Knowledge-Based Method for Word Sense Disambiguation by Using Hindi WordNet

Pooja Sharma
Department of Computer Science
Banasthali University
Tonk, India
poojasharma_bu@outlook.com

Nisheeth Joshi
Department of Computer Science
Banasthali University
Tonk, India
jnisheeth@banasthali.in

***Abstract*—The purpose of word sense disambiguation (WSD) is to find the meaning of the word in any context with the help of a computer, to find the proper meaning of a lexeme in the available context in the problem area and the relationship between lexicons. This is done using natural language processing (NLP) techniques which involve queries from machine translation (MT), NLP specific documents or output text. MT automatically translates text from one natural language into another. Several application areas for WSD involve information retrieval (IR), lexicography, MT, text processing, speech processing etc. Using this knowledge-based technique, we are investigating Hindi WSD in this article. It involves incorporating word knowledge from external knowledge resources to remove the equivocalness of words. In this experiment, we tried to develop a WSD tool by considering a knowledge-based approach with WordNet of Hindi. The tool uses the knowledge-based LESK algorithm for WSD for Hindi. Our proposed system gives an accuracy of about 71.4%.**

Keywords-word sense disambiguation; Lesk; WordNet

I. INTRODUCTION

Ambiguity is one of the characteristics of a natural language. A word can have multiple meanings, varying with the context in which it is used. WSD has always been characterized as an AI-complete issue which is related to NP-completeness. WSD is a process to determine the exact sense of the word with respect to the context it is used in. Examples:

- उसने लाल कलम सें लिखा (Usne lal kalam se likha). Here, the '**कलम**' meaning is a red pen. Sentence example: He wrote with a red pen.
- गुलाब के फूल की कलम को माली नें काट दिया (Gulab ke phool ki kalam ko mali ne kaat diya). Here, the '**कलम**' meaning is graft cutting of rose. Sentence example*:* The gardener has cut the stem of a rose flower.
- नाई ने कलम को काटा (Naayi ne Kalm ko kaata). Here, the '**कलम**' meaning is graft cutting of hairs. Sentence example*:* The barber trims the temple.

WSD mainly relies on knowledge. The WSD system can be interpreted to work as follows: it takes as input an arrangement of words or a sentence, then NLP techniques are applied which utilize at least one source of knowledge to identify the most appropriate senses of the words regarding the context. Words possess different meanings that vary with the context in which they are used and our undertaking is to figure out which sense of word is intended in a particular context. This is one of the basic issues which are often experienced by any NLP framework. WSD is an area of exploration in NLP, which is gaining popularity nowadays. WSD makes use of a dictionary to achieve better identification of all word senses. Knowledge resources can also vary depending on corpora of textual content, labeled or unlabeled with word senses or more arranged resources like machine-readable dictionaries, semantic systems etc. Knowledge of the words is the key ingredient to distinguish the meaning whether we are referring to about humans or machines. WSD finds applications in many areas such as speech recognition (SR), information extraction (IE) and information retrieval (IR). WSD is a critical element for word knowledge. There are many different approaches to deal with WSD, for example, knowledge-based approach, selectional restriction, machine learning approach in order, which may further incorporate four types of approaches: supervised, unsupervised, semi-supervised, or hybrid. In this paper, we revolve around approaches like selectional restriction with Hindi WordNet for WSD in case of Hindi language.

II. RELATED WORK

Author in [1] proposed a knowledge-based WSD algorithm which selects the word that most overlaps with words in context. Many knowledge-based systems are based on this Lesk algorithm. Authors in [2] estimated semantic separation amongst lexemes and ideas. Authors in [3] tried to figure the most helpful knowledge sources and whether their mix prompts enhanced outcomes. Authors in [4] elaborated the SSI algorithm, which is a structural algorithm for WSD used for pattern matching. This algorithm can be refined to get high precision (possibly low recall), a benefit that will be more sensible for viable applications using WSD. As of now, the framework is tuned for disambiguating nouns by extending semantic graphs related to verbs. Authors in [5] offered a new graph-based method, which uses knowledge in LKB (based on WordNet) to do unsupervised WSD. They showed that the algorithm can easily be used with good results in other

Corresponding author: Pooja Sharma

languages, with the only requirement of having a WordNet. Authors in [6] investigated the WSD frameworks to particular areas using knowledge-based techniques, based on the graph based WSD framework that uses the data in WordNet. Three corpora were utilized i.e. one adjusted corpus (BNC), and two area particular corpora. The outcomes demonstrated that in every one of the three corpora, the knowledge-based WSD calculation enhances the past outcome and gets better results by using words related to the context than actual context words on the domain-based datasets. The outcomes are better for an area dependent corpus than for a general one, raising the chances for improving current WSD frameworks when connected to particular domains. Authors in [7] tried to develop WordNet for Punjabi by using extension approach from Hindi WordNet under the Indradhanush WordNet Project. They explained the morphology, symbols, origin and syntactic features of Punjabi language and lexical semantic relations used in Punjabi WordNet. Authors in [8] presented an algorithm for domain specific all-word WSD with an approximate accuracy of 69% (F1-score) with the domain corpus baseline standing at 65. Authors in [9] dealt with graph-based algorithms for WSD. They reported results on standard data sets, and demonstrated that graph-based approach performs comparably to the state of the art.

Author in [10] utilized a knowledge-based way to deal with WSD with Hindi language. Dictionaries and thesauri were the outside lexical assets used. Appropriate senses were assigned to a word in Hindi by using statistical strategy. Their work depicts the use of selectional restriction for Hindi language which is again a knowledge-based approach. Authors in [11] used the genetic algorithm including elitism to find the appropriate meanings of polysemous nouns in the given context. A lexical knowledge base for Hindi was used. The obtained outcomes depict the optimized disambiguation of Hindi words. Based on the personalized page rank on the generated graph using WordNet, authors in [12] presented an unsupervised WSD approach. It is produced by contributing substance words to the WordNet graph and associating to that synset in which they are visible in the form of strings. Then, the personalized page rank (PPR) algorithm was used to calculate the corresponding weight of the synset according to their relative structural significance, and as a result, for each content word, synset with the highest PPR weight is taken as the appropriate meaning. Authors in [13] further enhanced Lesk by using word embeddings to calculate the similarity between sense definitions and words in the context. Authors in [14] presented a graph-based unsupervised WSD framework which amplifies complete combined probability of all the senses in the context by displaying the WSD problem as a Markov random field developed utilizing the WordNet and a dependency parser and utilizing a maximum a posteriori (MAP) query for deduction.

Knowledge-based systems have been proposed for WSD for any low asset language [15]. As knowledge-based, Bengali WordNet has been utilized as a part of this work. The conventional Lesk algorithm has been utilized as the standard information-based approach.

III. PROPOSED METHODOLOGY

WSD is a process for assigning the appropriate sense of an ambiguous word according to context. A fixed class in WSD is portrayed by knowledge-based methods. Such strategies for WSD are generally connected to all words in a continuous text, in-spite of corpus-based methods, which are available for those words only for which annotated corpora are accessible. WSD makes use of dictionaries or thesauri to achieve better identification of word senses. Knowledge resources can also vary depending on corpora of textual content, labeled or unlabeled with word senses or more arranged resources like machine-readable dictionaries, semantic systems etc. Knowledge of the words is the critical ingredient to distinguish the meaning. WSD finds its applications in many areas such as information retrieval (IR), information extraction (IE), and speech recognition (SR). WSD has seen different approaches. Most of the approaches are established on various statistical techniques. Few approaches include corpora which has been tagged for senses and others employ unsupervised learning. In this paper, we utilize Lesk approach [1], which involves appearing for overlap among the words in given definitions with words from the text surrounding the word to be disambiguated. The flowchart is shown in Figure 1.

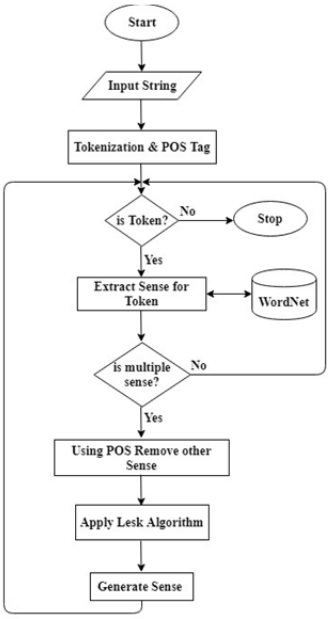

Fig. 1. Flowchart of WSD utilizing Lesk algorithm

A. Hindi WordNet

WordNet is a giant database consisting of nouns, verbs, adjectives, and adverbs. These are clustered into equivalent word sets or synsets. These sets are related by methods of applied lexical and semantic relations. WordNet is finishing for Hindi and it is being designed for Marathi at IIT Bombay. It is an electronic, huge, lexical database and it is a perfect mixture of thesaurus and dictionary which is made and handled by the cognitive science lab of Princeton University. The WordNet of Hindi is similar to the WordNet of English. The words are assembled together as indicated by their comparability of

implications. Two or more words that may be different in a context can be synonymous in some other context. Similar synsets are present within the Hindi Wordnet for every word, denoting a lexical thought. This is achieved to expel ambiguity in situations where a solitary word has numerous implications. The fundamental building squares of WordNet are the synsets. The Hindi WordNet manages the content words or words with open class categories. Thus, the Hindi WordNet comprises of the accompanying category of words: noun, adjective, verb and adverb. All entries in the Hindi WordNet consist of the following attributes:

- Synset: It is a collection of words with similar meanings i.e. synonyms. For example "पेन, कलम, लेखनी" (pen, kalam, lekhanee) describes an instrument for writing with ink. The words are arranged in the synset according to the popularity of usage.
- Gloss: It elaborates the concepts. It contains 2 parts: text definition which describes the concepts denoted by the synset (for example: "स्याही के सहयोग से कागज आदि पर लिखने का उपकरण" (syaahee ke sahayog se kaagaj aadi par likhane ka upakaran) explains the concept of a device for writing or drawing with ink), and example sentence which gives the importance of words in a sentence. Generally, the words in a synset can be easily replaced in a sentence (for example: "यह पेन किसी ने मुझे उपहार में प्रदान की है|" (yah pen kisee ne mujhe upahaar mein pradaan kee hai) gives the utility of words in the synset describing an instrument for writing or drawing with ink).
- Semantic Relations: It affirms the connection between the form and the meaning of a word. The Hindi WordNet is influenced through the English WordNet, and it is utilising lexical data. They are primarily used to construct the lexicon, and the semantic relations are shown in Table I.

B. Lesk Algorithm:

Lesk algorithm [1] is a traditional knowledge-based WSD algorithm which disambiguates a word by choosing a sense whose definition overlaps the most with the words in its context. Alongside there are considered different sources like synonyms, hypernyms, homonyms, meronyms, and example sentences. Numerous ensuing knowledge-based systems are situated on the Lesk algorithm. A knowledge-based approach utilizes outer lexical assets like thesauri and WordNet.

Suppose W i.e. '**राज्य**' (Rajyaa) is an ambiguous word and there is a group of a context word sets i.e. C in the window surrounding it. We search W in WordNet and then discover the senses of W. By using LESK algorithm, we search the most appropriate sense of W. The maximum overlapping sense will be the output of the algorithm. We can use simulated annealing to remove ambiguity if there is more than one word in the sentence. The steps of the Lesk Algorithm are:

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

Figures 2-4 show the multiple senses of an ambiguous word as shown in Hindi WordNet according to the various parts of speech.

TABLE I. RELATIONS IN HINDI WORDNET

Relation	Meaning	Example
Hypernymy/ Hyponymy	Is-a (kind-of)	बेलपत्र (belapatr) is a kind- of पत्ता (patta)
Entailment/ Troponymy	Manner-of (for verbs)	खराटा लेना, नाक बजाना → सोना (kharaata lena, naak bajaana → sona)
Meronymy/ Holonymy	Has-a (part-whole)	जड़ (jad) is the part- of पेड़ (ped)
Antonymy	holds between two words that express opposite meanings	मोटा→पतला (mota→patala)
Causative	the pattern of making causation	खाना→ खिलाना (khaana→ khilaana)

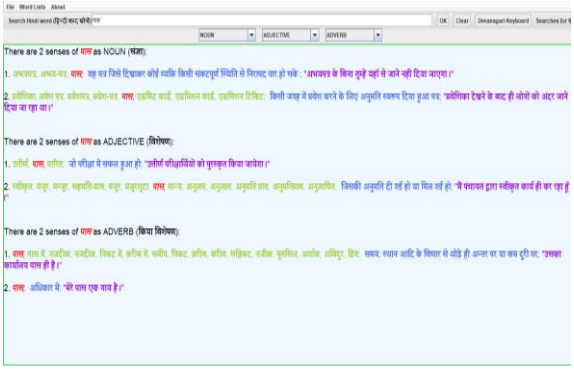

Fig. 2. Senses of the word 'पास' (paas)

As an example let's find the correct meaning of this sentence: 'मोहन के पास एक अच्छी कलम है।' (Mohan ke paas ek achchi kalam hai). In the given sentence 'पास' (paas), 'अच्छा' (achcha) and 'कलम' (kalam) contain multiple senses. The 7 different senses in Hindi WordNet for the word 'पास' (paas) are shown in Figure 2. If we link that 'पास' (paas) word with 'अच्छा' (achcha) and 'कलम' (kalam) word then there will be several senses for the sentence 'पास अच्छी कलम' (paas achchi kalam). The 14 different senses in Hindi WordNet for the word 'अच्छा' (achcha) are shown in Figure 3 and the 11 different senses of word 'कलम' (kalam) are shown in Figure 4. Lesk algorithm helps us in fetching the actual meaning of the word according to its usage. This reduces the number of

iterations required for WSD. According to Lesk algorithm, the best sense of 'पास' (paas) is 'अधिकार' (adhikaar), the best sense of 'अच्छा' (achcha) is 'सुंदर' (sundar), and the best sense of 'कलम' (kalam) is 'पेन' (pen). In such a process we are avoiding stop words such as का, के, की, है, हो, से, मे, ने (ka, ke, kee, hai, ho, se, me, ne) etc.

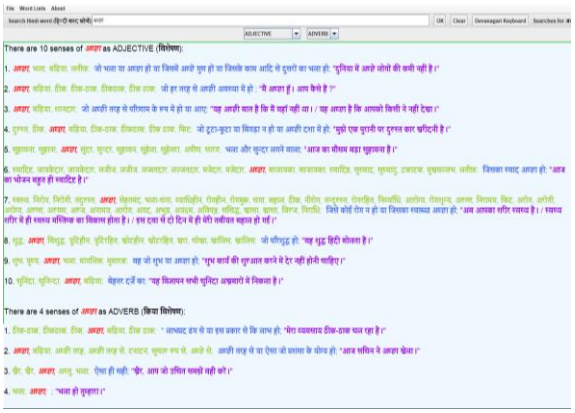

Fig. 3. Senses for the word 'अच्छा' (achcha)

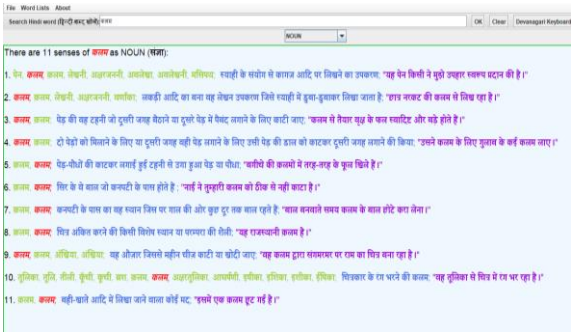

Fig. 4. Senses for the word 'कलम' (kalam)

IV. EVALUATION

We developed a tool for WSD by using the WordNet for Hindi. For this purpose, we considered a small corpus that contains 3000 ambiguous sentences. From this, we correctly identified 2143 sentences. The accuracy of the system was calculated by dividing the correctly identified senses with the total ambiguous sentences as shown in (1). The overall accuracy of our system is 71.43% and the error was 28.57%. as calculated by (2). The summary of the evaluation is shown in Table II and Figure 5.

$$Accuracy = \frac{Correctly\ Identified\ Senses}{Total\ Ambogious\ Senses} \quad (1)$$

$$Error = \frac{Incorrectly\ Identified\ Senses}{Total\ Ambiguous\ Senses} \quad (2)$$

TABLE II. EVALUATION SUMMARY

Total Ambiguous Senses	3000
Correctly Identified	2143
Accuracy	71.43
Error	28.57

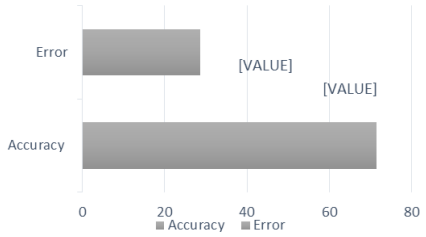

Fig. 5. Accuracy in error measure

V. CONCLUSION

In this paper, focus was given in Hindi language. WSD has been a significant issue of NLP and AI, but the issue was to find the correct senses of words in respective contexts. We utilized a knowledge-based approach by implementing Lesk algorithm for Hindi. Our system gave 71.4% accuracy for finding the correct sense of an ambiguous word. We plan to add more knowledge sources like part-of-speech (POS) tagger to improve accuracy. It would be interesting to see how selectional restrictions can improve the accuracy of the current system.

REFERENCES

[1] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone", in: Proceedings of the 5th Annual International Conference on Systems Documentation, pp. 24-26, AMC, 1986

[2] J. J. Jiang, D. W. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy", arXiv preprint cmp-lg/9709008, 1997

[3] M. Stevenson, Y. Wilks, "The interaction of knowledge sources in word sense disambiguation", Computational Linguistics, Vol. 27, No. 3, pp. 321-349, 2001

[4] R. Navigli, P. Velardi, "Structural semantic interconnections: a knowledge-based approach to word sense disambiguation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 7, pp. 1075-1086, 2005

[5] E. Agirre, A. Soroa, "Personalizing page rank for word sense disambiguation", in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 33-41, Association for Computational Linguistics, 2009

[6] E. Agirre, O. L. De Lacalle, A. Soroa, I. Fakultatea, "Knowledge-Based WSD and Specific Domains: Performing Better than Generic Supervised WSD", Twenty-First International Joint Conference on Artificial Intelligence, Pasadena, USA, July 11-17, 2009

[7] R. Kaur, R. K. Sharma, S. Preet, P. Bhatia, "Punjabi WordNet relations and categorization of synsets", 3rd National Workshop on IndoWordNet Under the Aegis of the 8th International Conference on Natural Language Processing (ICON 2010), Kharagpur, India, December 8-11, 2010

[8] M. M. Khapra, S. Shah, P. Kedia, P. Bhattacharyya, "Domain-specific word sense disambiguation combining corpus based and wordnet based parameters", 5th International Conference on Global Wordnet, January 31-February 4, Mumbai, India, 2010

[9] R. Navigli, M. Lapata, "An experimental study of graph connectivity for unsupervised word sense disambiguation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 4, pp. 678-692, 2010

[10] P. Bala, "Knowledge Based Approach for Word Sense Disambiguation using Hindi Wordnet", International Journal of Engineering and Science, Vol. 2, No. 4, pp. 36-41, 2013

[11] S. Kumari, P. Singh, "Optimized word sense disambiguation in Hindi using genetic algorithm", International Journal of Research in Computrer & Communication Technology, Vol. 2, No. 7, pp. 445-449, 2013

[12] E. Agirre, O. L. de Lacalle, A. Soroa, "Random walks for knowledge-based word sense disambiguation", Computational Linguistics, Vol. 40, No. 1, pp. 57-84, 2014

[13] P. Basile, A. Caputo, G. Semeraro, "An enhanced lesk word sense disambiguation algorithm through a distributional semantic model", 25th International Conference on Computational Linguistics, Dublin, Ireland, August 23-29, 2014

[14] D. S. Chaplot, P. Bhattacharyya, A. Paranjape, "Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser", in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 2217-2223, Association for the Advancement of Artificial Intelligence, 2015

[15] A. R. Pal, D. Saha, A. Pal, "A Knowledge based Methodology for Word Sense Disambiguation for Low Resource Language", Advances in Computational Sciences and Technology, Vol. 10, No. 2, pp. 267-283, 2017

AUTHORS PROFILE

Pooja Sharma received her MCA degree from the Banasthali Vidyapith, Rajasthan, India in 2016 and is currently doing M.Tech in computer science in Banasthali Vidyapith, Rajasthan, India. Her research activity is mainly focused at the development of WSD algorithms and their techniques for Indian Languages. She also has interest in other fields of NLP and AI.

Nisheeth Joshi is an Associate Professor in the Department of Computer Science, Faculty of Mathematics and Computing, Banasthali Vidyapith, Rajasthan, India. He has done his PhD in the area of Natural Language Processing. He is involved in teaching and research for over 10 years and also has a vast experience of handling large scale research projects sponsored by various funding agencies He has authored several papers in international journals and conferences regarding AI and NLP.