Word Sense Disambiguation in Gujarati Language

Bhagyada Desai, Krinal Naik, Preeti Bhatt

Abstract- **Word Sense Disambiguation (WSD) – a challenge of Natural Language Processing, for Gujarati language. All natural languages have words that mean different thing in different contexts. Human beings are generally good at sensing those ambiguities but it is difficult for computers to understand that. But computers can sense that by following certain algorithms and rules. Four methodologies were discussed in this paper for implementing word sense disambiguation: AI-based method, Knowledge-based method, supervised method and unsupervised method. Seeming that knowledge based method is more accurate, the algorithms that are of knowledge based method, are used for further processing. It includes the standard Lesk algorithm, the simplified lesk algorithm, the lesk algorithm with synonyms and a baseline algorithm. With the comparison of these algorithms it is finally decided to use simplified lesk algorithm and it will be implemented on Java.**

Index Terms— **Word Sense Disambiguation, Knowledge-based method, the Lesk Algorithm, the Simplified lesk algorithm**

I. INTRODUCTION

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things.

NLP has ambiguities or we can say challenges like

1. Word Sense Disambiguity
2. Multiword Ambiguity
3. Part-of-speech Tagging
4. Named Entity Recognition
5. Coreference Ambiguity
6. Scope Ambiguity

The major task was to find ambiguities in NLP. In previous researches the main focus was on English language and most of the work has already been done in English language and also in Hindi language. Our effort is to work on Guajarati language.

In figure 1, NLP trinity has been shown. NLP trinity includes Language, Problem and algorithm. There are different natural languages such as Hindi, French, English, Gujarati, etc. And every language has problems or challenges such as Part-of-speech tagging, morph analysis, parsing, etc. Each problem can be solved with algorithms such as HMM, CRF, MEMM.

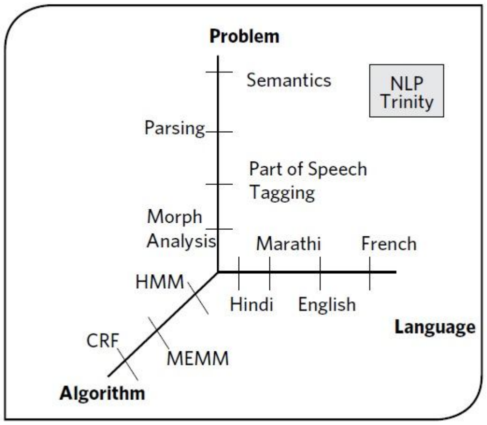

Fig 1: NLP Trinity

Word sense disambiguation is the task of automatically determining the correct sense of a word within a text. A word may be used in a variety of contexts with a variety of meanings. Each of these meanings is called a word sense. Now-a-days, determining the correct sense of a word has been the subject of research.

The belief is that if ambiguous words can be correctly disambiguated, IR performance will increase.

For example,

Son getup quickly.

In above example, the word Son has ambiguity of sense. There are two meanings of Son: male offspring and son of god. Here it is used in the context of first meaning.
Example,

રમેશભાઈ અને સુરેશભાઈ નાં પરિવાર વચ્ચે મૈત્રીપૂર્ણ સંબંધ

In above sentence, the word સંબંધ has ambiguity.

Manuscript received January 11, 2015

Bhagyada Desai, Department of Computer Science and Technology, Uka Tarsadia University, Bardoli, India, (+91)9537517755,
Krinal Naik, Department of Computer Science and Technology, Uka Tarsadia University, Bardoli, India, (+91)8238776563,
Preeti Bhatt, Department of Computer Science and Technology, Uka Tarsadia University, Bardoli, India, (+91)9898011188.

In Gujarati wordnet there are two meanings of the word “સંબંધ”. One is “નાતો”, “સગપણ” and other one is “લગાવ”. But here it is used in the context of first meaning.

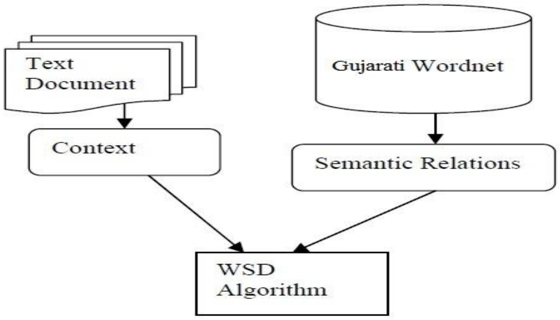

Fig 2: Strategy for WSD

In figure 2, it gives the pictorial description of the basic idea of the strategy. The idea behind using the intersection similarity measure is to capture the belief that there will be high overlap between the words in the context and the *related words* found from the wordnet lexical and semantic relations and glosses.

The applications of word sense disambiguation are:
- A search engine may use it to determine what a user wants to search for more accurately..
- A program that translates text in one language to another can find the correct translation of a homonym.

WSD is usually performed as a part of larger application. It plays a big role in applications that involves human languages. By improving the quality of WSD, we can improve the quality of all applications that utilize WSD.

II. RELATED WORK

In research for NLP, the initial step was to understand about NLP. i.e. what is NLP? Challenges of NLP, Solution towards it. For all the basic information about NLP, read this book [1] and it was helpful in understanding basic concepts about NLP.
The next introduction paper was giving brief idea about NLP that it is an area of research and application that depicts how computers can be used to understand natural languages [4]. It also describes what NLP researchers are doing to process further on this domain.
Natural Language Processing was described from the perspective of ambiguity. It was mostly based on English language but the examples from other different languages were also given. Then they have selected one particular challenge i.e. WSD and proposed a solution for it using specific algorithms.
Because of WSD, IR performance is being poor. If ambiguous words are correctly disambiguated, then IR performance can be increased. In this paper, they conclude present research to go beyond.
Till date most of the work has been done in Hindi and English language but very less work has been done in Gujarati language. To go further for research, Hindi wordnet as well as Gujarati wordnet is important to study. Now-a-days, WSD is a wide area which is very useful. Some of work has been done in Hindi Language and they try to disambiguate Hindi words. Likewise doing the same for Gujarati Language is perspective to initial step of this paper.
Five algorithms : The standard Lesk algorithm, the simplified Lesk algorithm, a Lesk algorithm variant using hypernyms, a Lesk algorithm variant using synonyms, and a baseline performance algorithm are there. While the baseline algorithm should have been less accurate than the other algorithms, testing found that it could disambiguate words more accurately than any of the other algorithms.
Various issues like scalability, ambiguity, diversity (of languages) and evaluation pose challenges to WSD solutions. The aim of this project is to develop a WSD technique which can handle all these issues with better accuracy and performance.
This paper study on various aspect of graph based approaches in word sense disambiguation. [13] Overview of supervised, unsupervised and knowledge based method is given in the paper.
A new algorithm for WSD which extends variations of Lesk WSD is explained that is simplified lesk and implemented on SemSevel. [15]. another paper [16] presents an adoption of Lesk dictionary-based WSD algorithm. The lexical database WordNet is employed. That provides rich hierarchy of semantic relations that algorithm can exploit.
Paper [17] gives a brief on three categories under which WSD algorithms can be classified: AI-based, Knowledge-based and corpus-based methods.
Approaches to WSD are often classified according to the main source of knowledge used in sense differentiation. Methods that rely primarily on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence, are termed dictionary-based or knowledge based. Natural language tends to be ambiguous.
One of the several approaches proposed in the past is Michael Lesk’s 1986 algorithm. This algorithm is based on two assumptions. First, when two words are used in close proximity in a sentence, they must be talking of a related topic and second, if one sense each of the two words can be used to talk of the same topic, then their dictionary definitions must use some common words.

III. ALGORITHMS

Four methodologies are there for WSD: AI-based method, knowledge-based method, supervised method and unsupervised method.

A. *AI-based Method*

In 1960's and 70's, majority of the systems were based on AI methodology. These systems tried to produce semantics for the whole sentence to determine its meaning and from this semantic representation, word ambiguity problem would be solved. But a sentence can have several possible interpretations; therefore the AI based system adopted a strategy for combining syntactic and semantic constraint.

Algorithm Category	**Advantage**	**Disadvantage**	**Example**
AI Methods	Some ideas formed the basis of all further work on the subject e.g. word window	Very domain Specific	
Knowledge Based	Accuracy	Rely on precompiled lexical knowledge resources	*The Lesk Algorithm*
Supervised	Accuracy	Dependent on pre-annotated corpora for training data	*The Naïve Bayesian Classifier*
Unsupervised	No pre-training necessary Works on multiple languages with no modification to the algorithm	Merely discriminates between word senses; not disambiguate word senses	

B. *Knowledge-based Method*

It depends on lexical resources to disambiguate word senses. The most common lexicon used in this method are: machine readable dictionaries, thesauri and computational lexicons. Most predominant machine readable dictionary is WordNet. One of the popular algorithms that utilizes machine readable dictionary is lesk algorithm. A number of variant of lesk algorithm have also been implemented and used. Comparison between the lesk algorithms are shown in table 2.

Algorithm	**Accuracy(%)**	**Words sent(%)**	**Recall(%)**
The lesk algorithm	33	25	32
Simplified Lesk	41	22	28
Hypernym Lesk	28	15	20
Synonym Lesk	34	27	34
Baseline algorithm	39	20	22

C. *Supervised Methods*

They are similar to AI methods. Supervised method to identify patterns and rules that are concerned with word senses and that can be applied by the algorithm to disambiguate words. It uses naïve Bayesian algorithm.

D. *Unsupervised Methods*

These methods are less accurate than others. These methods are only able to distinguish between senses and uses of words, not what the difference is.

E. *Selection of methodology and algorithm*

As the knowledge based method is more accurate, the lesk algorithm will be used for word sense disambiguation. As there are variants of lesk algorithms also, it is preferable to use simplified lesk algorithm.

F. *Protocols Comparison*

IV. CONCLUSION

Word Sense Disambiguation is sensing the correct word from the sentence. To sense the correct word from Gujarati Language, the Simplified lesk algorithm can be used. Actually all five variants of Lesk can be used but simplified algorithm is more accurate in sensing words than others. Its accuracy is almost 93%. It will be implemented on Java.

REFERENCES

[1] Daniel Jurafsky and James H. Martin, *Speech and Language Processing*, Prentice Hall, Englewood Cliffs, New Jersey (1999)

[2] Alexander Gelbukh, *Special issue: Natural language processing and its Application*, vol.46, Instituto Politécnico Nacional, Maxico (2010)

[3] Gobinda G. Chowdhury , "*Natural Language Processing*" , Dept. of Computer and Information Sciences University of Strathclyde, Glasgow

[4] Pushpak Bhattacharyya, "*Natural Language Processing: A Perspective from Computation in Presence of Ambiguity, Resource Constraint and Multilinguality*," CSI Journal of Computing, Vol.1 No2, 2012

[5] Mark Sanderson, "*Word Sense Disambiguation and Information retrieval,*" Proceedings of the 17th annual International ACM SIGIR conference on Research and development in information retrieval, July,1994

[6] Manisha Gupta, Seema Yadav, Shraddha Sharma and Dr.Surendra Yadav, "*Word Sense Disambiguation using Hindi Wordnet and Lesk Approach,*" IPASJ International Journal of Computer Science, Vol.1, Issue 6, pp. 13-16, December 2013

[7] Preeti Yadav and Sandeep Vishwakarma, "*Mining Association Rules Based Approach to Word Sense Disambiguation for Hindi Language,*" International Journal of Emerging Technology and Advanced Engineering, Vol.3, Issue 5, pp. 470-471 May 2013

[8] Manish Sinha, Mahesh Kumar Reddy .R, Pushpak Bhattacharyya, Prabhakar Pandey and Laxmi Kashyap, "*Hindi Word Sense Disambiguation,*" Department of Computer Science and Engineering Indian Institute of Technology, Bombay

[9] David Justin Craggs, "*An Analysis and Comparison of Predominant Word Sense*

Disambiguation Algorithms," Faculty of Computing, Health and Science Edith Cowan University

[10] Esha Patel, "*Word Sense Disambiguation,"* Kanwal Rekhi School of Information Technology Indian Institute of Technology, Powai, Mumbai

[11] Amal Zouaq, Michel Gagnon and Benoit Ozell, "*Can Syntactic and Logical Graphs help Word Sense Disambiguation?,"* Ecole Polytechnique de Montréal, Canada

[12] Andres Montoyo, Armando Suarez, German Rigau and Manuel Palomar, "*Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods,"* Journal of Artificial Intelligence Research, Vol.23, pp.299-330 March 2005

[13] Neha R. Kasture and Avinash Agraval, "*Graph based Algorithms for Word Sense Induction and Disambiguation,"* Proc. of the Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering, Nagpur

[14] Roberto Navigli, "*Word Sense Disambiguation: A Survey,"* ACM Computing Surveys, Vol. 41, No. 2, Article 10, Publication date: February 2009

[15] Pierpaolo Basile, Annalina Caputo and Giovanni Semeraro, "*An Enhanced LeskWord Sense Disambiguation Algorithm through a Distributional Semantic Model,"* Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1591–1600, Dublin, Ireland, August 23-29 2014

[16] Satanjeev Banerjee and Ted Pedersen, "*An Adaptive Lesk Algorithm for Word Sense Disambiguation Using WordNet,"* University of Minnesota, Duluth, MN, USA

[17] Onder Eker, "*DEVELOPING METHODS FOR WORD SENSE DISAMBIGUATION,"* Bogazici University 2007

[18] Neetu Sharma, Samit Kumar and Dr. S. Niranjan, "*Using Machine Learning Algorithm for Word Sense Disambiguation: A brief survey,"* International Journal of Computer Technology and Electronics Engineering Vol.2 Iss. 2

[19] Satanjeev Benerjee, "*Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet,"* University of Minnesota, Duluth, MN, USA