STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 5, Issue 1
www.stmjournals.com

An Analytical Review on Word Sense Disambiguation

Manali Parekh*, Prem Balani, Deven Gol
Department of Information Techology, G.H. Patel College of Engineering and Technology, Anand, Gujarat, India

Abstract

Due to high semantic ambiguity that is associated with language, word sense disambiguation is an open problem of natural language processing and ontology in computer linguistics. For instance, the word 'head' can mean the part of body and a person in charge of organization. The task of automatically assigning the most suitable meaning to a polysemous word can be defined as Word Sense Disambiguation. Several approaches have been done, from dictionary- and knowledge-based method that uses lexical resource like WordNet, to supervised-machine learning method that use trained classifier on manually sense-annotated corpus, to cluster based un-Supervised methods. These approaches have been applied for several languages like German, English, French, Chinese, and some Indian languages like Assamese, Hindi, Marathi and Malayalam. This survey aims to present about some aspect of word sense disambiguation and its approaches, focusing more on unsupervised graph based approach.

Keywords: *Word sense disambiguation, natural language processing, ambiguity, WordNet*

****Author for Correspondence*** E-mail: parekh.manali.mp@gmail.com

INTRODUCTION

Word Sense Disambiguation (WSD) is notorious problem in Natural Language Processing (NLP) referred as lexical semantic ambiguity [1]. Mainly there are two types of ambiguous words: (1) Polysemy, single word having multiple meanings and (2) Homonymy, different words having different meaning but spell or sound the same. For example, face is polysemy word having meaning body part and to deal with, while the words allowed and aloud are homonymy words. The procedure of automatically assigning the most suitable meaning to a ambiguous word can be delineated as Word Sense Disambiguation (Figure 1). There is no fix boundary in which a particular sense can be used. Hence it is known as AI complete problem [1]. Word sense disambiguation is task of natural language understanding in natural language processing [2].

Natural language processing is a buzz word in today's technology era. People spend more time on web to find and share information, however information is available in natural languages. In natural languages, words having multiple meanings are considered as polysemous. Normally, humans are using their common sense for labelling correct sense of polysemous words, but for machines, it is a complex task to convert unstructured data into appropriate data structure and to process it. Such ambiguous words originate huge problem in machine translation and other natural language processing tasks. Automatically assigning the most relevant label to the ambiguous word is known as word sense disambiguation. Word sense disambiguation used as in between steps for many applications [3]. In machine translation, computer needs to know meaning of every single word to translate from one natural language to another [4]. Same as for other applications like information retrieval [4], spelling correction [4], speech processing [5], text mining [6], lexicography [6], text summarization [6], question answering [7], speech synthesis [8], text processing [8], and grammatical analysis [8]. Until now, many techniques and enhancements have been proposed for word sense disambiguation problem [9]. These techniques and approaches can be broadly classified into:

- Knowledge Based Approach (Those that are based on machine readable dictionaries, Thesauri, WordNet).
- Corpus Based Approach (Those that uses raw corpus or Sense-tagged corpus).
- Hybrid Approach (Using best of knowledge based and corpus based approach).

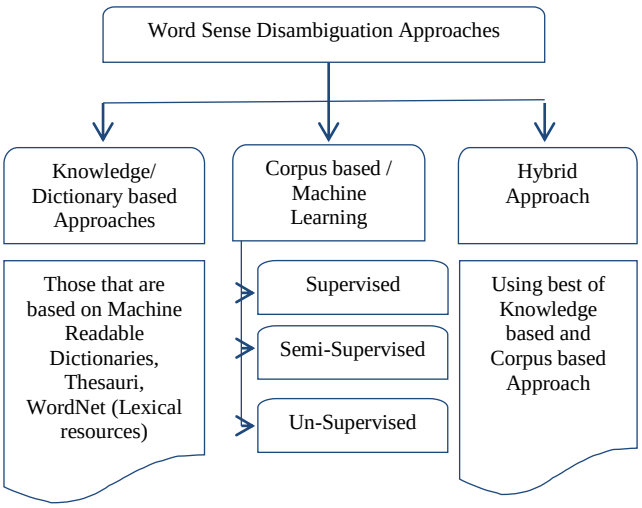

Fig. 1: *Word Sense Disambiguation Approaches.*

This survey aims to present word sense disambiguation approaches in general, focusing more on unsupervised graph based approaches for word sense disambiguation and also a review of the main research problems in recent articles in this field is presented.

The rest of the paper is organized as follows: Next part of the paper contains the literature survey done on word sense disambiguation approaches and their challenges. Concluding remarks are mentioned afterwards.

LITERATURE REVIEW

The literature survey indicates that, there are different approaches through which word sense disambiguation can be solved.

Knowledge/Dictionary Based Approach

It uses information provided by machine readable dictionaries, thesauri or lexical resources [4]. It can exploit grammar and manual rules for disambiguation.

Ledo-Mezquita *et al.* combined the original Lesk with simplified Lesk algorithm [10]. They calculated the score of the senses using this combined method with additional lexical resources like synonyms, dictionaries and ontology. It is local domain approach. They applied preprocessing steps, such as morphology normalization, and auxiliary word removal. The results obtained by the method are better than the results obtained by the baseline methods, namely, it gives 63% of precision.

Minca and Diaconescu proposed an algorithm of extended Lesk algorithm by building lexicon network from semantically tagged glosses [11]. They described several methods that transform lexicon network to better describe the natural language. Further, they explained a method to create semantic trees for each sense using such a lexicon network with reducing the computational volume for the best variant detection.

Sawhney and Kaur used modified Lesk algorithm with elimination of stop word and stemming [12]. They showed that precision can be increased if we change window size.

Singh and Siddiqui scrutinized the effect of stemming, stop word removal and context window size in the process of word sense disambiguation [4]. They evaluated the effect of varying window size with applying and without applying stemming and stop word removal on Hindi word sense disambiguation using supervised Lesk approach. They assumed that stop word removal and stemming will increase the number of content words and its chances of matching with the sense, and will have positive effect. The approach is evaluated on manually sense tagged corpus containing 10 polysemous nouns. They used Hindi WordNet for sense inventory.

Agarwal and Bajpai used knowledge based learning on correlation analysis [5]. They have done correlation analysis on word sense disambiguation features like co-occurrence

and collocation for each sense of ambiguous word. They used correlation matrix to determine the meaning of the target word, which provides the sense of target word with respect to correlated words. They have done stop word removing, punctuation removing and stemming as pre-processing tasks. The sense of target word is taken from Hindi WordNet. The algorithm is evaluated on 60 polysemous nouns, with the window size of 7, 8 and 9. They obtained 88.92% precision and showed that proposed approach gives better efficiency.

As we can see there are many variants applied for knowledge based. But in knowledge based, we only used definitions from dictionaries that are generally small and also require large amount of knowledge, and machine readable dictionary does not have proper nouns [8]. These are the challenges that the Knowledge based approach faces.

Corpus based Approach

Corpus based approach can be divided into three categories: Supervised, Semi-Supervised and Unsupervised. It uses information gathered from a semantically tagged or raw corpus [4]. These are machine learning based approaches.

Supervised

It falls under machine learning algorithms. It trains classifier using manually sense tagged corpora, dictionaries, WordNet, syntactic analysis and use this classifier to disambiguate new word in corpora. There are many supervised algorithms which have been invented, such as decision tree, decision list, Neural Network, Naïve Bayes, Support vector Machine (SVM) *etc.* [13].

Hao *et al.* proposed a novel word sense disambiguation approach based on SVM with automatic feature selection [9]. This learning method employs rich contextual feature to predict the proper sense. The method has achieved an excellent performance on the data released during the senseval-2 English lexical sampling task. Also, the proposed method is language independent, so can applied to other languages also.

Zhang *et al.* proposed WSD system based on Bayesian model in which vocabularies besides an ambiguous word are used in disambiguation process [14]. Then, a Bayesian classifier is built according to disambiguation features. A word segmentation tool is employed to segment an input sentence into many word units. This system was implemented on client-server framework.

Zopon and Joshi compared most of supervised approaches like decision tree, decision list, Adaboost, Naïve Bayes, Support vector Machine (SVM) and described that among all this approaches decision list is better having 62.12% of highest accuracy [13].

This supervised approach requires large amount of manually sense tagged corpora which are highly expensive and need human resource [15]. Because of that supervised approach faces knowledge acquisition bottleneck problem [16].

Semi-Supervised

It is hybrid procedure of supervised and unsupervised approach but it requires less sense-tagged corpora. It implements on unlabeled corpora. As supervised approach needs sense annoted corpora which are costly and on the other hand unsupervised provides less accuracy, so semi-supervised approach is proposed, that needs less amount of human interaction and provides higher accuracy. This system was implemented on client-server framework.

Broda and Piasecki proposed method that is based on clustering text snippets including ambiguous word [17]. For each cluster they found a core, the core is tagged with sense by human and used to create classifier. This classifier that was constructed for each word is then applied to corpora. They used two features lexical and topical feature to improve classifier.

Yu *et al.* solved two main problems that are acquisition and data sparseness of WSD using parallel corpora [18]. According to word alignment on the parallel corpora, the senses are defined.

Unsupervised

It is cluster based approach to analyze a pattern in enormous dataset and uses this pattern to classify data into clusters. It also known as Word Sense Induction. The selection of the most appropriate sense per word occurrence is then made through the use of graph processing algorithms that offer a degree of importance among the graph vertices [19]. Many graph based methods were applied with some improvements on different languages [6, 7, 20–22].

Jain and Lobiyal proposed an unsupervised word sense disambiguation method for Hindi sentence based on network agglomeration [7]. The approach can be given as they first created the sentence graph G for the given sentence. This sentence graph collectively represented all the interpretations of the sentence. Now from this sentence graph G, they created the interpretation graph G' for each of the interpretation of the sentence. To identify the desired interpretation they have computed network agglomeration for all the interpretation graphs. Thus the relevant interpretation having highest value of network agglomeration is identified. They evaluated their propose approach on sense tagged corpus. They compared their performance with unsupervised graph approach and showed their superiority.

Nandanwar proposed the graph-based Unsupervised Word Sense Disambiguation Algorithm to resolve the ambiguity of a word in Hindi Language [6]. The proposed algorithm proceeds incrementally on a sentence by sentence basis. There are two steps of proposed approach:

Step 1: Build the graph G=(V, E) for each target sentence. These sentences are part-of-speech tagged, so algorithm only considers the content words.

Step 2: The given graph G, they select the most appropriate sense of the word from the WordNet by performing Depth First Search and ranking each vertex in the graph G according to its importance which optimizes in terms of graph connectivity.

Rintyarna and Sarno compared the Lesk algorithm with Adapted Lesk algorithm to improve the performance of weighted graph-based word sense disambiguation [23]. They used Lesk and Adapted Lesk algorithm because both have the highest coverage of part-of-speech, as they can measure with the different part-of-speech. In this approach, they applied POS tagging and filtering as pre-processing. Then every sense of word from WordNet becomes edge of the graph. The similarity measure between word senses is then assigned by several similarity measures. Whenever the similarity between the senses is not zero, an edge is created to connect the node associated with the senses. The value of the similarity is then assigned as the weight of such vertices. The final step is to determine the most important sense by determining the edge with the maximum score using Indegree Algorithm.

Zhang *et al.* used unsupervised evolutionary Genetic algorithm for WSD to maximize the semantic similarity of the words that took from WordNet [24]. They compared the waighted Genetic WSD with GWSD.

Vaishnav applied Genetic Word Sense Disambiguation for Gujarati Language [25]. It was the first effort taken for Gujarati Language. They use Indo WordNet.

Sinha and Mihalcea also used semantic similarity using graph centraity based method [26].

Preotiuc-Pietro and Hristea described the problem of feature selection for unsupervised WSD performed with underlying Naïve Bayes model [27]. They proposed N-gram feature. While creating features from unlabeled data, they helped a simple, basic knowledge-lean disambiguation algorithm to significantly increase its accuracy as a result of receiving easily obtainable knowledge.

CONCLUSION

There are various approaches for word sense disambiguation that can be broadly classified into three groups (knowledge based approach, corpus based approach and hybrid approach). Dictionary based approach depends upon large amount of knowledge, also dictionary explanations are small, machine readable

dictionaries do not include proper noun, so knowledge based technique suffers from knowledge acquisition bottleneck problem.

If we applying word sense disambiguation on language in which we can get sense tagged corpus, then supervised machine learning method is better. This technique requires large amount of manually sense tagged corpora which are highly expensive and need human resource. However, unsupervised approach is difficult as compared to supervised approach because of no labelled data; so it gives lower accuracy than supervised approach.

Also, if we apply pre-processing steps including stop word removing, stemming, tokenization, part-of-speech tag with features like collocation and co-occurrence then these approaches gives better results with higher efficiency. Also, many languages remain for automatic word sense disambiguation process like Gujarati. Finally, we can conclude that, the field of word sense disambiguation is open for researchers to be further researched. Therefore, to improve the performance of word sense disambiguation, there is need for semantically rich, context aware, widely applicable word sense disambiguation methods, to disambiguate ambiguous word with regards to the contexts and concepts.

REFERENCES

1. Pathak R, Thankachan B. Natural Language Processing Approaches, Application and Limitations. *International Journal of Engineering Research & Technology (IJERT)*. Sep 2012 , Vol-1(7).
2. Indurkhya, Nitin, Damerau Fred J, editors. *Handbook of Natural Language Processing*. Chapman and Hall/CRC; 2010. Vidhu Bhala R, Abirami S. Trends in Word Sense Disambiguation. Springer. Artificial Intelligence Review.sssAugust 2014, Volume 42, Issue 2, pp 159–171.
3. Singh S, Siddiqui T. Evaluating Effect of Context Window Size, Stemming and Stop Word Removal on Hindi Word Sense Disambiguation. *Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on.* IEEE. 2012.
4. Agarwal M, Bajpai J. Correlation Based Word Sense Disambiguation. *Contemporary Computing (IC3), 2014 Seventh International Conference on.* IEEE. 2014.
5. Nandanwar L. Graph Connectivity for Unsupervised Word Sense Disambiguation for HINDI Language. *IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems*. 2015.
6. Jain A, Lobiyal D. Unsupervised Hindi Word Sense Disambiguation based on Network Agglomeration. *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on.* IEEE. 2015.
7. Sarika, Sharma D. A Comparative Analysis of Hindi Word Sense Disambiguation and its Approaches. *International Conference on Computing, Communication and Automation*. IEEE. 2015.
8. Hao W, Guilin C, Lianxian X. Highly Accurate SVM Model with Automatic Feature Selection for Word Sense Disambiguation. *J Syst Eng Electron*. 2004; 15: 723–727p.
9. Ledo-Mezquita Y, Sidorov G, Cubells V. Combined Lesk-based Method for Words Senses Disambiguation. *Proceedings of the 15th International Conference on Computing*. 2006.
10. Minca A, Diaconescu S. An Approach to Knowledge-Based Word Sense Disambiguation Using Semantic Trees Built on a WordNet Lexicon Network. *Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on.* IEEE. 2011.
11. Sawhney R, Kaur A. A Modified Technique for Word Sense Disambiguation Using Lesk Algorithm in Hindi Language. *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on.* IEEE. 2014.
12. Zopon B, Joshi S. Empirical Comparative Study to Supervised Approaches for WSD Problem: Survey. *IEEE Canada International Humanitarian Technology Conference*. 2015.

13. Zhang C, He S, Gao X. A Word Sense Disambiguation System Based on Bayesian Model. *4th International Conference on Computer Science and Network Technology*. IEEE. 2015.
14. Agirre E, Martinez D, Lacalle O, *et al.* Two Graph-Based Algorithms for State-of-the-Art WSD. *ResearchGate*. Jan 2008.
15. Chandra G, Dwivedi S. A Literature Survey on Various Approaches of Word Sense Disambiguation. *2nd International Symposium on Computational and Business Intelligence*. IEEE. 2014.
16. Broda B, Piasecki M. Semi-supervised Word Sense Disambiguation Based on Weakly Controlled Sense Induction. *Proceedings of the International Multiconference on Computer Science and Information Technology*. 2009; 4: 17–24p.
17. Yu M, Wang S, Zhu C, *et al.* Semi-supervised Learning for Word Sense Disambiguation using Parallel Corpora. *Eighth International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE. 2011.
18. Tsatsaronis G, Varlamis I, Norvag K. An Experimental Study on Unsupervised Graph-based Word Sense Disambiguation. Unpublished.
19. Cho J. A Graph-based Word Sense Disambiguation Using Collocation. *Adv Sci Technol Lett*. 2015; 86: 77–80p.
20. Zungre N, Dhopavkar G. Sense Disambiguation For Marathi Language Words Using Decision Graph Method. *IEEE Sponsored World Conference on Futuristic Trends in Research and Innovation for Social Welfare*. 2016.
21. Arab M, Zolghadri M, Mostafa S. A Graph-based Approach to Word Sense Disambiguation: An Unsupervised Method Based on Semantic Relatedness. *24th Iranian Conference on Electrical Engineering*, IEEE. 2016.
22. Rintyarna B, Sarno R. Adapted Weighted Graph for Word Sense Disambiguation. *Fourth International Conference on Information and Communication Technologies*, IEEE. 2016.
23. Zhang C, Zhou Y, Martin T. Genetic Word Sense Disambiguation Algorithm. *Second International Symposium on Intelligent Information Technology Application*. IEEE. 2008.
24. Vaishnav Z. Gujarati Word Sense Disambiguation using Genetic Algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC)*. Jun 2017; 6: 635–639p.
25. Sinha R, Mihalcea R. Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. Unpublished.
26. Preotiuc-Pietro D, Hristea F. Unsupervised Word Sense Disambiguation with N-Gram Features. Springer; *Artificial Intelligence Review, 2012; 41(2): 241–260p.*

Cite this Article
Manali Parekh, Prem Balani, Deven Gol. An Analytical Review on Word Sense Disambiguation. *Journal of Artificial Intelligence Research & Advances*. 2018; 5(1): 32–37p.