RESOLVING AMBIGUITY IN MARATHI LANGUAGE TEXT: A RULE BASED SOLUTION

[1]Gauri Dhopavkar, [2]Manali Kshirsagar
[1]Research Scholar, GHRCE, Asstt. Professor, YCCE
[2]VP(Academics), ADCC Infocad Pvt. Ltd., Nagpur
Email: [1]gauri.ycce@gmail.com, [2]manali_kshirsagar@yahoo.com

Abstract— 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. Ambiguity poses many difficulties in major natural language processing tasks like machine translation, named entity recognition, information retrieval, information extraction etc. This paper discusses various ambiguities that arise in a natural language. It also presents explanation of a rule based technique to tackle different ambiguities in Marathi language which is the official state language of Maharashtra in country India.

Index Terms— Word Sense Disambiguation, WSD, ambiguity, natural language.

I. INTRODUCTION

All living things/organisms need to communicate with each other for their survivor. They do it using different methods. Many communicate through various event specific sounds, gestures etc. Homo sapiens are intelligent among all the species and they share their thoughts through Natural Language. Through natural language, all ideas, thinking, views are communicated accurately. Still, due to some features of language, meanings of words may shift leading to misunderstanding and miscommunications. Ambiguity being the inherent property of a language, either it may be overlooked as an obstacle, or we may need to resolve it for better communication.

For humans, they being intelligent may overcome misunderstanding and miscommunication caused by language ambiguities, by using various ways of analysis naturally. But in Natural Language Processing (NLP), various tasks are performed using machine. Here in processing, ambiguity is a crucial task. Ambiguity poses problems in majority of the NLP tasks like Machine Translation, Text Summarization and Named Entity Recognition etc. In order to deal with the problem of ambiguity, it becomes necessary to understand the linguistic view of nature of ambiguity.

Disambiguation task involves identifying and applying the knowledge of the relationship between various features of the words with the techniques available and applicable to resolve ambiguity.

II. NATURAL LANGUAGE PROCESSING

A Natural language (NL) is a language used by humans for writing and speaking purpose. Language serves as the primary means by which people communicate amongst themselves and record information. It has the potential for expressing and sharing all sorts of ideas, and conveying complex thoughts. By studying language, we can get more detailed knowledge about the world.

For example, languages like Hindi, Marathi, English, and Chinese are some natural languages.

Natural Language Processing (NLP) is a process of computer analysis which is used for sound communication between humans[1]

III. WORDS IN A LANGUAGE

In order to deal with multiple senses of the words in given natural language text, we need to understand about a word. There are a number of definitions of "word" in linguistics based on different views and various theories. Some of them are worth mentioning here to create background for Word Sense Disambiguation. They are listed as below:

1. "WORD" is the smallest free-standing sign in a language[4]
2. Word is combination of number of characters in a specific sequence which carries meaning.
3. As per Wikipedia Word is considered as sound. Word is also understood as a representation in writing or printing that symbolizes and communicates a meaning and may consist of a single morpheme or of a combination of morphemes[3]
4. Word may be treated as a single distinct meaningful element used in speech or writing used with others (or sometimes alone) to form a sentence[2]

In the effort to define and understand "word", one needs to determine where is the end of one word and starting of another word, thus identifying word boundaries[4] semantic relatedness and pragmatic features (veena Dixit).

A. Constituent Structure of Words

Like mentioned earlier, word is a combination of morphemes (basic units of meaning). These morphemes which are constituents of the word can be organized into a branching or hierarchical (tree) structure. For example, the word "*undivided"* contains three morphemes like:

1. prefix "un-"
2. verb stem "divide"
3. suffix "-d"

There are two combinations of this structure. First combination is "divide" + "-d" to make "undivided", then combined with "un-" to make "undivided"? Second combination is "un-" + "divide" to make "undivide", and then combined with "-d" to make "undivided". Since "undivide" doesn't exist in English, while "divided" does, the first structure, sounds correct.

B. Word Classes[9]:

Basically in linguists considered only two word classes, noun and verb. Further more classes were obtained, noun, verb, adjective, pronoun, preposition, adv erb, conjunction and interjection.

The words are categorized in various types. Usually, following four definitions of *word* are considered in linguistics, as mentioned below.

1. Orthographic word,
2. Phonological word
3. Lexical word
4. Grammatical word

The above word categories represent four levels of analysis in NLP. In each of these definitions, word is considered to be the central unit of focus for analysis of a natural language.

IV. MARATHI LANGUAGE

Marathi is the official state language of Maharashtra state, a state in Bharat(India). It is one of the top most languages at world level because of its total number of speakers worldwide. More than 90 million people in India speak Marathi language.

Marathi is spoken in Maharashtra, a Southern state in india[10,11]

Key Dialects of Marathi include Nagpuri Marathi, Varhadii, Cochin, Gawdi of Goa, Kosti, Kudali, Malwani, Kasargod, Dangii etc. Apart from Bharat, Marathi is spoken in Israel and Mauritius (Outside India)

Marathi language follows Subject, object, verb order. Nouns inflect for gender (feminine, masculine, neuter), number (plural, singular), and case (nominative, genitive, accusative, oblique, locative, instrumental,). As far as historical record of Marathi is concerned, Marathi is the only Indo-Aryan language. It is a language of Sanskrit origin, which preserves locative case. Like Sanskrit, Marathi language also has three genders (masculine, feminine, neuter). Generally Marathi language adjectives do not show inflection unless they end in long a. Marathi verbs show inflection for all tenses i.e. past, present and future. If verb do not show

agreement with either subject or object a special type of voice is generated. Suffixation is mainly Language and postpositions are attested. Over the years, words from different languages are accepted in Marathi language. The majority of foreign words accepted and popularly used in Marathi come from Urdu, Persian, and Arabic languages.

A. Word Categories in Marathi Language

According to the part they play and based on functional as well as formal differences in a sentence, words are categorized into different word classes/categories, called "parts of speech". In Marathi language total 8 main parts of speech are available.

Here, we consider eight major grammatical classes for the purpose of understanding the techniques for sense disambiguation. While devising the word sense disambiguation method, details of each part of speech are taken into consideration.

The grammatical categories of words are:
naama (नाम): Noun , Kriyaapada (क्रियापद): Verb, *vishheshaNa* (विषेशण): Adjective, *kriyaavishheshaNa* (क्रियाविषेशण): Adverb, *sarvanaama (*सर्वनाम): Pronoun, *shabdayogii avyaya* (शब्दयोगी अव्यय): Postposition, *ubhayaanvayii avyaya* (उभयान्वयी अव्यय): Conjunction , *kevalaprayogii avyaya* (केवलप्रयोगी अव्यय): Interjection

V. WORD SENSE DISAMBIGUATION

Sometimes two completely different words are spelled the same in a language.
For example: English language word "Fly", can be used as a modal verb, in the sentence, "You can fly high". It is used as noun in "The fly is poisonous". Here "fly" corresponds to insect type.
Process of WSD is described in many ways: Some are as below:

1. **Word sense disambiguation (WSD)** is the process of determining in which sense a word (having a number of distinct senses) is used in a given sentence. [12]
2. **The Word Sense Disambiguation (WSD)** is a particular task that removes the lexical or semantic ambiguity from the set of ambiguous words[13]
3. The computational recognition of meaning for ambiguous words in a particular context is termed as **WSD.**

In general, people are reluctant of the ambiguities in the language they use; because they are naturally intelligent at tackling with them using the common sense knowledge. On the contrary, computer systems do not have this common sense knowledge. So it cannot perform well to resolve ambiguity.

There are TWO types of ambiguity at Semantic level of NLP. They are:

1. Structural Ambiguity: If the ambiguity is in a sentence or clause, it is called **structural (syntactic) ambiguity**.

   Consider the example,
   "The boy saw the lady with the camera."
2. When ambiguity occurs in a single word, due to multiple senses it is called **lexical ambiguity**.

   For example, the words "Fly" or "book" in English or word "kar" in Marathi.

WSD is a important step in all major applications of NLP like Machine translation, information retrieval, automatic text summarization etc. Machine Translation is the most dependent application on WSD. In machine translation, WSD process is required at two stages of understanding the source language text and generating sentences in the target language. This is because a word in the source language may have more than one possible sense in the target language. For example, the English word *"can"* may be translated into Marathi language as "karu shakane*" or* for its sense of *"ability"* or as *"charavi or bhaanda"* for its sense of *"Container"* depending on the context. For the accurate translation, the system needs to know which sense is the intended sense in the text.

VI. LITERATURE REVIEW

Machine Translation (MT) is a very useful NLP application. The success of MT systems depends on how accurate is the translation carried out by the system. The important requirement of machine translation is that the

translated output of the text in the source language should preserve the semantic relatedness according to source language rules, locale of the language etc. In view of this requirement WSD module design plays a vital role in MT systems.

From the literature review carried out, it is very clear that much of the work related to Word Sense Disambiguation for NLP is done for languages like English, Chinese, Japanese, Swahili etc.

Good efforts have been carried out by various researchers in India also towards solution of WSD for Indian Languages. (Major work is carried out for Indian languages like Hindi, Punjabi, Bangla etc.) A detailed literature review has been carried out by [6,7, 8].

The reason behind choosing Marathi language is that a little work related to WSD is completed/ongoing in Marathi. Very few researchers are reported to work on Marathi language (which is official language of Maharashtra, a state in India.)

Most of the methods carried out in Marathi focus on Corpus based technique to resolve ambiguity problem which has its own limitations.

The major limitations of Corpus based methods are:

1. The method becomes Domain Specific as corpora are generally available based on certain domains.

VII. OUR APPROACH

In our approach, Marathi Word sense Disambiguation is carried out as unsupervised approach using Rule Based method. The rules are framed based on Marathi language ambiguity sources.

The WSD Method is implemented using Wordnet 1.3 and machine learning approach

A. Structure of Marathi Wordnet 1.3

Marathi Wordnet 1.3 is developed by IIT Mumbai [5]. It has following files-

1) Index_txt: Provides information about all words.
2) Data_txt: Provides details of every word in index file.
3) Onto_txt : Provides ontology details of the words in data file.

B. Design of our System using Rules

Rules work on Ontological relationship of words given in Wordnet 1.3 and grammar of Marathi Language.
We have designed, New Marathi parsers for Noun, adjective, verb, adverb.
Rules are word based and sentence based rules
The GUI consists of Morphological Analyzer, WSD buttons for rules, virtual keyboard, etc.
User may add new ambiguous words anytime.
User may frame new rules as and when needed just by identifying relationship between the words.
The system shows all verbs, nouns, adjectives adverbs in separate categories.
Description of Rules used in our system:
Word Rule: These rules are applicable only on individual words.
It includes rules under various classes like following.

*C. Rules to resolve ambiguity caused due to **postpositions** in words:[6,7]*

- **aalaya Rule**
- **aata Rule**
- **bhara Rule**
- **laya Rule**
- **mule Rule**
- **sheel Rule**
- **taa Rule**
- **vaadi Rule**
- **var Rule**

Explanation of "mule" (मुळे) Rule:
1. Rule Type: This rule comes under word rule.
2. In the word rule, a prefix word with suffix word forms a single word.
For example:-
अभाव (Prefix) + मुळे (Suffix) = अभावामुळे

3. Here, Prefix word forms a meaningful word which could be an adjective or noun.
4. According to the prefix word we make some conclusion.
5. "mule" (मुळे) Rule works s below:
- "mule" (मुळे) rule defines that a word with prefix cannot be a name of a person,

* It could be कारणदर्शक (Reason).

- Here we consider examples,

a)अभावामुळे

apply = Sense Count: 1

Possible sense: कारणदर्शक (Reason)

Correct Sense: कारणदर्शक (Reason)

D. Other word rules where word senses have different/ same parts of speech category:

-paatra Rule
-patha Rule
-kara Rule

E. Rules to identify co-reference of Proper nouns:

Consider the example, "raama ne aambaa khaallaa. to bhukella hotaa."
In this example, it is ambiguous to know about whether word "to" is referring to "raama" or "aambaa".

F. Rules to identify ambiguity due to prepositions

These rules tackle ambiguity generated due to lack of proper prepositions.

G. Rules for two verbs that appear in adjacent positions in a sentence

When in a sentence two verbs appear adjacent to each other, the first verb acts as "dhaatusaadhita" (which indicate incomplete action) and the second verb acts as actual verb(that indicates completed action).

VIII. CONCLUSION

Thus using variety of rules, Un-supervised WSD using Machine Learning system has been designed.
Here our system works directly on Marathi Wordnet 1.3. and Corpora are not used. It gives the advantage of large database.
Parsers are designed to identify detailed information about the words present in the input sentence.
A word is decided to be ambiguous if it has multiple senses.
To resolve ambiguity, all words present in context of the target word are considered.
Word rules and sentence rules are written to resolve ambiguity as per grammar of Marathi language. The system provides around 80% accuracy when counter checked with manual disambiguation.

REFERENCES

[1] NPTEL material, IIT Kharagpur
[2] www.oxforddictionary.com
[3] www.thefreedictionary.com
[4] https://en.wikipedia.org/wiki/Word
[5] http://www.cfilt.iitb.ac.in/wordnet/webmwn/
[6] G.M.Dhopavkar, M.M.Kshirsagar, L. G. Malik, "Handling Word Sense Disambiguation in Marathi using a Rule Based Approach", International Journal of Engineering Research and Applications(IJERA), ISSN: 2248-9622, pp 13-16, April 2013.
[7] G.M.Dhopavkar, M.M.Kshirsagar, "Word Sense Disambiguation using modified Maximum Entropy Approach", CIIT 2013, 3rd International conference on Computational Intelligence and Information Technology, Oct 18-19, 2013 ,Mumbai, India, pp 278-282.
[8] G.M.Dhopavkar, M.M.Kshirsagar, "Word Sense Disambiguation: A modified Maximum Entropy Approach", 4th International Conference Confluence 2013, The Next Generation Information Technology Summit on the theme S. M. A. C. (Social, Mobile, Analytics, and Cloud), Amity University, U.P.
[9] http://www.ling.upenn.edu/courses/Fall_2007/ling001/morphology.html
[10] www.languageinindia.com
[11] Zapata Becerra, Trabajo de Ascenso sin publicar. Mérida, Venezuela: Escuela de Idiomas Modernos, A. A. (2000), Handbook of general and applied linguistics, Universidad de Los Andes.
[12] S. Parameswarappa, V.N. Narayana, "Sense Disambiguation of Simple Prepositions in English to Kannada Machine Translation", IEEE conference, 2012 International Conference on Data Science & Engineering (ICDSEE) 978-1-4673-2149-5/12/$31.00 ©2012 IEE.
[13] Mitesh M. Khapra, Pushpak Bhattacharyya, Chauhan Shashank, Soumya Nair, Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting, Proceedings of ICON-2008: 6th International Conference on Natural Language Processing Macmillan Publishers,

India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008.

[14] Sudha Bhingardive, Samiulla Shaikh, Pushpak Bhattacharyya, Neighbors Help: Bilingual Unsupervised WSD Using Context, Sofia, Bulgaria, August 4-9, 2013. The 51st Annual Meeting of the Association for Computational Linguistics - Short Papers (ACL Short Papers 2013)