ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.12

A Novel Natural Language Processing to Solve Word Sense Disambiguation by Using Supervised Machine Learning Technique

Chandrakanth Rathod Karahari[1], Assoc.Prof. Dr.Nandini N[2]
Department of Computer Science
Dr.AIT Institute of Technology Bangalore, India

Abstract:
Ambiguity in natural language is a big problem in language processing tasks by machine. The ambiguity can be found in sentence level or word level. There are various types of ambiguities such as lexical, syntactic, semantic and anaphoric that can be found in written text and/or spoken text of a language. In our research proposal focused its study on the lexical ambiguity in the written text and use of knowledge based contextual overlap count WSD method utilizing the information from the WordNet for sense disambiguation. The knowledge-based WSD approaches use information from the resources such as dictionaries, thesauri, ontology, collocation etc to disambiguate the sense of a polysemy word in a given context

I. INTRODUCTION

Ambiguity in natural language is a big problem in language processing tasks by machine. The ambiguity can be found in sentence level or word level. There are various types of ambiguities such as lexical, syntactic, semantic and anaphoric that can be found in written text and/or spoken text of a language.

In our research proposal focused its study on the lexical ambiguity in the written text and use of knowledge based contextual overlap count WSD method utilizing the information from the WordNet for sense disambiguation. The knowledge-based WSD approaches use information from the resources such as dictionaries, thesauri, ontology, collocation etc to disambiguate the sense of a polysemy word in a given context [1]. In early days during 1980s, it was noticed that these resources contains less information about a word and this less information was not sufficient to disambiguate the correct sense. When the lexical database WordNet was developed in 1990s, at Princeton University, the problem of lack of sufficient information was seemed solved [2], [3]. It is because the WordNet provides the more information than dictionaries do. The WordNet is a lexical database for English language [4]. It organizes the nouns, verbs, adjectives and adverbs into the groups of synonyms each describing a distinct concept [5]. After the popularity of WordNet, WordNets in other foreign languages such as German, French, Spanish, etc. were developed and are massively being used in various Natural Language Processing (NLP) tasks including Word Sense Disambiguation.

The WordNet organizes the information using the various relations such as synset, gloss, hypernym, hyponym, meronym etc [6]. During the sense disambiguation process of the contextual overlap count knowledge-based WSD approaches, context and sense bags are formed by collecting the information from these relations. Then, overlaps between the context bag and each sense bag are counted [1]. The sense which has the maximum overlap with the context is determined as the correct sense. During this disambiguation, the information taken from the WordNet may induce noise information in the sense bag of wrong sense of the polysemy word. The noise information is the information induced from the WordNet in the sense bag of the wrong sense of a polysemy word, due to which the wrong sense of the polysemy word will have the maximum overlap with the context resulting in the wrong sense disambiguation. This noise information is the main cause for contextual overlap count of existing knowledge-based WSD approaches to obtain low accuracy.

This was the main motivation towards our research proposal. In this research, we investigated the cause of this noise information and resolved the problem in our new lexical database Polysemy WordNet.Problem Identified in using WordNet for Word sense disambiguation:

We looked at various contextual overlap count knowledge-based WSD approaches that uses the information taken from the various relations such as synsets, glosses, examples, hypernyms, holonymy, hyponymy, troponymy, meronymy, attribute etc taken from the WordNet for sense disambiguation. From the detailed study on these approaches, we found the following important points: The WordNet relations- synset, glosses of synset, attribute relation, hypernyms, hyponym, troponym, holonym, meronym, also see, similar to and pertainym, domain relations for nouns, verbs and adjectives have been used to collect the information for sense disambiguation to form the context and sense bags.

The hyponymy relation does not seem to contribute to the sense disambiguation. The inclusion of definitions from hyponymy relation decreased the accuracy [7].The higher levels of the WordNet hierarchy are less semantically related than a lower level. The relatives in a synonym class tend to share similar context at higher level hierarchy. For these reasons, use of glosses of relatives of a word in higher level is not appropriate and the definitions in the WordNet still dont

contain sufficient information for sense disambiguation. The results from experiments indicate the higher level hypernyms/hyponyms are not useful for all words for sense disambiguation [9].

Very few words are overlapped with context even the full hypernymy hierarchy is used [10]. Information from synonyms, hyponyms, hypernyms, definitions of its synonyms and hyponyms and its domains are not sufficient for sense disambiguation [8], [11].

When the deeper level hypernyms [12] are used, the accurately disambiguated words are also started to inaccurately disambiguated. If only first level hypernyms are used, they contain very less information for sense disambiguation. If all level of hypernyms are used, they contain more common information. This common information for each sense do not help to disambiguate the sense rather they introduce the noise information which causes the wrong disambiguation of the sense [12] and [13].

Problem Statement:

Noise Information and Wrong Disambiguation:

- The use of common hypernym introduces the noise information and this noise information causes the wrong sense disambiguation. In addition, the information excluding the hypernym hierarchy contains less information which is insufficient for sense disambiguation. These findings forced us to think" can we organize the polysemy words and related words in such a way that it resolves the problems stated (WordNet) in previous section and increases the accuracy of WSD algorithms?" which is our research question.
- Disambiguation Depends on the Gloss's Words:
- While defining the gloss of a word, the words which are used to define it, does not have any rule. The words, that are used to define the gloss, determine the degree to which the conceptual overlap count share common words with given context. This does not seem fair for all contexts. Therefore, the sense disambiguation using the gloss in the WordNet depends on which words are being used to define the glosses of the WordNet. We, therefore, intend to develop a lexical resource where the sense disambiguation does not depend on the words used to define the gloss of a word in a lexical database.

II. Literature survey

2004

Paper 1:

It takes time to process purchases and as a result a queue of customers may form. The pricing and capacity (service rate) decisions of a monopolist who must take this into account are characterized. They find that an increase in the average number of customers arriving in the market either has no effect on the price, or else causes the firm to reduce the price in the short run. In the long run the firm will increase capacity and raise the price. When customer preferences are linear, the equilibrium is socially efficient. When preferences are not linear, the equilibrium will not normally be socially efficient.

REF 14: - H. Chen and M. Frank, "Monopoly Pricing When Customers Queue," IIE Trans., vol. 36, no. 6, pp. 569-581, 2004.

2008

Paper 1

This paper proposes a novel approach to improve the kernel-based Word Sense Disambiguation (WSD). Author first explain why linear kernels are more suitable to WSD and many other natural language processing problems than translation-invariant kernels. Based on the linear kernel, two external knowledge sources are integrated. One comprises a set of linguistic rules to find the crucial features. For the other, a distributional similarity thesaurus is used to alleviate data sparseness by generalizing crucial features when they do not match the word-form exactly

REF 15:- Peng Jin, Fuxin Li, Danqing Zhu, Yunfang Wu and Shiwen Yu, "Exploiting external knowledge sources to improve kernel-based Word Sense Disambiguation," Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. International Conference on, Beijing, 2008, pp. 1-8

2011

Paper 1:

Concept Description Language (CDL) is a common language that represents the semantics of content in a simple and structured manner. In particular, it is intended to describe Natural Language (NL) texts in a format that can be understood and processed by computers. Since words with multiple meanings can be found from texts, it becomes necessary to perform Word Sense Disambiguation (WSD) in order to achieve a correct representation. This paper presents a WSD approach that determines best candidates for word meanings and contributes to a semi-automatic conversion of NL into CDL. Author perform preliminary experiments by evaluating the approach with some test sentences and comparing with other WSD methods

REF16:- F. Tacoa, H. Uchida and M. Ishizuka, "A Word Sense Disambiguation Approach for Converting Natural Language Text into a Common Semantic Description," Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on, Pittsburgh, PA, 2010, pp. 478-486

2012

Paper 1

Word Sense Disambiguation (WSD) has become even more important research area in recent years with the widespread usage of Natural LanguageProcessing (NLP) applications. WSD task has two variants: "Lexical Sample" and "All Words" approaches. Lexical Sample approach disambiguates the occurrences of a small sample of target words that were previously selected, while in the latter all the words in a piece of text are disambiguated. In the scope of this work, a Lexical Sample Dataset for Turkish has been prepared. As a first step, highly ambiguous words in Turkish have been selected. Collection of text samples for chosen words has been completed. Five taggers have annotated the word senses.

REF 17:- B. İlgen, E. Adali and A. C. Tantuğ, "Building up Lexical Sample Dataset for Turkish Word Sense

Disambiguation," Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on, Trabzon, 2012, pp. 1-5.

2014

Paper 1

Hindi WordNet, a rich computational lexicon is widely being used for many Hindi Natural Language Processing (NLP) applications. However it does not presently provide exhaustive list of senses for every word, which degrades the performance of such NLP applications. In this paper, author propose a graph based model and its associated techniques to automatically acquire words' senses. In the literature no such method is available which is capable of automatically identify the senses of the Hindi words. Author use a Hindi part of speech tagged corpus for building the graph model. The linkage between noun-noun concepts is extracted on the basis of syntactic and semantic relationships. All of the senses of a word including the sense which is not present in Hindi WordNet are extracted.

REF 18:- A. Jain and D. K. Lobiyal, "A new method for updating word senses in Hindi WordNet," Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on, Ghaziabad, 2014, pp. 666-671

Paper 2

Word Sense Disambiguation (WSD) is the process of selecting the correct sense for a word in a context. WSD has become a growing research area in the field of Natural Language Processing (NLP). Over the decades, lot of studies had been carried out to suggest different approaches for WSD process. A break-through in this field would have a significant impact on many relevant web-based applications, such as information retrieval (IR), information extraction etc. This paper describes various approaches of WSD like knowledge based approach, supervised approach, unsupervised approach and semi-supervised approach. It also describes various applications of WSD like information retrieval (IR), machine translation (MT), speech recognition, computational advertising, text processing, classification of documents and biometrics.

REF 19:- G. Chandra and S. K. Dwivedi, "A Literature Survey on Various Approaches of Word Sense Disambiguation," Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium on, New Delhi, 2014, pp. 106-109

2015

Paper 1

Word Sense Disambiguation (WSD) is an important and challenging task in the area of Natural Language Processing (NLP) where the task is to find the correct sense of an ambiguous word given its context. There have been very few attempts on WSD in Bengali or in Indian languages. The k- Nearest-Neighbor (k-NN) algorithm is a very well-known and popular method for text classification. The k-NN algorithm determines the classification of a new sample from its k nearest neighbors. In this paper, author present how k-NN algorithm can be effectively applied to the task of WSD in Bengali.

REF 20: - R. Pandit and S. K. Naskar, "A memory based approach to word sense disambiguation in Bengali using k-NN method," Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on, Kolkata, 2015, pp. 383-386

Paper 2

Word Sense Disambiguation (WSD) is crucial and its significance is prominent in every application of computational linguistics. WSD is a challenging problem of Natural Language Processing (NLP). Though there are lots of algorithms for WSD available, still little work is carried out for choosing optimal algorithm for that. Three approaches are available for WSD, namely, Knowledge-based approach, Supervised approach and Unsupervised approach. Also, one can use the combination of given approaches. Supervised approach needs large amounts of manually created sense annotated corpus which takes computationally more amount of time and effort. Knowledge-based approach requires machine readable dictionaries, sense inventories, thesauri, etc, which are dependent on own interpretation about word's sense; Whereas unsupervised approach uses sense-unannotated corpus and it is based on the phenomenon of working that words that co-occur have similarity. This research is for Hindi language which uses Hierarchical clustering algorithm with different similarity measures which are cosine, Jaccard and dice, the result of clusters is overlapped with Hindi WordNet a product of IIT Bombay which improves result of word sense disambiguation as clustering does grouping of words which are similar.

REF 21:- N. Patel, B. Patel, R. Parikh and B. Bhatt, "Hierarchical clustering technique for word sense disambiguation using Hindi WordNet," 2015 5th Nirma University International Conference on Engineering (NUiCONE), Ahmedabad, 2015, pp. 1-5

.

Paper 3

The internet has become the most important knowledge source, because the popularity of computers and networks. Always the users use some keywords to find out the related topics, at same time spend a lot of time to finding what they really want, and because of the imprecise results of search in the internet, most studies of web mining method are trying ring to improving the accuracy of the information gotten from web search. In this work author implemented the approaches of our proposed implement master-slave voting model, and in this model author selected five robust supervised algorithms, so in this paper author implemented empirically these approaches using WordNet, Senseval-3, and tried make comparative between them

REF 22:- B. F. Zopon AL Bayaty and S. Joshi, "Empirical comparative study to supervised approaches for WSD problem: Survey," Humanitarian Technology Conference (IHTC2015), 2015 IEEE Canada International, Ottawa, ON, 2015, pp. 1-7.

Paper 4

Word Sense Disambiguation (WSD) has become a popular method for solving the ambiguous meaning of the words in Information Retrieval (IR) field area. Under the Natural Language Processing (NLP) community, WSD has been described as the task which able to select the appropriate

meaning among the ambiguous meanings to a given word. Among three approaches, supervised based, unsupervised based and knowledge based approaches to WSD, this paper focuses on both supervised based and knowledge based approaches by proposing new Jaccard coefficient-based WSD algorithm to overcome the vocabulary miss match problem. WordNet and corpus external knowledge resources are utilized as the sense repositories by linking up with the new WSD algorithm to consider additional semantic for WSD.

REF 23:- S. M. Tyar and T. Win, "Jaccard coefficient-based word sense disambiguation using hybrid knowledge resources," 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE), Chiang Mai, 2015, pp. 147-151.

III. Objectives

The goal of the research work is to present an accurate word sense disambiguation for English and Indian language.To adopt NLP to solve word sense disambiguation for various Indian languages including English.Knowledge based model is presented for Word sense disambiguation.

To propose a model that is robust to noisy data and changes in sense frequency distributions and corpus domain (or genre).

To develop an accurate PolysemyWordNet in order to overcome the accuracy of WordNet.The proposed Knowledge based WSD model will be evaluated considering different corpus (language).

The accuracy of proposed Knowledge based WSD will be evaluated considering both PolysemyWordNet and WordNet Dictionary.The proposed Knowledge based WSD will be compared with supervised model.

IV. Methodology

The proposed WSD model consisting of following phase:

Related Words:

We believe that each sense of polysemy word has some distinct words related only to that sense and describe the sense. We call these words as related words for the sense. We call these words as related words for the sense. For example, the words ”copy”, ”write” etc. are the related words for the sense ”writing implement with a point from which ink flows” of polysemy word ”pen”. Therefore, the word ”pen” is used with ”copy”, then it is a sufficient evidence to understand the meaning of the word ”pen” as a writing implement. This does not require the overlap counting that may cause the introduction of noise information. Only the task we need to do is that to find the such related words for each sense of the polysemy word and organize in a new lexical database. These words are linked with appropriate senses of polysemy words and the resulted net of senses of polysemy words and related words can be used for sense disambiguation with high accuracy. The related words for a word can be a noun, a verb, an adjective and an adverb.

Finding Related Words:

These related words are the main key for our disambiguation method and form the main part of our lexical database PolyWordNet. Finding a set of good related words is a difficult task. A good related word must possess two essential features. Firstly, it must not lead to create another ambiguity during sense disambiguation process. That is, a related word must be avoided to connect with more than one sense of the same polysemy word as much as we can do. Secondly, a related word(s) must be capable enough to disambiguate the sense.

Organization of Words

Once the related words are generated, the next task is to organize them by linking the related words with respective senses of polysemy words. The lexical database that we developed to organize the senses of polysemy words with their related words is named as ’PolyWordNet’. PolyWordNet organizes multiple senses of a polysemy word in such a way that each sense of the polysemy word is linked with its related words by dividing these related words into verbs, nouns, adverbs and adjectives.

Disambiguation Process using Knowledge based algorithm:

Our algorithm does not count the word overlaps between the context and the sense bags for sense disambiguation. Instead, our algorithm searches the paths or links of context words with the senses of a target word. We keep the track of each path or link that connects a context word and a sense of the target word. If the paths, thus, obtained connect only one sense of the target word, the algorithm output the linked sense as the correct sense of the target word for the given context. If there are paths that link more than one sense, then the algorithm counts the number of paths or links for each linked sense. Then the sense for which the number of connection paths is maximum is selected as a correct sense. If the two or more senses have the equal number of paths, the first sense in the array maintained by algorithm is selected as correct sense. If no connection path is found, the algorithm displays an information indication failure of disambiguation.

V. Possible outcome

The accuracy of the PolysemyWordNet can achieve better accuracy than WordNet. That is, if the senses of a polysemy word and the related words are organized so that they get linked to each other, it resolves the ambiguity in the ambiguity (source of noise information) as it is produced with the use of common relations from the WordNet. In addition, this increases the accuracy of the proposed WSD approaches. In other words, if we interlinked the senses of polysemy word and the contextual related words, it removes the unwanted noise information that causes the wrong disambiguation of sense. The removal of this noise information increases the accuracy of the proposed WSD approaches. We will further carry out evaluation considering supervised approach to further evaluate which model is best suitable for WSD. Presenting an accurate WSD model will aid in solving many online prediction models such sentiment analysis, recommendation system and so on.

REFERENCES

[1] N. Ide and J. Vronis, “Word sense disambiguation: The state of the art,” Computational Linguistics, vol. 24, pp. 1–40, 1998.

[2] E. Agirre and P. Edmonds, Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology). Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[3] S. Banerjee and T. Pedersen, “An adapted lesk algorithm for word sense disambiguation using wordnet,” in

Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, ser. CICLing '02. London, UK, UK: Springer-Verlag, 2002, pp. 136–145. [Online]. Available: http://dl.acm.org/citation.cfm?id=647344.724142

[4] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database*," International Journal of Lexicography, vol. 3, no. 4, pp. 235–244, 1990. [Online]. Available: http://ijl.oxfordjournals.org/content/3/4/235.abstract

[5] G. A. Miller, "Wordnet: A lexical database for english," Commun. ACM, vol. 38, no. 11, pp. 39–41, Nov. 1995. [Online]. Available: http://doi.acm.org/10.1145/219717.219748

[6] G. A. Miller and C. Fellbaum, "Semantic networks of english," Cognition, vol. 41, no. 13, pp. 197 – 229, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0010027791900364

[7] K. Fragos, Y. Maistros, and C. Skourlas, "Word sense disambiguation using wordnet relations," in In Proc. of the 1st Balkan Conference in Informatics, Thessaloniki, 2003.

[8] S. Liu, C. Yu, and W. Meng, "Word sense disambiguation in queries," in Proceedings of the 14th ACM International Conference on Information and Knowledge Management, ser. CIKM '05. New York, NY, USA: ACM, 2005, pp. 525–532. [Online]. Available: http://doi.acm.org/10.1145/1099554.1099696

[9] H.-C. Seo, H. Chung, H.-C. Rim, S. H. Myaeng, and S.-H. Kim, "Unsupervised word sense disambiguation using wordnet relatives," Computer Speech & Language, vol. 18, no. 3, pp.
253 – 273, 2004, word Sense Disambiguation. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230804000166

[10] A. Montoyo, M. Palomar, G. Rigau, and A. Suarez, "Combining knowledge- and corpus-based word-sense-disambiguation methods," CoRR, vol. abs/1109.2130, 2011. [Online]. Available: http://arxiv.org/abs/1109.2130.

[11] U. Dhungana and S. Shakya, "Word sense disambiguation in nepali language," in Digital Information and Communication Technology and it's Applications (DICTAP), 2014 Fourth International Conference on, May 2014, pp. 46–50.

[12] U. Dhungana, S. Shakya, K. Baral, and B. Sharma, "Word sense disambiguation using wsd specific wordnet of polysemy words," in Semantic Computing (ICSC), 2015 IEEE International Conference on, Feb 2015, pp. 148–152.

[13] U. R. Dhungana and S. Shakya, "Hypernymy in wordnet, its role in wsd, and its limitations," in Computational Intelligence, Communication Systems and Networks (CICSyN), 2015 7th International Conference on, June 2015, pp. 15–19.

[14] H. Chen and M. Frank, "Monopoly Pricing When Customers Queue," IIE Trans., vol. 36, no. 6, pp. 569-581, 2004.

[15] Peng Jin, Fuxin Li, Danqing Zhu, Yunfang Wu and Shiwen Yu, "Exploiting external knowledge sources to improve kernel-based Word Sense Disambiguation," Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. International Conference on, Beijing, 2008, pp. 1-8

[16] F. Tacoa, H. Uchida and M. Ishizuka, "A Word Sense Disambiguation Approach for Converting Natural Language Text into a Common Semantic Description," Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on, Pittsburgh, PA, 2010, pp. 478-486

[17] B. İlgen, E. Adali and A. C. Tantuğ, "Building up Lexical Sample Dataset for Turkish Word Sense Disambiguation," Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on, Trabzon, 2012, pp. 1-5.

[18] A. Jain and D. K. Lobiyal, "A new method for updating word senses in Hindi WordNet," Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on, Ghaziabad, 2014, pp. 666-671

[19] G. Chandra and S. K. Dwivedi, "A Literature Survey on Various Approaches of Word Sense Disambiguation," Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium on, New Delhi, 2014, pp. 106-109

[20] R. Pandit and S. K. Naskar, "A memory based approach to word sense disambiguation in Bengali using k-NN method," Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on, Kolkata, 2015, pp. 383-386

[21] N. Patel, B. Patel, R. Parikh and B. Bhatt, "Hierarchical clustering technique for word sense disambiguation using Hindi WordNet," 2015 5th Nirma University International Conference on Engineering (NUiCONE), Ahmedabad, 2015, pp. 1-5.

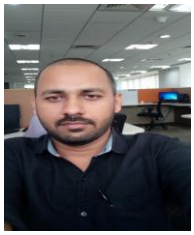
[22] B. F. Zopon AL Bayaty and S. Joshi, "Empirical comparative study to supervised approaches for WSD problem: Survey," Humanitarian Technology Conference (IHTC2015), 2015 IEEE Canada International, Ottawa, ON, 2015, pp. 1-7.

[23] S. M. Tyar and T. Win, "Jaccard coefficient-based word sense disambiguation using hybrid knowledge resources," 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE), Chiang Mai, 2015, pp. 147-151

AUTHOR'S PROFILE

Dr.Nandini N

Department of Computer Science, Dr.AIT Institute of Technology, Bangalore, India



Chandrakanth Rathod

Department of Computer Science, Dr.AIT College, Bangalore, India

Severing Ph.D. in Computer Science from Dr.AIT College Bangalore

SUPPORT PERSONS

Vishwas Yellappa, M.Tech, Department of Computer Science, R.V College of Engineering, Bangalore, India

Big Data Engineer. in Reliance Jio, Bangalore