# WORD SENSE DISAMBIGUITION FOR DEVNAGARI LANGUAGE

Chitra Liladhar Mahajan[1] Sandip S. Patil[2]
*Department Of Computer Engineering,*
*SSBT's College Of Engineering And Technology, North Maharashtra University,*
*Jalgaon, Maharashtra, India*

***Abstract* - In natural language processing and understanding, semantic processing is an important task. In semantic processing some words has multiple senses (meanings) which are unrelated with each other. These multiple senses possess critical problems to linguists and they create ambiguity in sentence. Word sense disambiguation accepted this challenge. It is one of the central challenges in NLP and occurs in all the languages. Human can easily disambiguate the words but machine can not. WSD has numerous applications in machine translation, information retrieval, question-answering etc. The ambiguity can be lexical and semantic. In NLP, Word Sense Disambiguation(WSD) is the task of perfectly assigning the acceptable, correct sense (meaning) to the words having multiple senses in the given natural language text. WSD is categorized in three types wiz. Knowledge base, machine learning and Hybrid approach. The work carried out on Marathi language is limited. In the proposed work, we are resolving the ambiguity in Marathi words based on their senses and their context hybrid approach. Hybrid approach consist of modified Lesk with support Vector Machine.**

***Index Terms:* WSD, Sense annotated corpus, Ambiguity, Context, polysemous word, Context window, SVM, Wordnet**

## I . INTRODUCTION

Today is the era of information technology. Everyone is using web to share and find information. But, the information is present in natural languages. As know that natural languages are ambiguous i.e single word denotes the different meaning. Ambiguity is something which can be understood in two or more ways. So, to use information technology efficiently need to remove ambiguity from the sentences with the help of tool called Word Sense Disambiguation. Word Sense Disambiguation is one of the task of identifying correct meaning of polysemous word given in context.

For example:

1. □□□□ □□ □□□□□□ □□□ □□□□.
2. □□□□ □□□□□□ □□□.
In sentence 1, the word □□□□ indicates: name of the “Person” and in sentence 2, it indicates “Sky” sense.

Ambiguity is one of the problems which have been a great challenge for computational linguists. Something is ambiguous when it can be understood in multiple possible ways or when it has more than one meaning. Sometimes two completely different words are spelled the same. Word Sense Ambiguity makes it tough for Computers automatically carry out Natural Language applications like machine translation, information retrieval, question-answering etc. Every natural language suffers from sense ambiguity problem. Word sense disambiguation (WSD) is the problem of determining in which sense a word is used in a given context. Ambiguity is something which can be understood in two or more ways. So, to use information technology efficiently need to remove ambiguity from the sentences with the help of tool called Word Sense Disambiguation (WSD). The significant work are exists on Word Sense Disambiguation for many different languages using various methods. But the work carried out on Marathi language is limited. Keeping this reality in mind, in proposed work the worked will be done for Marathi language using Hybrid approach.

In the understanding of natural language, processing an ambiguous word is an major challenge. Word Sense Disambiguation(WSD) is having the ability to identify the correct sense of the ambiguous word used in the sentence. The problem of identification of specific sense of given word seems to be easy for a human being by using common sense, but for machines, it is difficult task as it requires processing of huge amount of unstructured information present in natural languages to identify the correct meaning. In Literature, WSD is categorized in three types wiz. Knowledge base, supervised and unsupervised. Knowledge based WSD requires overlapped approach, supervised requires tagged corpus and unsupervised gives less accuracy. However in literature, Marathi WSD has not taken under consideration. The proposed WSD approach, disambiguates the Marathi words by using hybrid approach, which resolves the ambiguity from the words based on their senses and their context in the Marathi sentence. In hybrid approach for marathi ambiguous words, considered the two words previous and two words after ambiguous words. The system works on only single sentence at present and identify the ambiguity.

## II. RELATED WORK

Navigli et al. [1], 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.

Mincã et al. and Diaconescu et al.[2], This paper proposed knowledge based approach derived from Lesk algorithm. This method considers an extension of the definition domain of the Lesk algorithm by creating a lexicon network. In this paper, use of semantically tagged glosses to create a lexicon network. Also, described a method to build semantic trees for each sense using such a lexicon network. This method was reduce the computational volume for the best variant detection. Testing was done on WordNet and it showed good results for WSD using only semantic trees similarity costs.

Singh et al[3], This paper introduced a word sense disambiguation system that was developed for the Manipuri language. Supervised approach is used for this purpose. A decision tree based method is used. The database used consisted of 672 sentences of Manipuri language. It consisted of 2,000 polysemous words which were of main concern. Context based and conventional positional features were used to figure out the sense of the polysemous words. The system gave an accuracy of 71.75%. But In this system, the input data set was limited.

Prity Bala[4], This author studied the Word Sense Disambiguation problem which is an open research area in both computational linguistics and Natural Language Processing. A system was proposed to find the sense of an ambiguous word in some given context. The knowledge based approach was used along with selected restrictions method in the proposed system. Hindi WordNet was used for the development of WSD system. WordNet was built from collocation and co-occurrence, and included the synonyms belonging to noun, adjective, verb or adverb. The accuracy of the system developed was 66.92%. The accuracy of the system was limited because some words were not tagged correctly with POS tagger.

Gopal et al. and Haroon et al.[5], This paper presented Word Sense Disambiguation using Naïve Bayes Classifier for Malayalam language. Lack of good Corpus is one of the major problem has been solved in this approach. The proposed system found that the quality of WSD system is directly proportional to the quality of corpora in which used. This proposed system provide us 95% reliability using a corpora of 1 lakh words.

Kalita et al. and Barman[6], This paper presents implementation of Knowledge Based Walker algorithm for disambiguation of Assamese language. The algorithm has been tested separately for manually designed and randomly taken sentences to examine the behavioral change in performance. Most of the manually designed sentences are in lack of words with a wordnet entry. Walker algorithm with Large window size get accurate result for sentences.

Tayal et al[7], This author developed an approach to handle the disambiguation for the Hindi language. The unsupervised approach was used. Hyperspace Analogue to Language(HAL) vectors had been used for the training purpose where each word is mapped into the high-dimensional space. Fuzzy C–means algorithm was used for making clusters denoting the occurrences of the various polysemous words. Finally the test data was mapped to the high dimensional space. It was also concluded that this approach is not language specific. The most important advantages of this approach involves the word knowledge and the context specific word sense disambiguation.

Sarika et al[8], This author studied the Word Sense Disambiguation system for Hindi language. The proposed work used the cosine similarity for developing a Hindi Word Sense Disambiguation system using a supervised approach. A dataset of 90 ambiguous words were used for the experimentation. Based on the cosine similarity, a sense was assigned to the ambiguous words. An average recall of 72.58% and an overall average precision of 78.99% were obtained. The only drawback was that disambiguation for the 65 parts of speech except noun was not done properly because they have shallow networks of relation in Hindi WordNet.

Sharma et al. and Niranjan et al.[9], This paper presents a comparison between supervised Machine Learning Algorithms Random Forest and combination of unsupervised machine learning algorithm K-Means clusterer and Random Forest Machine learning algorithm. The dataset file poach.arff is used as input to WEKA. Both the concepts and the representation appear very sparsely. Accuracy of Random Forest classifier with K-Means clusterer is high i.e. 82.3% which is highly required.

Pal et al. and Saha et al.[10], This paper has been surveyed on the different approaches adopted in different research works and also made a survey on WSD in different international and Indian languages. The different Word Sense Disambiguation approaches along with their different methods or techniques has been discussed in this paper and also comparison was done between them. Applications of WSD has been presented. This paper discussed the progress of Asian languages, especially in Indian languages was good, due to large scale of morphological inflections, development of WordNet, corpus and other resources.

Srinivas et al. and Rani et al[11], This survey paper presented unsupervised Graph based WSD for English which have been ability to remove the accuracy gap from the supervised methods. Then, discussed the various efforts accomplished by several researchers to develop WSD systems for Indian languages like Hindi, Kannada, Malayalam, and Assamese. Finally, discussed about WSD for other Asian languages like Nepali, Arabic and Myanmar. In Asian languages, development of corpus, WordNet and other resources is progressing slowly due to the more morphological inflections. The accuracy of the WSD and performance of the system depends on size of the corpus.

Garje et al, Kharate et al., Kulkarni et al[12], This paper introduced the architecture of a Machine Translation System with source language as English and target language as Marathi. The rule based approach was used for this purpose. The mainly focuses on the grammar structure of the target language that will produce better translations. The number of rules formed is large for target language generation to achieve better quality translations. However, there exist many exceptions in the language which do not conform to these rules. The exceptions can be handled but it will increase the size and the complexity of the knowledge base.

In literature survey, The supervised approach requires huge amount of sense-tagged corpora which is expensive to create and required large amount of manual efforts. The unsupervised approach focuses on various knowledge sources to build their models. But, The performance is not as good as the other approaches and the algorithms are difficult to implement. The knowledge based approach used the available information in large lexical database such as Word-net. The performance depends on the dictionary and it is overlap based.

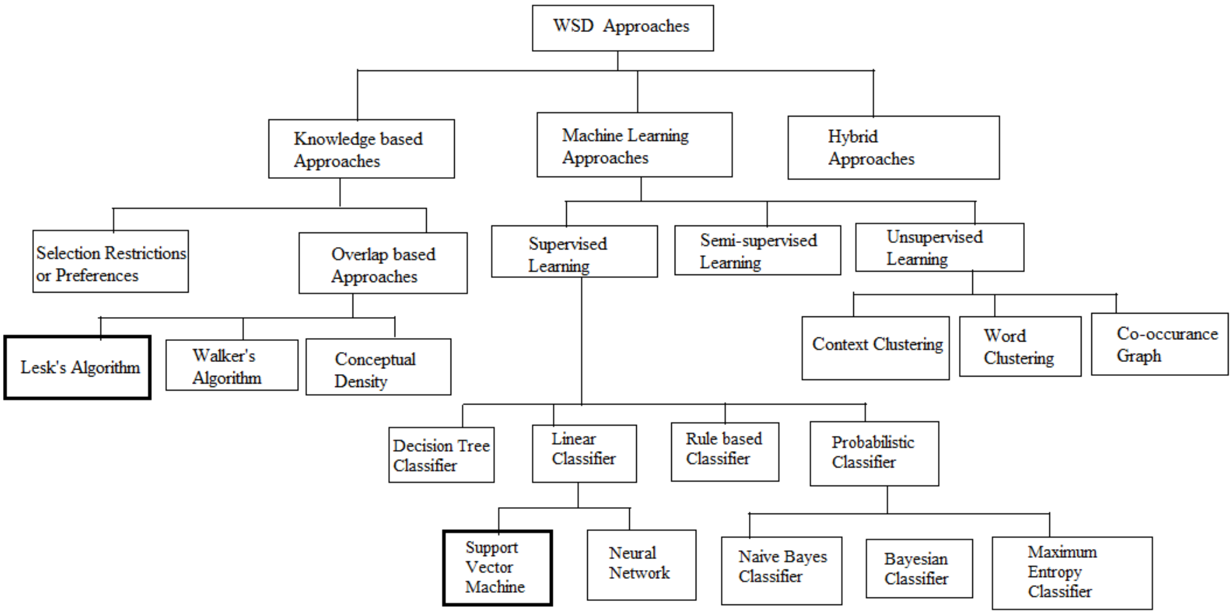

Fig.1 Structure of Literature Survey

## III. PROPOSED APPROACH

### *I. Architecture*

The Proposed approach is based on the mixture of Knowledge based approach and supervised learning approach and is used for word sense disambiguation of Marathi language. The Modified Lesk Algorithm uses the dictionary. It matches the dictionary definition of the target words with the most closely related left and the right words of the target ambiguous words. These target words are also known as neighboring words. In this proposed work, simply give the instance output of modified lesk approach to Support Vector Machine and get the instance output of Support Vector Machine to find better result of disambiguation.

The algorithm uses the concept of dynamic context window. The context window is considering the left and the right words of the target ambiguous word and which is chosen dynamically. The removal of special tokens like “|” & “,” is the first step of the proposed algorithm. At every step increase the size of the context window, since greater the size of the context window better will be the precision of this algorithm. Precision and Instance of the algorithm are calculated after determining the number of an ambiguous word. In this proposed system, mainly three metrices will be used to calculate the performance of the system which are the Precision and Recall and F-Measure. The system works on only single sentence at present and identify the ambiguity. ILTI Sense annotated corpus will be used to obtain the exact features of each word in the sentence.

Text is an unstructured source of information, to make it a suitable input to the automatic method it is usually transformed into a structured format. To this end, a preprocessing of the input text is usually performed, which is includes the following steps: The figure 2 shows the architecture of Proposed system.

- Tokenization: Tokanization is the process of breaking the sentence into words. Various grammatical rules are applied for tokenization of sentences into words. The tokens act as a particular word which is extracted from the sentence. Generally in any language the sentence boundary is identified when stop word occur, similarly the words are separated by using space.

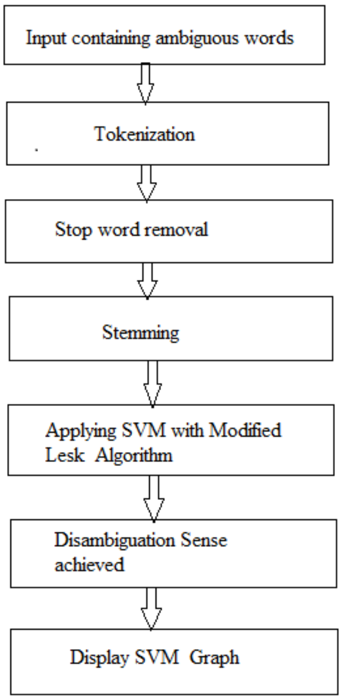
Input containing ambiguous words
Tokenization
Stop word removal
Stemming
Applying SVM with Modified Lesk Algorithm
Disambiguation Sense achieved
Display SVM Graph

Fig.2 Architecture of Proposed System

- Stop Word Removal: Stop words are identified after stemming is done. They show higher occurrence frequency and are relatively small in length. These two features help in their detection. For each cluster, its size is normalized as Number of word in a cluster and Total number of word in all clusters. Stop words mostly occupy the upper positions of the list. So, detection of stop words is easy.

- Stemming and Lemmatization: The main purpose of stemming is to get the radix positive integer value of each and every significant word and frequency is improved with the help of stemming. Stemming is a examining process that crops the ends of words in the hope of achieving this goal correctly most of time, and it often includes the removal of derivational affixes. Lemmatizations is the reduction of morphological variants to their root form. Lemmatization is doing things properly with the use of a morphological analysis of words and vocabulary, its target to delete inflectional endings only and its return the dictionary, theasures form of a word, which is known as lemma. In proposed system uses the Rule-Based stemmer for Marathi.

- SVM Classifier**:** Support Vector Machine classifier is used to obtain the better result. Support vector machine is a best technique to solve the word sense disambiguation. For this purpose compare the two approach name SVM and New Lesk approach in which first is based on knowledge or Decision based approach and second is based on supervised approach. To find the accurate result of marathi sentences disambiguation, give the instance output of modified lesk approach to SVM and get the instance output of SVM. Modified Lesk algorithm based on overlap based approach simple tell the no of senses in given context but no tells the exact meaning of the target word. So goal is to provide the traning to the the system to with the help of SVM so that machine easily find the exact meaning of the target word.

*II. Algorithm*

- SVM with Modified Lesk Algorithm

WSD is a technique by which single word having more than one meaning is clubbed together. The system is then trained to find the exact meaning of that word. Modified Lesk Algorithm is under the category of overlap based approach. The precision of a word can be calculated as the product of 100 and the ratio of the number of words and the instances of that word.

Precision P = No. of instances / Number of average words

Input: Text containing ambiguous words

Output: Actual sense of the ambiguous words

**Step 1:** Words <− no. of words in a sense after removal of special tokens
**Step 2:** Sense Computation <− no. of sense of word
**Step 3:** Determine Instance Count
**Step 4:** If word sense overlap target word sense
**Step 5:** Instance count = +1
**Step 6:** Calculate Precision P for every target word
**Step 7:** If P< Threshold Value
**Step 8:** Increase Context Window Size
**Step 9:** Calculate Precision P again

In Step1, Calculate the number of words in a sentence and also removes the special tokens like ‘,’ or ‘|’ followed by all the specialized symbols. In Step2, Calculating the number of senses of the word. In Step 3, determine the instance count. Then, In step 4 and 5 , Calculating the instance output of every target word . The context window is dynamic. Context window are the number of left and the right words of the target words. If given word sense overlaps the target word sense, then, instance Count is +1. In Step 6, calculating the precision P. In Step 7, if precision value less than threshold value then increase the context window size and Again calculate the precision.

## IV. CONCLUSION

In this paper,. The hybrid approach will be used for proposed system and this approach is the combination of the knowledge based approach and machine learning approach. The Lesk approach has modified which uses a dynamic context window. The context window which forms the left and the right words of the target word is chosen dynamically. The Modified Lesk algorithm is used with support vector machine classifier will be used in proposed work to get the better result for disambiguation. The whole work will be carried out on Marathi language.

## REFERENCES

[1] R.Navigli, “*Word sense disambiguation:a survey*", ACM Computing Surveys(CSUR), Vol. 41 , no .2, p.-10, 2009.

[2] Andrei Mincă, Ştefan Diaconescu, “An Approach *to Knowledge-Based Word Sense Disambiguation Using Semantic Trees Built on a WordNet Lexicon Network*”, 6th Conference on Speech Technology and Human-Computer Dialogue(SpeD), IEEE, p.1-6, 2011.

[3] Singh and R. L. Ghosh, K. and Nongmeikapam, K. and Bandyopadhyay, S., “*A decision tree based word sense disambiguation system in manipuri language*”, Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, page no- 17-22, July 2014.[4] Prity Bala, “Word Sense Disambiguation Using Selectional Restriction”, International Journal of Scientific and Research Publications, Volume-3,Issue 4,p.1-5, 2013.

[5] Sreelakshmi Gopal and Rosna P Haroon, “Malayalam *Word Sense Disambiguation using Naïve Bayes Classifier*”, IEEE,International Conference on Advances in Human Machine Interaction (HMI - 2016), R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India, Page no.1-4, March 2016.

[6] Purabi Kalita and Anup Kumar Barman, “*Implementation of Walker Algorithm In Word Sense Disambiguation for Assamese language*”, International Symposium on Advanced Computing and Communication (lSACC), pg.1-5, 2015.

[7] Devendra K. Tayal et al, “Word *Sense Disambiguation in Hindi Language Using Hyperspace Analogue to Language and Fuzzy C-Means Clustering*”, International Conference on Natural Language Processing, p.247-256, 2013.

[8] Sarika and Dilip Kumar Sharma, “Hindi *Word Sense Disambiguation Using Cosine Similarity*”, Proceedings of International Conference on ICT for Sustainable Development, p.801-808, 2016

[9] Neetu Sharma, Dr. S. Niranjan, “OPTIMIZATION *OF WORD SENSE DISAMBIGUATION USING CLUSTERING IN WEKA*”, Int.J.Computer Technology & Applications,Vol 3 (4), 1598-1604, p.1-7 July-August 2012.

[10] Alok Ranjan Pal and Diganta Saha, "*WORD SENSE DISAMBIGUATION: A SURVEY*", International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3, p.1-16, 2015.

[11] Mulkalapalli Srinivas and B. Padmaja Ran, "Word *Sense Disambiguation Techniques for Indian and other Asian Languages: A Survey*", International Journal of Computer Applications (0975 – 8887)
Volume 156 – No 8,p.1-7, 2016.

[12] G V Garje, G K Kharate, Harshad Kulkarni, " *Transmuter: An Approach to Rule-based English to Marathi Machine Translation*", International Journal of Computer Applications (0975 – 8887) Volume 98 – No.21, July 2014.