Sense Disambiguation Using Decision Graph for Marathi Language Words

Nutan B. Zungre[1], Gauri M. Dhopavkar[2], Nagmani Wanjari[1]
[1]P.G., Department of Computer Science and Engg., YCCE, Hingna Road, Nagpur, India.
[1]nutanzungre22@gmail.com, [3]*nagmani.wanjari@gmail.com*
[2]Asst. Prof., Department of Computer Science and Engg., YCCE, Hingna Road, Nagpur, India.
[2]*gauri.ycce@gmail.com*

***Abstract*—"Sense Disambiguation" of a word is a simple way of selecting proper sense (meaning) for an ambiguous word in a given context. Sense disambiguation of a word is very crucial, and its importance is used in every application of computational linguistics. There are many methods available in Natural Language Processing for performing the sense disambiguation. In our work, we are proposing the Graph-based algorithm through which the ambiguity is resolved for the words based on their senses and the context domain. Graph-based algorithm creates a graph comprising the word to be disambiguated along with their corresponding candidate sense. In the proposed work, sense disambiguation of word has been done for Marathi language words. Multiple meanings of Marathi word has been explored with the help of Marathi WordNet [4] prepared by IIT-Bombay.**

Keywords: Natural Language Processing; Word Sense Disambiguation; Graph-based Algorithm; Ambiguity; Marathi WordNet.

I. INTRODUCTION

In (NLP) Natural Language Processing, Word Sense Disambiguation (WSD) is the task of determining in which "sense" (meaning) a word is emerged by the use of the word in a particular context. Sense disambiguation of word is basically a process of identifying the sense of an ambiguous word. In computational linguistics, sense disambiguation of word is an accessible problem of NLP and an ontology.

For computational linguistics, one of the great challenge is the problem of Ambiguity [6]. It can be said that a word is ambiguous when it is understood in more than two possible ways or when it has more than one meaning, called as Polysemy. For example, 'fix' word has many meanings like Attach, Arrange, Get ready (food or drinks), repair, punish, set right (the hair), etc. Sometimes two words either spells the same or sounds the same or both but has different meanings, this is called as Homonymy. For example, Made/Maid, Allowed/Aloud, Mail/Male, Lone/Lone, etc. So ambiguity is one of the major problem that often arises in the computational linguistics. Disambiguation of senses of word can greatly help in minimizing the ambiguity in the computational linguistics. Word Sense Disambiguation (WSD) is a function or a task of determining in which way a word (having meaning in more than two possible ways) is used in a given sentence. Sense disambiguation of word basically identifies the sense of word when the word has lots of meanings. Sense disambiguation is an Artificial-intelligence problem. Word sense disambiguation greatly depends on knowledge. Sense disambiguation plays a crucial part in finding the meaning of words. Word Sense Disambiguation(WSD) can be defined as a function of determing the correct meaning of the word in a given context. The function needs large amounts of lexical and word knowledge.

For example, lets take the word '**आकाश**' in the following two Marathi sentences.

1. **आकाश खेळत आहे**.
2. **आकाश निरभ्र आहे**.

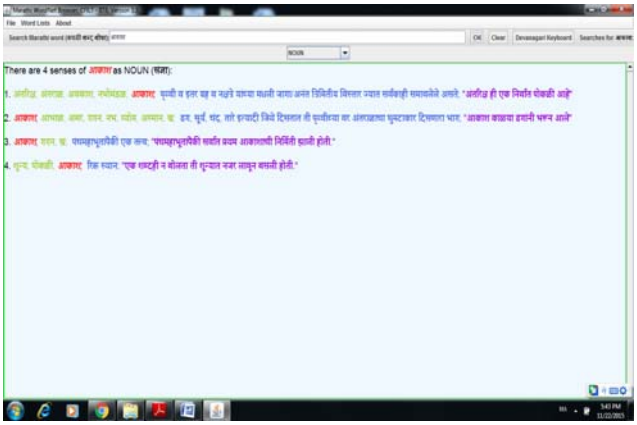

Figure I.1: sense of word '**आकाश**' obtained from Marathi WordNet

In the above two sentences the word '**आकाश**' has different meanings respectively. In the first sentence, the word '**आकाश**'denotes a person (boy), who is playing. In the second sentence, the word '**आकाश**' denotes the sky, which is cloud free.

II. AMBIGUITY

Ambiguity for computational linguistics, one of the great challenge is the problem of Ambiguity (word having multiple

meanings). This problem of ambiguity is not only recognized by humans but also by computers.

A. Ambiguity for humans:

Communication is the part of humans life, so ambiguity is rarely a problem for humans; But, ambiguity as seen in written work or newspaper cannot be sometimes resolved by humans.

For example,

माणसाने दूर्बीन सोबत एका मुलाला पाहीले.

The above example creates the ambiguity like whether a man is handling the binocular or a boy.

B. Ambiguity for computers:

Ambiguity for computers is often a problem iff it does not have word knowledge or the WordNet attached with it. Otherwise, it is a rare problem depending on the situation or the context in which the ambiguous words are used.

For example,

फूलेंचे फूल फुलले.

In this example the ambiguity is created, where the first word in the sentence denotes a person or a flower. The second word denotes the flower or child or a favourite thing or work and the third word denotes to arise.

III. APPROACHES OF WSD

Approaches[9] of WSD are:-

A. Knowledge-based approach:

This approach uses a large corpus or machine readable dictionaries or sense inventories. Four main types of knowledge-based methods which are as follows:

i) The LE

SK Algorithm:
-- This algorithm depends on knowledge and dictionary of its context at which sense disambiguation of word is going to perform.
-- Main idea of this algorithm is to identify overlapping among the senses of words which are to be disambiguate.

ii) Semantic Similarity:
-- According to semantic similarity[2] words that are related shares a common context and thus an appropriate sense is chosen by those meanings that are found within the smallest and closest semantic distance.

iii) Selectional Preferences:
-- Selectional preferences finds information of the likely relations of the word types, and denotes the common sense knowledge of the classes.

iv) Heuristic Method:
-- Word sense is based on the heuristics that are drawn from the linguistics function or properties identified on large texts.

B. Supervised:

Supervised approach has number of algorithms for sense disambiguation of words and uses machine learning techniques for the purpose of disambiguation. Sense-annotated corpus is used by this approach. Methods for Supervised WSD are:-

i) Decision List:
-- A decision list is a list or set of if-then-else rules.
-- Training set is used in decision list to induce set of features for the given words.

ii) Decision Tree:
-- It is used to represent the division rules in a tree structure that iteratively divides the given training data set.
-- The internal nodes of a decision tree [1] represents a test which is going to be tested on a feature value, and each subsection (branch) indicates an output of the test.
-- When it reaches to its leaf node, forecasting about the meaning is made.

iii) Naïve Bayes:
-- This classifier is a possibility classifier, which uses an application of Bayes's theorem.
-- It depends on the calculation of the conditional probability of each sense of a word given the features in the context.

iv) Neural Networks:
-- Artificial neurons are grouped together and uses a computational model for data processing using a connectionist approach is a neural network.

v) Exemplar-Based/Instance-Based Learning:
-- This type of supervised algorithm draws a classification model from examples.
-- This model is used to store examples in memory as point in feature space and a new examples are considered for purpose of classification, and then they are gradually added to this model.

vi) Support Vector Machine:
-- Support Vector Machine is based on the features of Structural Risk Minimization from the theory of statistical learning.
-- Main idea in this is to separate positive example from negative example with the max. margin and margin is the measuring length of hyperplane to the closest of the positive or negative examples.
-- The positive & the negative examples nearest to the hyperplane are called as support vector.

C. Unsupervised:

Unsupervised approach assumes that the nearest or the closest words have similarities and that uses an un-annotated corpus. Main approaches of unsupervised are:-

i) Context Clustering:

-- This method is based on the clustering techniques where the first context vector is created and then they are grouped into clusters to find out the meaning of the word.

-- This approach uses a vector-space as word-space and the dimensions are used as words only.

ii) Word Clustering:

-- Word Clustering technique is very similar to context clustering in terms of identifying the sense but it clusters those words that are semantically same and so it gives a similar meaning of the word.

iii) Co-occurrence Graph:

-- Co-occurrence Graph method create a co-occurrence graph with vertex and edge.

-- vertex tends to the words in the text and edge is joined if the words co-occur in their relation according to the syntax in the same paragraph or in text.

IV. MARATHI WORDNET

In the proposed system, Marathi WordNet [12] has been used. Marathi WordNet has been developed at CFILT Lab, Dept of Computer Science & Engineering, IIT Bombay. It has been developed for the purpose of facilitating Natural Language Processing applications. The whole system contains:

1. A browser (Offline Java Swing based browser)
2. The database (text files in similar format as Princeton English WordNet)

The Marathi Wordnet Browser is a Graphical User Interface (GUI) to access the Marathi Wordnet lexical database. The user inputs the words in Unicode and the results are also shown in Unicode. The results that appears are the synsets i.e. synonyms of the entered word. The semantic relations of the word are displayed in the drop down boxes.

Synsets (synonymy set) are the basic building blocks of the WordNet. The Marathi WordNet contracts with the content-words or open-class type of words. Thus, the Marathi WordNet consists of the following types (category) of words-Noun(N), Verb(V), Adjective(ADJ) and Adverb(AD). Each entry in the Marathi WordNet consists of following elements:-

1) Synset: synset is a set of synonymous words. Similar words are displayed according to the frequency of usage. For example, '□□□' has similar meanings like □□□(financial), □□□□-□□□, □□□□□□, etc.

2) Ontology: It defines the concept being used. It contains two parts:

- ***Text definition***: Text definition determines the concept represented by the synset. For example, **पैसे व दागिने सुरक्षित ठेवण्याचे व जेथून कर्ज घेता येते ती सरकारी किंवा खाजगी संस्था**.
- ***Example sentence***: It gives the usage/function of the words in the sentence. For example, **त्याने बँकेत दहा हजार रुपये जमा केले**.

V. RELATED WORK

For Word Sense Disambiguation, the present system uses different approaches. For Marathi Word Sense Disambiguation, Marathi WordNet has been used. The use of the WordNet for Marathi developed at IIT-Bombay is a very important lexical knowledge base for Marathi. Our proposed work has used Graph-Based Method for disambiguation of Marathi word. One of the earlier work for this approach is that of *Ravi Sinha and Rada Mihalcea [2]*, their paper described an unsupervised graph-based method for WSD, and presented a comparative evaluations using various measures of word semantic similarity and various algorithms for graph centrality. *Richard Laishram Singh, Sivaji Bandyopadhyay, Krishnendu Ghosh, and Kishorjit Nongmeikapam [1]*, proposed the same but for Manipuri language. Word sense disambiguation has been done using set of contextual and positional features. *Ehsan Hessami, Faribourz Mahmoudi, and Amir Hossien Jadidinejad [3]* suggested Word Sense Disambiguation along with an algorithm based on weighted graph has been discussed. *Amita Jain & DK Lobiyal [4]*, used a Marathi part of speech tagged corpus for building the graph model. Sense disambiguation has been done for all POS in Marathi. *Sandeep Kumar Vishwakarma and Chanchal Kumar Vishwakarma [5]* mentioned the graph-based approach to word sense disambiguation and an ambiguity for humans and computers. *Gauri Dhopavkar, Manali Kshirsagar and Latesh Malik [6]* used a rule-based approach and mentioned that ambiguity problem is solved by 3 steps: i) Collecting datasets from different domains. ii)Creating features iii)Max. entropy model. Sense disambiguation has been done for Marathi language. *Preeti Yadav & Mohd. Shahid Husain [7]* says that the words in the polysemous context have differentiated roles to describe the polysemous sense. Sense disambiguation has been done for marathi words. *Ioannis P. Klapaftis & Suresh Manandhar [8]* says that the Total sense score(TSS) increases as frequency increases which in turn increases the confidence on correct disambiguation. Sense disambiguation has been done for Noun POS tagging. *Nirali Patel, Bhargesh Patel, Rajvi Parikh, Brijesh Bhatt [9]* presents a study-based paper which says that Supervised Approach is the better approach, but it requires large sense-annotated data. *Eneko Agirre, David Mart´ınez, Oier L´opez de Lacalle and Aitor Soroa [10]* has used Hyperlex algorithm in their paper. The HyperLex algorithm builds a co-occurrence graph for all pairs of words co-occurring in the context of the target word.

VI. WSD ALGORITHM

Ambiguity is one of the major problem for computational linguistics. The main aim of our work is to implement a system that can perform sense disambiguation for Marathi language words. The work of our proposed system focuses on developing a method used to resolve semantic ambiguity for Marathi language words. In fact, some Marathi word has more than one meanings.

For example, consider the word, -- '**कर**' -- It refers 3 meanings which is obtained from Marathi WordNet.

1) "**रावणाला दहा तोंडे आणि वीस कर होते असे म्हणतात**."

   Here the meaning of '**कर**' is hand/arm.

2) "**शेत विकत घेतल्यावर आम्ही दोन हजार रूपये कर भरला**."

   Here the meaning of '**कर**' is tax/bill.

3) "**त्याचा कर मशीनी खाली आला**."

   Here the meaning of '**कर**' is upper and lower part of hand span.

In this paper, a graph-based algorithm [5] for Marathi WSD has been discussed. The algorithm moves on incrementally on the sentence-by- sentence basis. The algorithm illustrates all the words in text by exploiting similarities identified among the senses of words. In this paper, it gives a conditional evaluation of various depth of the word semantic-similarity by using a graphical architecture.

Steps to Graph construction process:
1. Proceed sentence-by- sentence basis.
2. Initially, a graph G= (V, E) is constructed for each target sentence which is included from the graph of reference lexicon.
3. We assume that the sentences in Marathi language are part of speech tagged. So the proposed algorithm considers context word only.
4. In the graph, node represents word sense and the edge represent semantic relation.
5. With the help of Marathi WordNet, final graph is constructed.

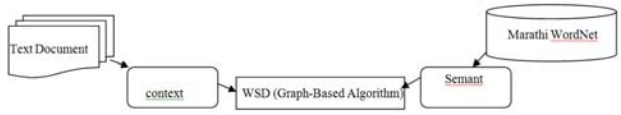

Figure VI.1:- Proposed System Architecture

The above Figure shows the flow of the proposed work. The graph-based algorithm goes through each context of the given text document and moves on incrementally on the sentence-by- sentence basis. It identifies the ambiguous words and finds the accurate sense for a given context by graphing and decision making. The algorithm uses the Marathi WordNet to find the semantics (meanings) of the ambiguous words in the context.

Algorithm Example:-

भारत हा एक <u>विशाल</u> कृषीप्रधान देश आहे. शेतकरी लोकं येथे <u>खूप</u> धान पिक्वितात. हे धान त्यांचासाठी <u>धन</u> सारखे असते.

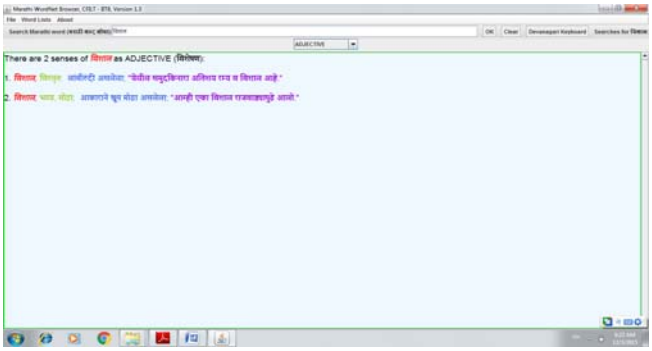

Figure VI.2 : Different senses for the word '**विशाल**' in Marathi WordNet

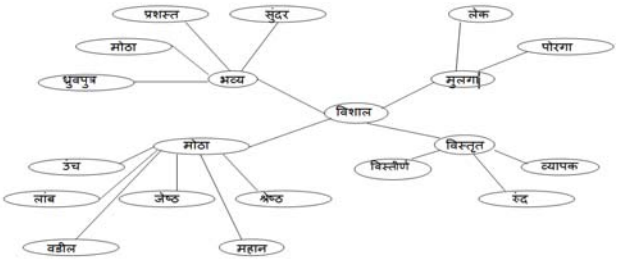

Figure VI.3: Graph generated for the word '**विशाल**'

The algorithm moves on sentence-by-sentence basis. It builds a graph by searching ambiguous words in each sentence. The constructed graph represents the context word and its related semantics from the Marathi WordNet.

VII. RESULT

The completed modules has used the Google input tool. Google input tool is a transliteration tool that is used to convert English language to Marathi. The created modules

Fetches the input Marathi word and identifies whether the given word is ambiguous or not. In every module, disambiguation part is done for each Marathi word.

In one of the module of the proposed system, text field has been created where Marathi sentence is entered and each Marathi word in the sentence is tokenized and disambiguated along with its POS tagging [11], root word and gender. A clustering algorithm is used to tokenize the words in the Marathi sentences. Marathi WordNet attached with the module displays every information like POS tagging, root word and gender after submitting it.

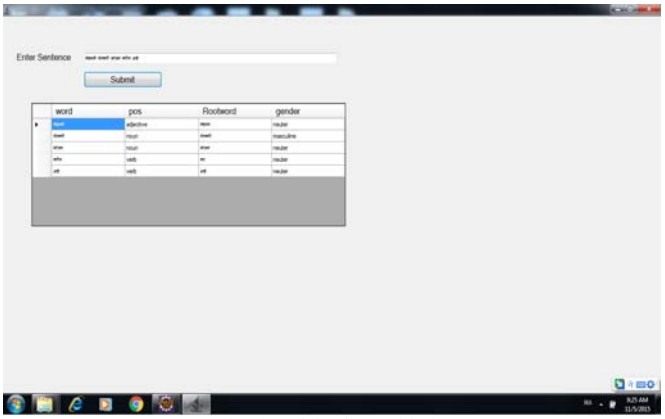

Figure VII.1: Example of module along with the output

For example, as displayed in the above figure, Marathi sentence "**मेहनती शेतकरी भोजन करित आहे**." has been taken as an input. Now in this case, each and every word in the sentence are tokenized and their corresponding features are represented respectively.

VIII. CONCLUSION

The proposed system is used to determine the correct sense of word in Marathi language using Decision graph algorithm. This system uses Source Language as Marathi. Input Text is obtained through Google Input Transliteration. Marathi WordNet is used to get the detailed features of each word in the sentence. Target language is the Marathi language for which the sense disambiguation is done.

REFERENCES

[1] Richard Laishram Singh, Sivaji Bandyopadhyay Krishnendu Ghosh and Kishorjit Nongmeikapam, "A Decision Tree Based Word Sense Disambiguation System In Manipuri Language", Advanced Computing: An International Journal (ACIJ), Vol.5, No.,4 July,2014.

[2] Ravi Sinha and Rada Mihalcea,"Unsupervised Graph- based Word Sense Disambiguation Using Measures Of Word Semantic Similarity", International Conference on Semantic Computing, IEEE. DOI 10.1109/ICSC.2007.87, 2007.

[3] Ehsan Hessami**,** Faribourz Mahmoudi, Amir Hossien Jadidinejad, "Unsupervised Weighted Graph For Word Sense Disambiguation", 2011 World Congress on Information and Communication Technologies,2011.

[4] Amita Jain & DK Lobiyal, "A New Method for Updating Word Senses In Marathi WordNet", International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014.

[5] Sandeep Kumar Vishwakarma, Chanchal Kumar Vishwakarma, "A Graph Based Approach To Word Sense Disambiguation for Marathi Language", International Journal of Scientific Research Engineering & Technology (IJSRET) Volume 1 Issue5 pp 313-318 August 2012 www.ijsret.org ISSN 2278 – 0882.

[6] Gauri Dhopavkar, Manali Kshirsagar, Latesh Malik, "Handling Word Sense Disambiguation In Marathi Using a Rule Based Approach", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 & International Conference on Industrial Automation and Computing (ICIAC- 12-13th April 2014).

[7] Preeti Yadav, Mohd. Shahid Husain, "Study Of MarathiWord Sense Disambiguation Based On MarathiWorldNet", International Journal For Research In Applied Science And Engineering Technology (IJRASET) Vol. 2 Issue V, May 2014 ISSN: 2321-9653.

[8] Ioannis P. Klapaftis and Suresh Manandhar, "Google & WordNet Based Word Sense Disambiguation", 22nd ICML Workshop on Learning & Extending Ontologies. Bonn, Germany, 2005.

[9] Nirali Patel, Bhargesh Patel, Rajvi Parikh, Brijesh Bhatt, "A Survey: Word Sense Disambiguation", International Journal of Advance Foundation and Research in Computer (IJAFRC) Volume 2, Special Issue (NCRTIT 2015), January 2015.

[10] Eneko Agirre, David Mart´_nez, Oier L´opez de Lacalle and Aitor Soroa, "Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm", 2006.

[11] Pallavi Bagul, Archana Mishra, Prachi Mahajan, Medinee Kulkarni, Gauri Dhopavkar, "Rule Based POS Tagger for Marathi Text",(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1322-1326.

[12] Naik Ramesh Ram, C. Namrata Mahender, "Marathi WordNet Development", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume - 3 Issue - 8 August, 2014 Page No. 7622-7624