RESEARCH ARTICLE OPEN ACCESS

Handling Word Sense Disambiguation in Marathi Using a Rule Based Approach

Gauri Dhopavkar[1,2], Manali Kshirsagar[2], Latesh Malik[1]
[1]*Department of Computer Science and Engg., GHRCE, RTM Nagpur University, CRPF Gate, Hingna Road, Nagpur, India.*
[1]gauri.ycce@gmail.com
[3]lateshmalik@rediffmail.com
[2]*Department of Computer Technology, YCCE, Hingna Road, Nagpur, India.*
[2]manali_kshirsagar@yahoo.com

Abstract- In this paper, we have presented a detailed overview of the Word Sense disambiguation (WSD) efforts undertaken in India related to Indian Languages. Also in remaining sections we have discussed the method used by us for Marathi Language- WSD. This approach of WSD uses combination of Rules to disambiguate a word to provide a suitable sense with respect to the context.
Keywords*:* WSD, ambiguity, NLP, annotated corpora, rule based approach

I. INTRODUCTION

Natural Language Processing (NLP), is a branch of Artificial Intelligence. Language serves as the primary means by which people communicate and record information. It has the potential for expressing an enormous range of ideas, and for conveying complex thoughts. Natural Language Processing deals with the study of Natural Language computations. It tries to enable computers to be used as aids in analyzing and processing natural language, and to understand, by analogy with computers, more about how people process natural language. By understanding language processes in procedural terms, we can give computer systems the ability to generate and interpret natural language. This would make it possible for computers to perform linguistic tasks (such as translation), process textual data (books, journals, newspapers), and make it much easier for people to access computer-stored data. In the field of computational linguistics, some results have already been obtained however, a number of important research problems have not been solved yet.

Ambiguity is one of these problems which have been a great challenge for computational linguists. Something is ambiguous when it can be understood in two or more possible ways or when it has more than one meaning. Sometimes two completely different words are spelled the same. Word sense disambiguation (WSD) is the problem of determining in which sense a word (having a number of distinct senses) is used in a given sentence.

To a human it is very easy to take out suitable meaning from an ambiguous sentence because of Natural Intelligence. Although this seems obvious to a

human, developing algorithms to replicate this human ability is a difficult task.In general, people are unaware of the ambiguities in the language they use because they are very good at resolving them using context and their knowledge of the world. But computer systems do not have this knowledge, and consequently do not do a good job of making use of the context.If the ambiguity is in a sentence or clause, it is called **structural (syntactic) ambiguity**. If it is in a single word, it is called **lexical ambiguity**.

II. LITERATURE REVIEW

[1] In this paper authors report the use of supervised word sense disambiguation (WSD) mechanism with very less use of annotation. Here testing is done on Tourism and Health domain and mixed domain SemCor corpus and the approach is not restricted to specific set of target words. The concept of domain adaptation is used in which training happens in one domain and testing is in another domain which helps achieve good level of performance. Here 4 adaptation scenarios are used where adaptation using injection is of main interest. Authors of [2] claim that "I Can Sense It" is a simple online system interface developed to perform the exhaustive comparison between current state-of-the-art algorithms and existing state-of-the-art algorithms. This system supports 3 languages viz. ENGLISH, HINDI, MARATHI. This system is implemented using PHP5 and Javascript framework. As per

authors the main advantage of this system is that it can run

multiple algorithms in parallel, the whole system is user friendly, and runs variety of algorithms like IWSD, IMS, PPR, RB and so on along with the notification to the user when execution is done and output can be seen online in UKB format.

In the paper [3], authors showed that a multilingual system obtained on average a substantial relative error reduction when compared to the monolingual system. Three approaches to word sense disambiguation are presented that uses Wikipedia as a source of sense annotation. The three approaches are basic monolingual approach, knowledge-base approach, and data-driven approach. In this paper author addressed the sense tagged data bottleneck problem by using Wikipedia as a source of sense annotation. Here they built monological sense tagged corpora for four languages, using Wikipedia hyperlinks as sense annotation. Monolingual WSD system were trained on these corpora and were shown to obtained relative error reduction between 28% and 44% with respect to most frequent sense baseline. In order to reduce the reliance on the machine translation system during training, they explored the possibility of using the multilingual knowledge available in Wikipedia through its inter lingual links.

Authors of [4] report their method for performing verb disambiguation. In this paper author proposed a modified EM (Expectation Maximization) formulation using context and semantic relatedness which is used to determine the sense frequencies which gives 17% to 35% of more efficiency than unsupervised bilingual EM algorithm which have accuracy of 25-38% only. In the approach proposed here, we tackle with the problem of sense ambiguity by taking into account the words from the context of target word. This formulation solves the problem of "inhibited progress due to lack of translation diversity" and "uniform sense assignment, irrespective of context" that the previous EM based approach by khapra suffers from.

In paper [5], authors claim that Krudanta is the phenomenon of participial construction that needs adroit handling for high quality output in machine translation (MT). In this paper the krudanta in Marathi is processed using Finite State Machine (FSM). Authors have the opinion that Marathi Hindi MT can be improved with krudanta processing. The claim that FSM technique has high accuracy morphological analyzer for krudanta. Efficiency of 95% can be observed.

In this paper, the authors followed entropy based selective query disambiguation approach for Hindi language information retrieval[6]. The approach identifies the ambiguity in the query which is further disambiguated. This study summarizes the ambiguity detection approach as the prior ambiguity detection leads to conserve computation power. The survey of results concludes that several times even if the query consists of polysemous word, it is detected as

unambiguous. At the end, authors claim that Human intervention in lexical query disambiguation can be an effective tool for information retrieval applications. Detecting the ambiguity using the concept of Entropy and Threshold is found quite successful. [7] In this paper, the authors have presented an algorithm for domain specific all-words WSD. The scoring function which is used to rank the senses is inspired by the quadratic energy expression of Hopfield network and is employed by a greedy iterative disambiguation algorithm that uses only the words-disambiguated-so-far to disambiguate the current word in focus. Authors have used various parameters used for domain-specific WSD as below-

Wordnet-dependent parameters : belongingness-to-dominant-concept, conceptual-distance , semantic-distance

Corpus-dependent parameters : used are sense distributions and corpus co-occurrences. So here the authors represented three algorithms which combine the parameters described above to arrive at sense decisions as

Algorithm-1: Iterative WSD (IWSD),

Algorithm-2: Exhaustive graph search algorithm,

Algorithm-3: Modifying Page Rank to handle domain-specificity.

[8]In this paper, the authors discussed challenges involved in one of the toughest annotation tasks - sense marking.

The sense markers had the following options:

a. Marking the word with the exact sense

b. Marking with a subsuming sense

c. Marking with the closest sense

d. Marking with the exact sense even if the sense does not mention the particular word as a synset member

e. Creating a new sense

In the work reported in this paper, the corpus is taken from tourism domain and the Princeton wordnet (Version 2.1) is used as the sense inventory for English text while the Hindi and Marathi wordnets have been used for Hindi and Marathi texts respectively.

The corpus was independently tagged by different sense-markers and it was found that the inter annotator agreement on word sense disambiguation was about 80 % across the three languages, *i.e*., English, Hindi and Marathi.

In this paper [9], the authors made use of the Wordnet for Hindi, developed at IIT Bombay, which is a highly important lexical knowledge base for

Hindi. The main idea in the paper is to compare the context of the word in a sentence with the contexts constructed from the Wordnet and chooses the winner.The mentioned Wordnet contexts are built from the semantic relations and glosses, using the

Application Programming Interface created around the lexical data. The authors used some implementation modules such as BuildContext, NounSemanticExtractor, Tokenizer, Intersection, Rank. At the end, the authors found accuracy values ranging from about 40% to about 70%. The authors conclude that performance can surely be improved if morphology is handled exhaustively.[10]In this paper, the authors worked on Domain Specific Iterative Word Sense Disambiguation (WSD) for nouns, adjectives and adverbs in the trilingual setting of English, Hindi and Marathi. Starting from monosemous words they iteratively disambiguate bi, tri and polysemous words. After that they have combined corpus biases for senses along with information in wordnet graph structure to arrive at the sense decisions.The authors worked on the features including Domain Specific Sense Distributions, Dominant Concepts within a domain, Corpus co-occurrence frequency of senses, Conceptual distance between senses, Semantic Graph Distance.In this, the authors proposed the algorithm in which the logic is, "At each stage, the input to the algorithm consists of a set of disambiguated words." The algorithm used is an Iterative WSD. Also they used two different parameter settings for this. [11] In this paper, the authors showed empirically that a feedback term is neither good nor bad in itself in general; the behavior of a term depends very much on other expansion terms.As a principled solution to the problem, they proposed spectral partitioning of expansion terms using a specific term-term interaction matrix. For this, they used Partitioning algorithm. They demonstrated on several test collections that expansion terms can be partitioned into two sets and the best of the two sets gave substantial improvements in retrieval performance over model-based feedback.

III. OUR APPROACH

Our approach for solving ambiguity problem in Marathi text has following steps:

1. Collecting dataset from different domains.
2. Creating Features
4. Max. Entropy Model

a) Implementation of training dataset using Rules for typical case relations and special example based training set for cases without any clear case relation. Further Max. Entropy Model is applied to find the closest relationship between number of senses.

b) Implementation of testing data on trained system.

Examples:

In the following sentence the word विशाल has got two senses.

1. विशाल खेळत होता. (Vishal was playing)
2. विशाल आकाशात पक्षांप्रमाणे उडण्याची इच्छा आहे.
   (In the Big sky, (I/we/they) wish to fly like a bird)

First sense indicates that the word under consideration is a Proper Noun. (Name of some person)
Second sense indicates that it is an Adjective which adds more information about sky(that is **Big** Sky)
Other example includes: word जागा

1. चुका सुधारून आता तरी जागा हो.
   (Improve on mistakes and get up or be alert)
2. ही जागा व्यवसायासाठी अतिशय मोक्याची आहे.
   (This place is very much suitable for business)

In sentence 1 the word जागा indicates:
To remain awake, alert sense, and in sentence 2, it indicates "place" sense.

Initially we focused on framing the rules which were written to disambiguate an ambiguous word which has clear distinction of POS tags (like one sense is noun and other sense is verb or adjective or adverb).

Later on we found that some words have same POS tagging but different morphological structure and based on this we can find the accurate sense suitable in the sentence.

Following conventions are used –
W1, W2 indicate word 1 and word 2, P1, P2 indicate POS tagging, R1, R2 indicate root words etc.
For dealing with the ambiguity we have framed certain rules. Sample rules are as given below:

1. W1=W2
   P1!=P2
2. W1=W2
   P1=P2
   R1!=R2
3. W1!=W2
   R1!=R2
   P1!=P2
   R1=W2
   R2!=W1

IV. CONCLUSION

Thus using Rule Based approach we have designed system of Word Sense Disambiguation for

Marathi language. This system uses Source Language as Marathi. Input Text is obtained through any source. Marathi WordNet[12] is used to get the detailed features of each word in the sentence. Target language can be any other language. The approach has been tested using around 50 generic rules which apply to many sentence cases and it gives the accuracy of 80%.

REFERENCES

[1] 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)

[2] Salil Joshi, Mitesh M. Khapra, Pushpak Bhattacharyya, I Can Sense It: a comprehensive online system for WSD, COLING 2012, Mumbai, December 2012. Proceedings of COLING 2012: Demonstration Papers, pages 247–254.

[3] Mitesh Khapra, Anup Kulkarni, Saurabh Sohoney, Pushpak Bhattacharya, All Words Domain Adapted WSD: Finding a Middle Ground between Supervision and Unsupervision, Uppsala, Sweden, 11-16 July 2010. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1532–1541.

[4] Ganesh Bhosale, Subodh Kembhavi, Archana Amberkar, Supriya Mhatre, Lata Popale, Pushpak Bhattacharyya, Processing of Kridanta (Participle) in Marathi, International Conference on Natural Language Processing (ICON 2011), Chennai, December, 2011.

[5] Bharath Dandala, Rada Mihalcea, Razvan Bunescu, Multilingual Word Sense Disambiguation Using Wikipedia, The People's Web Meets NLP: Collaboratively Constructed Language Resources", Springer book series- Theory and Applications of Natural Language Processing, 2012.

[6] S. K. Dwivedi and Parul Rastogi, An Entropy Based Method for Removing Web Query Ambiguity in Hindi Language., Journal of Computer Science, 4: 762-767.

[7] Mitesh M. Khapra, Sapan Shah, Piyush Kedia, Domain-Specific Word Sense Disambiguation combining corpus based and wordnet based parameters, 5th International Conference on Global Wordnet (GWC 2010), Mumbai, Jan, 2010.

[8] Jaya Saraswati, Rajita Shukla, Sonal Pathade, Tina Solanki, Pushpak Bhattacharyya, Challenges in Multilingual Domain-Specific Sense-marking, Essential English Dictionary, 2nd Edition 2006 - Collins – 1995.

[9] Manish Sinha, Mahesh Kumar Reddy .R, Pushpak Bhattacharyya,Prabhakar Pandey, Lakshmi Kashyap, Hindi Word Sense Disambiguation, International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India, November, 2004.

[10] 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.

[11] Raghavendra Udupa, Abhijit Bhole, and Pushpak Bhattacharyy, A term is known by the company it keeps: On Selecting a Good Expansion Set in Pseudo-Relevance Feedback, Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10-12, 2009 Proceedings, pp 104-115.

[12] J.Ramanand, Akshay Ukey, Brahm Kiran Singh and Pushpak Bhattacharyya, Mapping and Structural Analysis of Multilingual Wordnets, IEEE Data Engineering Bulletin, 30(1), March 2007.