यंत्र है जो जमीन की जुताई के काम आता है । इसकी सहायता से बीज बोने के पहले जमीन की आवश्यक तैयारी की जाती है । कृषि में प्रयुक्त औजारों में हल शायद सबसे प्राचीन है और जहाँ तक इतिहास की पहुँच है, हल किसी न किसी रूप में प्रचलित पाया गया है । हल से भूमि की उपरी सतह को उलट दिया जाता है जिससे नये पोषक तत्व उपर आ जाते हैं तथा खर-पतवार एवं फसलों की डंठल आदि जमीन में दब जाती है और धीरे-धीरे खाद में बदल जाते हैं । जुताई करने से जमीन में हवा का प्रवेश भी हो जाता है जिससे जमीन द्वारा पानी (नमी) बनाये रखने की शक्त बढ़ जाती है।

In both contexts, there is an ambiguous word “हल”. In context one the meaning of word “हल” represents “solution” or “समाधान“ whereas in context two the word “हल” represents “elder brother in law” or “जमीन जोतने का एक उपकरण”. Word sense disambiguation helps to judge the correct sense of an ambiguous word depending upon its context [3]. Hindi is considered the fourth most spoken language in the world. Among all researchers, most of the work is done for English language and many other natural languages, but work on Hindi language is minimal. This paper proposed a genetic algorithm based technique for Hindi word sense disambiguation [4]. Proposed work uses dynamic context window for finding the sense of the ambiguous word [5]. After this step genetic algorithm is applied to optimize the results to achieve the best quality or appropriate meaning. The paper is arranged in successive sections as

follows; related work is specified in section 2. Section 3 illustrates the proposed work for WSD using genetic algorithm. Section 4 describes results. Section 5 concludes this work and specifies the future work.

2. RELATED WORK

Word sense disambiguation is a current and arising research area. For Hindi language very limited research work is done. With every new research the accuracy of getting the correct sense of the polysemous word increases. Here are some of the research works that have already done so far.

Singh, S., & Siddiqui, T. J. [5] proposed a work that protrudes the context meaning of the target word from the dictionary meaning of the same word. They have used lesk algorithm in their approach for removing the ambiguity from ambiguous words.

Kumari, S., & Singh, D. P. [4] proposed a work that was the first attempt of using genetic algorithm (GA) for the Hindi word sense disambiguation. They suggested that genetic algorithm gives optimized result after disambiguation based on the fact that genetic algorithm makes sure that the best individual from each generation is assured to be in the next one. The algorithm has been applied to disambiguate nouns in a given sentence.

Sawhney, R., & Kaur, A. [7] used the dynamic context window that was described further as the modified lesk approach. This approach gives best overall performance for the case when both stemming and stop word elimination is already done. It is observed that as the size of the context window increases the precision of the algorithm is improved.

Jain, A., & Lobiyal, D. K. [8] presented an approach based on network agglomeration in which a graph is created for a given sentence, and the computation for the highest value of network agglomeration is determined.

Vaishnav, Z. B. [9] solves the problem of word sense disambiguation in Gujarati language by using the genetic algorithm in knowledge-based approach. A simple GA is applied in context after preprocessing of the whole context.

3. PROPOSED WORK

The proposed algorithm for Hindi word sense disambiguation comprises three stages as preprocessing, creation of context bag and applying genetic algorithm for output optimization. The complete approach for Hindi word sense disambiguation is shown in figure 1

3.1 Overview

Figure 1 describes the overall approach with the help of the flowchart. The amalgamation of a given word in context with a meaning which is different from meaning to that word is involved in Word Sense Disambiguation. The concept of the dynamic context window is the left and the right words of the target word. The work exhibited is done first by breaking all the words of the given sense and storing them which consists the removal of special tokens like ',' and '|' followed by all the specialized symbol and stop word. As the process continues the size of the context window increases. Although this compares the result with the static context window. The meaning of the words was fetched by the Hindi WordNet [10], considering all types of meanings like Hyponymy, Troponymy, Hyponymy this approach proceed using Onto_Nodes for all the words in context window. The comparison of static and dynamic context window is done. The concepts of onto_nodes provide the better result.

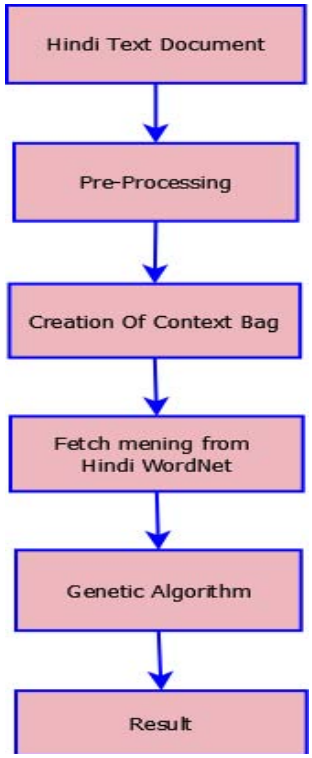

Figure 1: Proposed approach for Hindi word sense disambiguation

3.2 Algorithm

Step 1: Calculate the number of words in a sense. This covers the elimination of special tokens like ',' or '|' followed by all the specialized symbols.

Step 2: W[i] obtain the ambiguous word from a text file.
Step 3: W[k] will be left and right words of an ambiguous word.
Step 4: W[i], W[k] fetch meaning from Hindi WordNet[10].
Step 5: For both array determine the precision.

The algorithm moves ahead after the removal of stop word and the special characters, after the removal of stop words the whole paragraph is parted as words and stored in an array. Thus, the words having more than one meaning are acknowledged as the ambiguous words, i.e., W[i] (Static context) and the remaining words are considered as the W[k] (Dynamic context). The fetched meanings are compared, and the precision is determined.

Table 1. Array of Ambiguous word

Array	{" हल "}
Meaning from Hindi WordNet	{' समाधान, निबटारा, निपटारा, हल, निराकरण, अपाकरण ""}

Table 2. Remaining Words

Dynamic Context Window	Context 1	{ "इलाहाबाद उच्च न्यायालय लखनऊ पीठ तीस सितम्बर आए फैसले अंसारी बातचीत मसले "}
	Context 2	{' कृषि ,यंत्र ,जमीन ,जुताई बीज बोने जमीन इतिहास ..."}

4. RESULT

To check the exactness of the intended algorithm for Hindi WSD concluded experiments are executed and results are achieved for different domains. For measuring the performance of the algorithm, accuracy is used as the evaluation parameter. For testing the accuracy of the performed algorithm, data set of Hindi language is used from Indian Language Technology Proliferation and Deployment Centre (TDIL) [11] and due to the non-existing freely available large testing data a small testing data set is also created manually. The combination of both testing dataset contains data from various domains like sports, history literature, etc.

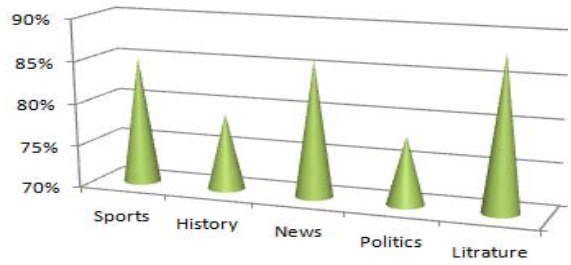

Figure 2: Accuracy for different domains for Hindi WSD.

Table 3. Experimental result generated for different nouns

Word	Synset	Answer	Comment
जेठ	अवधि(Period),व्यक्ति (Person),अवस्थासूचक (Stative)	Period	Correct
पानी	प्राकृतिक वस्तु (Natural Object), द्रव (Liquid), जातिवाचक संज्ञा (Common Noun), गुण (Quality), मानसिक अवस्थासूचक (Mental State)	Natural Object	Partially Correct
नजर	बोध (Perception), शारीरिक कार्य (Physical), अमूर्त (Abstract), गुणधर्म (property), घातक घटना (Fatal Event)	Perception	Correct
बरस	अमूर्त (Abstract), अवधि (Period),संप्रेषणसूचक (Communication), होना क्रिया (Verb of Occur), भौतिक अवस्थासूचक (Physical State)	Verb of Occur	Correct
मौसम	अवधि (Period), भौतिक अवस्था (Physical State)	Period	Correct

5. CONCLUSION

This paper uses dynamic context approach, and the matching is done by using the Onto_Nodes for every single word in sentence. Prior to performing the comparison stop word elimination is done. The sense having maximum appearance is considered as the correct sense of the target word. It is believed that as the size of context window increases the precision of algorithm gets better. From the results, therefore, it can be concluded that the sense of the ambiguous words can be detected more precisely as the target words contain more left and right words. For the proposed algorithm accuracy values range from about 75-80% for different domains. Our system, at present, will be dealing with nouns only.

REFERENCES

[1] N. Mishra, S. Yadav, & T. J. Siddiqui, "An unsupervised approach to Hindi word sense disambiguation", In Proceedings of the First International Conference on Intelligent Human Computer Interaction, Springer, New Delhi (2009), 327-335.

[2] R. Navigli, "Word sense disambiguation: A survey", ACM Computing Surveys (CSUR), (2009), 41(2), 10.

[3] R. Sharma and P. G. Bhatia, Word sense disambiguation for Hindi language (Doctoral dissertation), (2008).

[4] S. Kumari, D. P. Singh, "Optimized word sense disambiguation in Hindi using genetic algorithm", IJRCCT, (2013), 2(7), 445-449.

[5] S. Singh, and T. J. Siddiqui, "Evaluating effect of context window size, stemming and stop word removal on Hindi word sense disambiguation. In Information Retrieval & Knowledge Management (CAMP)", 2012 International Conference IEEE, Kuala Lumpur, Malaysia, (2012), 1-5.

[6] D. Modi, and N. Nain, "Part-of-Speech Tagging of Hindi Corpus Using Rule-Based Method", In Proceedings of the International Conference on Recent Cognizance in Wireless Communication & Image Processing, Springer, New Delhi. (2016), 241-247.

[7] R. Sawhney, and A. Kaur, "A modified technique for Word Sense Disambiguation using Lesk algorithm in Hindi language", In Advances in Computing, Communications and Informatics (ICACCI, 2014) International Conference on IEEE, New Delhi, India, (2014), 2745-2749.

[8] A. Jain, and D. K. Lobiyal, "Unsupervised Hindi word sense disambiguation based on network agglomeration", In Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on (pp. 195-200). IEEE (2015).

[9] Z. B. Vaishnav, "Gujarati Word Sense Disambiguation using Genetic Algorithm".

[10] P. Bhattacharyya, IndoWordNet: In The WordNet in Indian Languages, Springer, Singapore, (2017), 1-18.

[11] Indian Language Technology Proliferation and Deployment Centre,http://tdil-dc.in/index.php?lang=en