International Journal of Engineering & Technology, 7 (2) (2018) 458-461

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i2.9649
Research paper

A brief review on Word Sense Disambiguation Approaches

Mrs. Swati G.Kale *, Dr. Ujjwala Gawande

A Yeshwantrao Chavan College of Engineering ,Nagpur
**Email: s12_kale@yahoo.com*

Abstract

To deal with the problem of text retrieval, machine translation, query processing, speech processing, the word sense disambiguation (WSD) performs very important role. WSD is an AI complete problem. This paper presents the significance, approaches along with work done in the field of WSD. Apple amount of work has been done in WSD for foreign languages but for Indian languages this issue is still about to concern. This paper tries to find out different limitations and challenges in WSD, based on which better approach will be decided by the researchers to solve the problem of WSD.

Keywords: *Word sense disambiguation, Artificial Intelligence(AI)*

1. Introduction

Word Sense Disambiguation (WSD) is defined as the task of finding the correct sense of the word in a given context. Natural languages contain words bearing multiple meaning and this happens in almost all the languages. As human beings, it is easily possible for us to arrive at the correct sense (meaning) of a word using the context in which it is used. However, the dependency between meaning and context is not well understood and hence computational representation of context is difficult. The necessary condition for a word to be disambiguated is that it should have multiple senses. Generally, in order to disambiguate a given word, we should have a context in which the word has been used and knowledge about the word, otherwise it becomes difficult to get the exact meaning of a word. Also, if the concept of a sense is not well defined, then it becomes very elusive task for WSD. The senses of a word differ from dictionary to dictionary. Some of them are coarse, while other provides a fine-grained distinction between possible senses. This may be the reason why there does not exist any WSD classifier which can give an accuracy of 100%. Not even human experts can agree on the sense of some words during manual disambiguation tasks. Marathi word sense disambiguation is also one of the areas, as 10-15% population of the India speaks Marathi and similar kind of language. Following example shows the WSD for Marathi language

S1: मला उत्तर दिशेला जायचं आहे

(Mala ***Uttar*** dishela jayacha ahe)
(I have to go to north direction)
S2: मला उत्तर द्यायचे आहे

(Mala ***Uttar*** dyayache ahe)
(I want to given an answer)

In sentence S1, “Uttar” refers to the direction and in sentence S2, it refers to an answer of something. As mentioned earlier, identifying correct sense of a word requires consideration of the context. In NLP, definition of context is closely related to specific task, domain and application. Most of the WSD techniques consider context as the text surrounding an ambiguous word, usually in a fixed size window keeping ambiguous word in the middle. This context is utilized in a variety of ways. The simplest is to consider the number of matching words between dictionary definition of words appearing in a test instance and the dictionary definitions of various senses of the word being disambiguated. This is possible only for exact matches. To overcome the problem, it is necessary to extend the context being matched. Other ways include extension of matching context with the help of semantic relations like synonym, hypernym, etc., combining local context with semantic similarity, utilizing statistical information gathered over some corpus. The majority of work on WSD is focused on English and other European languages and standard test corpora are available for these languages. The lack of such standards put a major obstacle on WSD research for Marathi and other Indian languages.

This paper talks about different approaches of word sense disambiguation in different languages. The rest of the paper is organized as follows: Section 2 discusses different approaches of WSD. Word sense disambiguation for Indian languages is given in section 3, whereas different findings in Literature and its limitations are explained in section 4, followed by the possible areas and proposed approach in section 5. Finally conclusions are drawn in section 6.

2. Different approaches of WSD

There are three broad categories of existing WSD techniques: knowledge-based, supervised and unsupervised.

2.1. The paper should have the following structure

Knowledge-based approaches are based on different knowledge sources like machine readable dictionaries or sense inventories, thesauri etc. Wordnet (Miller 1995) is the mostly used machine readable dictionaries in this research field. The first machine readable dictionary based algorithm is LESK [1, 2]. Initially an ambiguous word is selected from the sentence. Then, dictionary defi-

nitions (glosses) for the different senses of the ambiguous word and the other meaningful words present in the phrase are collected from an online Dictionary. Next, all the glosses of the key word are compared with the glosses of other words. The sense, for which the maximum number of overlaps occurs, represents the desired sense of the ambiguous word. Semantic similarity [3] is a method of knowledge based WSD, which is dependent on semantic distance between two related words. Selectional preferences [4-5] find information of the likely relations of word types, and denote common sense using the knowledge source. For example, Modeling-dress, Walk-shoes are the words with semantic relationship. In this approach improper word senses are omitted and only those senses are selected which have harmony with common sense rules.

2.2 Supervised WSD

Machine learning techniques are the supervised approaches applied to WSD. Training set is used for classifier to learn and this training set consist examples related to target word. A decision list [6] is a kind of supervised WSD, where the set of "if-then-rules" are used. Training sets are used in the decision list to induce the set of features for a given word. A decision tree [7] is used to denote classification rules in a tree structure that recursively divides the training data set. Internal node of a decision tree denotes a test which is going to be applied on a feature value and each branch denotes an output of the test. When a leaf node is reached, the sense of the word is represented (if possible). Support Vector Machine based algorithms [8] is another supervised approach used for mainly classification algorithm. The goal of this approach is to separate positive examples from negative examples with maximum margin and margin is the distance of hyperplane to the nearest of the positive and negative examples In order to apply the SVM to the WSD task, each nominal feature with possible values was converted into binary (0 or 1) features. If a nominal feature took the i-th value, then the i-th binary feature was set to 1 and all the other binary features were set to 0. The default linear kernel was used. Since SVMs handle only binary (2-class) classification, They built one binary classifier for each sense class. Neural networks and Naïve bays [9,10] classifier also used for WSD.

2.3 Unsupervised WSD

Instead of assigning meaning to the words, unsupervised approach differentiates word meaning based on the information. Context clustering method [11] is based on clustering techniques in which first context vectors are created and then they will be grouped into clusters to identify the meaning of the word. Co occurrence method creates co-occurrence graph with vertex V and edge E, where V represents the words in text and E is added if the words co-occur in the relation according to syntax in the same paragraph or text. For a given target word, first, the graph is created and the adjacency matrix for the graph is created. After that, the Markov clustering method is applied to find the meaning of the word [12].

3. Word sense disambiguation for Indian languages

A lot of work is carried out in English and European languages, but less amount of work is done on Indian languages. In spite of less work, reasonable work is done on Hindi, but regional languages are less attracted, due to non-availability of corpus.

3.1. WSD for Hindi Language

Sudha Bhingardive and Pushpak Bhattacharyya proposed WSD using Indo wordnet [13] to find exact sense of word. It describes EM (Expectation Maximization) algorithm based and EM-C WSD approach. Two languages are used to find the meaning of word. Indo wordnet for Hindi language on health and tourist data base is considered here. Results show that EM-C (Context) outperforms EM especially in case of verbs in all language-domain pairs. MAXNET classifier for word sense disambiguation of Hindi language for Noun is used by Pratisha Mathur [14]. Sense Annotated Hindi Corpus was used as training data. It was found that using this approach maxnet will classify some words accurately. The overall accuracy for this classifier is 60-65%. To Find exact sense of hindi word noun for different domain, Hindi Word Sense Disambiguation approach is proposed by Manish Sinha [15]. In this methodology, a word was assigned a sense with the use of the context in which it has been mentioned, the information in the Hindi Wordnet and the overlap between these two pieces of information. The sense with the maximum overlap was considered as the winner sense. Hindi corpora from the Central Institute of Indian Languages (CIIL), Mysore is used as a dataset. Another approach for WSD based on genetic algorithm is proposed to find Exact sense of word for Hindi language [16]. WorldNet for Hindi developed at IIT Bombay and a lexical knowledge base for Hindi is sued as a dataset. This system works on nouns only. Nisheeth Joshi [17] proposed a HMM based POS Tagger for Hindi. Part-of-speech tagging for Hindi language is a very complex task which is required in Machine translation WSD and Information Retrieval. A Hidden Markov based POS tagger based on HMM is used, which assigns the best tag to the word by calculating the forward and backward probabilities of tags along with the sequence provided as an input. The accuracy of the approach is found to be 92% for tourism domain. A Rule Based Hindi Part of Speech Tagger was proposed by Navneet Garg [18]. The said algorithm takes the Input text using file upload button or manually enter by user. Then it tokenize the input text word by word followed by Normalized the tokenized words. i.e. Separate out the punctuation marks and the symbols from the text. Next it searches for number tag ,date tag, time tag and abbreviation and then search in database for different input words and tag the word according to corresponding tag. Then different rules are applied to tag the unknown words.

3.2. Other regional languages

For the first time, Richard Singh and K. Ghosh [19] have given a proposed architecture for Manipuri Language in 2013. In this work, raw data was processed to get the features, which was used for training and testing. A 5-gram window was considered, taking the key word and the four other co-locational words to represent the context information. From this contextual information the actual sense of the focused word is disambiguated. Haroon, R.P. (2010) has given the first attempt for an automatic WSD in Malayalam. This was based on a hand devised knowledge source and the concept of conceptual density by using Malayalam WordNet as the lexical resource. Rakesh and Ravinder [20] have proposed a WSD algorithm for removing ambiguity from the text document in Punjabi documents. The authors used the Modified Lesk Algorithm for WSD. Two hypotheses have been considered in this approach. First, the co-occurring words in a sentence are to be disambiguated by assigning the most closely related senses to them. The second hypothesis is considered as, the definitions of related senses haveing maximum overlapAll title and author details must be in single-column format and must be centred.

3.3. WSD for Marathi language

Jyoti Singh [21] has used statistical approach for POS tagging i.e. they train and test their model for POS Tagging. They have calculated frequency and probability of words of given corpus. For tranning to system they used 7000 sentences (1,95,647) words from tourism domain. Marathi data base for tourism domain was used for the experimentation here. Based on the study, it is found that three Marathi POS taggers viz. Unigram, Bigram, Trigram and HMM gives the accuracy of 77.38%, 90.30%, 91.46% and 93.82% respectively. H.B.Patil [22] proposed a rule-based Part-of-Speech tagger for Marathi Language. Here the sentences are taken

as an input and generates the tokens. Once the token was generated, apply the stemming process to remove all possible affix and reduce the word to stem word. The root-words that are identified from the stem word are then given to morphological analyzer. The morphological analysis is carried out by dictionary lookup and morpheme analysis rules. A rule based solution for WSD in Marathi language is proposed by Gauri Dhopavkar. Another method of modified maximum entropy approach for performing WSD also proposed [23]. Sense of the ambiguous word was obtained by considering the overall sense of the sentence containing ambiguous word and maximum closeness between context and the target word. Sharvari Govilkar [24] proposed a part of Speech Tagger for Marathi Language. This system assigns parts of speech to each word, such as noun, verb, adjective, adverb etc in a sentence. If word has more than one tag then it is an instance of ambiguity, so such a word can be disambiguating by using small set of context rule.

4. Different findings in Literature and its limitations

WSD is hard for many reasons, three main reasons are: A sense inventory cannot be task-independent, Different algorithms for different applications, Word meaning does not divide up into discrete senses. Still, numbers of algorithms in different languages are proposed to solve the problem of WSD. These techniques are based on neural networks, support vector machine, machine learning, etc. Wordnet is typically used to solve the problem of WSD. Main challenges in WSD are as follows.

5. No large scale broad coverage and highly accurate word sense disambiguation system has been build
6. Accuracy achieved in previous work is 65-80%
7. System with a common sense is difficult to prepare
8. Word meaning is infinitely variable and context sensitive. It does not divide up easily into distinct or discrete sub meaning
9. Mostly, word Net has been widely adopted as sense inventory is too fine grained for many tasks and this make disambiguation very difficult
10. Word meaning is not divide into discrete sense

Because of these challenges, very less work is done on Marathi language. Now a day, researchers from IIT, Bombay is working on creation of database for Marathi language. Marathi WordNet is created to work on Marathi Language. To deal with the problem and to find the solution, which is helpful for Marathi people, it is necessary to develop an algorithm, which will perform best for classifying words in Word sense Disambiguation. The main steps in the proposed algorithm will be preprocessing of query, Word sense Disambiguation, Classification of the test document into the related classifier by using machine-learning approach and Finding correct sense and Retrieve the desired information.

5. Conclusion

This paper explains about different approaches in word sense disambiguation. It starts its discussion about definition and importance of WSD, followed by WSD methods for various foreign languages. Significance of studying WSD for Indian languages is also explained. Various methods and approaches for regional WSD including Hindi and Marathi are discussed. Though various methods are discussed for English, still there is a problem in getting accurate results. This difficulty increases for Indian languages as the less availability of dataset and tagger, as well as complexity of languages. Our aim is to overcome these difficulties and propose the efficient algorithm to get the results for Marathi language.

References

[1] Banerjee, S., Pedersen, T.,(2002) "An adapted Lesk algorithm for word sense disambiguation using WordNet", In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February

[2] Lesk, M.,(1986) "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", Proceedings of SIGDOC.

[3] Mittal, K. and Jain, A.,(2015)"Word sense disambiguation method using semantic similarity measures and owa operator", ictact journal on soft computing: special issue on soft – computing theory, application and implications in engineering and technology, january, 2015, volume: 05, issue: 02

[4] Diana, M.C., Carroll, J., "Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences",

Computational Linguistics, Volume 29, Number 4, pp. 639-654.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] Patrick, Y. and Timothy, B.,(2006) "Verb Sense Disambiguation Using Selectional Preferences Extracted with a State-of-the-art Semantic Role Labeler", Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), pages 139–148

[6] Parameswarappa, S. and Narayana V.N,(2013) "Kannada Word Sense Disambiguation Using Decision List", Volume 2, Issue3 May – June 2013, pp. 272-278.

[7] Singh, R. L., Ghosh, K. , Nongmeikapam, K. and Bandyopadhyay, S.,(2014) "A decision tree based word sense disambiguation system in manipuri language", Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, July 2014, pp 17-22.

[8] Buscaldi, D., Rosso, P., Pla, F., Segarra, E. and Arnal, E. S.,(2006)"Verb Sense Disambiguation Using Support Vector Machines:Impact of WordNet-Extracted Features", A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 192–195.

[9] Le, C. and Shimazu, A.,(2004)"High WSD accuracy using Naïve Bayesian classifier with rich features", PACLIC 18, December 8th- 10th, 2004, Waseda University, Tokyo, pp. 105-114.

[10] Aung, N. T. T., Soe, K. M., Thein, N. L.,(2011)"A Word Sense Disambiguation System Using Naïve Bayesian Algorithm for Myanmar Language", International Journal of Scientific & Engineering Research Volume 2, Issue 9, September-2011, pp. 1-7.

[11] Niu, C., Li, W., Srihari, R. K., Li, H., Crist, L.,(2004) "Context Clustering for Word Sense Disambiguation Based on Modeling Pairwise Context Similarities", SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July 2004.

[12] Alok Ranjan Pal1 and Diganta Saha, "WORD SENSE DISAMBIGUATION: A SURVEY", International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3, July 2015, pp. 1-16.

[13] Sudha Bhingardive , Pushpak Bhattacharyya, "Word Sense Disambiguation Using IndoWordNet", 2016. IJARCSSE, Volume 6, Issue 2, February 2016, pp 654-657

[14] Madhuri Bansal, Dr. Pratistha Mathur, "Word Sense Disambiguation using Maxnet Approach for Hindi Language",

[15] Manish Sinha, Mahesh Kumar Reddy, .R Pushpak Bhattacharyya,Prabhakar Pandey, Laxmi Kashyap, "Hindi Word Sense Disambiguation", 2003.

[16] Sabnam Kumari1, Prof. (Dr.) Paramjit Singh, "Genetic algoritm based hindi word sense disambigaution", IJCSMC, Vol. 2, Issue. 5, May 2013, pg.139 – 144

[17] Nisheeth Joshi1, Hemant Darbari2 and Iti Mathur, "HMM BASED POS TAGGER FOR HINDI", Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA – 2013, pp. 341–349, 2013.

[18] Navneet Garg, Vishal Goyal, Suman Preet, "Rule Based Hindi Part of Speech Tagger", Proceedings of COLING 2012 Demonstration Papers, pages 163–174, COLING 2012, Mumbai, December 2012

[19] Singh, R. L., Ghosh, K. , Nongmeikapam, K. and Bandyopadhyay, S.,(2014) "A decision tree based word sense disambiguation system in manipuri language", Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, July 2014, pp 17-22

[20] Kumar, R., Khanna, R.,(2011) "Natural Language Engineering: The Study of Word Sense Disambiguation in Punjabi", Research Cell: An International Journal of Engineering Sciences ISSN: 2229-6913 Issue July 2011, Vol. 1, pp. 230-238.

[21] Jyoti Singh, Nisheeth Joshi, "Development of Marathi Part of Speech Tagger Using Statistical Approach" 978-1-4673-6217-7/13/$31.00_c 2013, IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI).

[22] H.B.Patil, A.S.Patil, "Part-of-Speech Tagger for Marathi Language using Limited Training Corpora", International Journal of Computer Applications (0975 – 8887) Recent Advances in Information Technology, 2014.

[23] Gauri Dhopavkar, "Word Sense Disambiguation: A Modified Maximum Entropy Approach", the Next Generation Information Technology Summit 26th - 27th Sept. 2013, Amity University Uttar Pradesh, India IEEE conference 2014.

[24] Sharvari Govilkar, Bakal J. W, "Part of Speech Tagger for Marathi Language", International Journal of Computer Applications (0975 – 8887), Volume 119 – No.18, June 2015.