IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-12, May 2019 E-ISSN: 2347-2693

Word Sense Disambiguation

Dhanashree Surkar[1], Vedika Limje[2], Bhavana Gopachandani[3*]

[1]Dept. Computer Science and Engineering, Jhulelal Institute of Technology, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, India
[2]Department of Science and Technology, Delhi University, Delhi, India
[3]Department of Computational Sciences and Technology, Delhi University, Delhi, India

Corresponding Author: bhavana.lalwani06@gmail.com

Available online at: www.ijcseonline.org

Abstract— In natural language processing, there exist a problem of determining which "sense" (meaning) of a word is activated by the use of the word in a particular context. Natural language is ambiguous. It is an easy task for a human to understand and disambiguate the ambiguous words but a trivial task for a computer to do so. Ambiguity can occur at various levels of NLP- lexical, syntactic, and semantic and discourse level. The project here concentrates on Lexical Semantic ambiguity task. Lexical Semantic ambiguity takes place when a word/lexicon or a phrase has multiple meanings associated with it. Word Sense Disambiguation (WSD) aims to disambiguate the words which have multiple sense in a context automatically. Sense denotes the meaning of a word and the words which have various meanings in a context are referred as ambiguous words. It is the task of understanding the sense of an ambiguous word in a piece of context. It basically assigns the appropriate sense to a word depending on the particular context. where it occurs in an automated manner. In this report, we propose supervised Machine Learning approach for Word Sense Disambiguation task in English language. Knowledge-based resource like machine readable dictionaries can be used for disambiguation task. We used Natural Language Toolkit(NLTK) to perform various operation on given input given by user. Natural Language Toolkit(NLTK) library, is free, open source tool. This library does the core functioning for our application. It provides training data sets, Wordnet corpus, various tokenizers , lemmatizers and stemmers.

Keywords— Natural Language Toolkit (NLTK), Ambiguous word, Supervised Machine Learning, Lexical Semantic.

I. INTRODUCTION

Word sense disambiguation has been implemented in many Indian languages like Assamese, Manipuri, Tamil, Malayalam, Hindi, Kannada, Nepali, and Punjabi using various approaches like Supervised ,Knowledge-based, Unsupervised and semi-supervised approaches. It is necessary that WSD is to be implemented in English language. Ambiguity can occur at various levels of NLP- lexical, syntactic, and semantic and discourse level. The project here concentrates on Lexical Semantic ambiguity task. For example, consider the two sentences.
"The bank will not be accepting cash on Saturdays." And "The river overflowed the bank."

The word bank in the first sentence refers to the commercial (finance) banks, while in second sentence, it refers to the river bank. It is easy for humans to understand the meaning of "bank" word by understanding logic behind sentence. But difficult for machine to understand actual meaning of such ambiguous words. The ambiguity that arises due to this, is tough for a machine to detect and resolve.
A famous example is to determine the sense of pen in the following passage:

Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy. Word Net lists five senses for the word pen: 1.pen: a writing implement with a point from which ink flows.
2. Pen : an enclosure for confining livestock.
3. Play pen, pen -a portable enclosure in which babies may be left to play.
4. Penitentiary, pen a correctional institution for those convicted of major crimes.
5. Pen : female swan.

To deal with this situation we proposed Word Sense Disambiguation (WSD). Word Sense Disambiguation (WSD) is desktop application that provides solution to the ambiguity which arises due to different meaning of words in different context. WSD is implemented by using supervised approach. This application enables the user to find exact meaning of ambiguous word in a sentence. WSD is the task of understanding the sense of an ambiguous word in a piece of context. It basically assigns the appropriate sense to a word depending on the particular context where it occurs in an automated manner.

II. RELATED WORK

ArielRaviv and **ShaulMarkovitch** introduce Concept-Based Disambiguation (CBD), [1] a novel framework that utilizes recent semantic analysis techniques to represent both the context of the word and its senses in a high-dimensional space of natural concepts. The concepts are retrieved from a vast encyclopedic resource, thus enriching the disambiguation process with large amounts of domain-specific knowledge. In such concept-based spaces, more comprehensive measures can be applied in order to pick the right sense. Additionally, they introduce a novel representation scheme, denoted anchored representation,that builds a more specific text representation associated with an anchoring word. We evaluate our framework and show that the anchored representation is more suitable to the task of word sense disambiguation (WSD).

Jumi Sarmah and **Shikhar Sarma** propose a supervised Machine Learning approach-Decision Tree for Word Sense Disambiguation task in Assamese language.[2] A Decision Tree is decision model flowchart like tree structure where each internal node denotes a test, each branch represents result of a test and each leaf holds a sense label. J48 a Java implementation of C4.5 decision tree algorithm is taken for experimentation in their case. A few polysemous words with different real occurrences in Assamese text with manual sense annotation was collected as the training and test dataset.

Simone Paolo Ponzetto and **Roberto Navigli** present a methodology to automatically extend WordNet with large amounts of semantic relations from an encyclopedic resource, namely Wikipedia. We show that, when provided with a vast amount of high-quality semantic relations, simple Knowledge-lean disambiguation algorithms compete with state-of-the-art supervised WSD systems in a coarse-grained all-words setting and outperform them on gold-standard domain-specific datasets.

Manish Sinha, Mahesh Reddy and **Pushpak Bhattacharyya** present an effective method of construction of the Marathi Word Net using the Hindi Word Net both of which are being developed at IIT Bombay. They present Word Sense Disambiguation (WSD) of nouns in Hindi. The system has been evaluated on the Corpora provided by Central Institute of Indian Languages and the results are encouraging.

III. METHODOLOGY

In this system we will use Natural Language Toolkit (NLTK) to perform various operation on given input given by user. Natural Language Toolkit (NLTK) library, is free, open source tool. This library does the core functioning for our application. It provides training data sets, Wordnet corpus, various tokenizers, lemmatizers and stemmers.

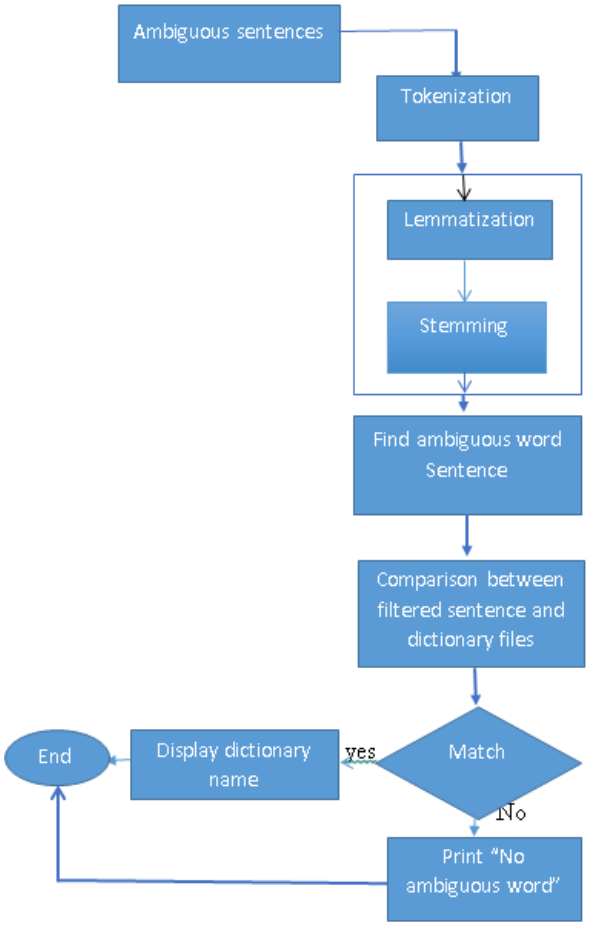

Fig III.I: Flow diagram of Word Sense Disambiguation

The application consists of five modules namely Tokenization, Lemmatization and Stemming, Ambiguous word, Dictionary creation and Training. Whenever user finds ambiguity in sentences , user will give those sentences as an input to the application. The application performs tokenization on inputs by removing stop words, special character and articles from sentences. Stopwords, are the high frequency words in a language which do not contribute much to the topic of the sentence. In English, such words include 'a', 'an', 'the', 'of', 'to' , etc.. We remove these words and focus on our main subject/topic to solve ambiguity.

Then application performs lemmatization on filtered sentence in which the verbs are converted into their first form. The plural words are converted into their singular form by performing stemming on filtered sentence. Now we have filtered sentence which does not contain plural words.

The third module identifies ambiguous word from filtered sentence. The application has been fed with data set files called dictionary files. The dictionary files are the text files. Dictionary Creation module contain list of dictionary files. The words inside dictionary files are compared with tokens of filtered sentence and if it match then application provide those dictionary file name as an output. At the end user get sense of the ambiguous word without human intervension.

III.I Natural Language Toolkit (NLTK)

Fig III.II: Natural Language Toolkit(NLTK)

Natural Language Toolkit (NLTK) library is free, open source tool. This library does the core functioning for our application. This toolkit is one of the most powerful natural language processing(NLP) libraries which contain packages to make machines understand human languages and reply to it with an appropriate response. It provides training data sets, Wordnet corpus, various tokenizers, lemmatizers and stemmers. NLTK also includes graphical demonstrations and sample data. It provides easy to use interface to over 50 corpora and lexical resources such as Wordnet.

IV. Results and Discussion

We will develop an application Word Sense Disambiguation(WSD) which disambiguate the word which have multiple sense in a context automatically by identifying and resolving conflict and provide output to user.Every time the user gives the input in terms of sentences, the application resolves the ambiguity and guesses the sense of the particular ambiguous word present in input sentence.

Thus we proposed an idea to train a machine in such a way that it can resolve ambiguity without any human intervention.

V .Conclusion and Future Scope

In this paper we proposed a supervised approach for English lexical semantic disambiguation task. In English language there exist many words which have multiple meanings. Such words are called ambiguous words. It is difficult for machine to understand actual meaning of such ambiguous words. The ambiguity that arises due to this, is tough for a machine to detect and resolve. To deal with this we proposed an idea that disambiguate word by identifying and resolving conflicts and provide output to user. In future, we intend to improve intelligent of system by working on logic based technique as sense dependencies are not always captured in syntactic structure of the sentence.

References

[1]Ariel Raviv and Shaul Markovitch, Concept-Based Approach to Word-Sense Disambiguation. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence.

[2]Jumi Sarmah and Shikhar Sarma, Decision Tree based Supervised Word Sense Disambiguation for Assamese. International Journal of Computer Applications (0975 – 8887)Volume 141 – No.1, May 2016

[3]Kalita, P. and Barman. AK, Word Sense Disambiguation: A Survey. International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 4 Issue 5 May 2015, Page No. 11743-11748V

[4]Ponzetto, S. P., and Navigli, R. 2010. Knowledge-rich word sense disambiguation rivalling supervised systems. In Proc. of ACL-1

[5]Sinha, M., Reddy R.M.K., Bhattacharyya, P., Pandey, P., P., Kashyap, L., www .cfilt .iitb .ac .in/wordnet/webhwn/papers/ HindiWSD.pdf

[6]Devendra Chaplot and Pushpak Bhattacharyya, Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser, AAAI 2015, Austin Texas, USA, 25-29 January, 2015

[7]https://towardsdatascience.com/a-simple-word-sense-disambiguation-application-3ca645c56357

[8]http://www.scholarpedia.org/article/Word_sense_disambiguation.

