iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VII **Month of publication:** July 2020

DOI: https://doi.org/10.22214/ijraset.2020.30599

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

Domain Driven Word Sense Disambiguation

Deeksha S[1], Niranjan S[2], Nithin S[3], Bhoomika P[4], Dr. Paramesha K[5]
[1, 2, 3, 4, 5]*Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India*

Abstract: ***Many times people use a single word with multiple senses, which provides a different meaning based on the sentence in which it has been used. So, the main goal of the work is to disambiguate an ambiguous word that has been located in certain sentences. Thus, it helps in exploring the actual sense of the word that it means. This process is itself named as Word Sense Disambiguation (WSD) which is an essential and on-going subject in NLP. However, the role of domain is very helpful in exploring the actual sense of an ambiguous word. There are a number of approaches to WSD which takes a wider semantic space of ambiguous words into account [10]. This semantic space can be represented as a specific domain, task or application [10]. The domain information which is one among them is more advantageous in the process of disambiguation, thus, the work explores the role of the domain in the disambiguation process. The work also includes a score allotment to all the senses of an ambiguous word based on the semantic relation of it with the ambiguous word. Later, this helps in obtaining the actual sense of the word.***

Keywords: Disambiguation, natural language processing, word sense

I. INTRODUCTION

In NLP, ambiguity has become a barrier to human language understanding. The best solution to overcome this barrier is the Word Sense Disambiguation process. For instance, consider an example word ‘bank’ which has two senses. One sense of the word is “financial reservoir” and another sense of it is “a river edge”. Now consider the following two statements.

“Willows lined the bank of the stream” and “I went to the bank to get a home loan”

To a human, the difference between these two senses of the word ‘bank’ will be clearly understandable as it means ‘river bank’ in the first sentence and a ‘financial reservoir’ in the second sentence. But this is not the same in case of a machine. Thus, it needs a different solution. There are different solutions of WSD that have been proposed. This can be generally divided into supervised and knowledge-based approaches. In comparison of these two approaches, the knowledge-based approach has gained a rapid development while compared to others in recent years [12]. Also, the availability of abundant information from a different knowledge resources has narrowed the gap between these two approaches [12]. Hence, the use of a knowledge-based approach has been considered as a useful one in our work. The main idea of this approach is to make use of WordNet and a semantic space such as specific domains to a greater extent to obtain the actual sense of the ambiguous word. Thus, the main purpose of our work is to make use of the domain information as a best possible way to explore the actual meaning of an ambiguous word that has been used in a sentence. There is a hypothesis that the domain information provides a powerful way to establish a semantic relation among the word senses [11]. Thus, it can be used in a profitable way in the process of disambiguation. We can refer to the domain, as a set of words that has a strong semantic relation between them. Thus, an approach of using domain information in obtaining the actual sense of an ambiguous word makes sense. Here, the basic prediction that we assumed to achieve the goal was to make use of a distinct score allotment for each of the senses that comes under the semantic space of domain of ambiguous word. This score allotment is not a random one, rather the sense that is closer to the word in a particular sentence will be allotted with a highest score while the remaining senses will be allotted in a decreasing manner accordingly. This prediction gave a way to attain the result later. Also, this approach helps in less time consumption than the normal disambiguation process. Therefore, the efficiency of the project will also be increased.

II. LITERATURE SURVEY

S.G Kolte and S.G Bhirud approaches is to Word sense disambiguation using wordnet domains, here they used wordnet as database to define domains and the words in the given sentence is taken as parameter which helps to detect domain by domain-oriented text analysis. For this they go through unsupervised method, and they trying to disambiguate nouns first using pos tag and getting results but here the drawback is that this approach is failed for a word having more than one sense in a same domain [1].

A Fully Unsupervised Word sense disambiguation system using dependency knowledge on specific domain is proposed in the year 2010, here they developed a fully unsupervised system using domain specific knowledge and this system performs above the first sense baseline. they showed that wsd can be achieve in unsupervised method without get compromised with supervised approach by using easily available unannotated text from internet and other sources and get a good result [2].

In the year 2009, Eneko Agirre, Oier Lopez de Lacalle and Aitor Soroa proposed a knowledge-based approach on specific domain which is better than generic supervised approach of WSD. Here they used information that are available in wordnet, and domain specific corpora and evaluate with publicly available domain specific datasets. They used totally three corpora, one is balanced corpora and two domain specific corpora. And this improves the overall previous results on this approach and also over the supervised system trained on semcor. Here they also used related words as context instead of occurring actual context words and obtained better results [3].

In the year 2010, Yuhang Guo, Wanxiang Che, Ting Liu and Sheng Li, proposed a semi supervised domain adaptation approach for WSD using word by word model selection. authors explored supervised and semi- supervised for domain adaptation. and then they trying to go with different models among which the target word is automatically selected from candidate model set which includes self-training and supervised models. This obtain higher accuracy than each single model and improves the performance than individual compared to supervised model. The result of this experiment is higher than SVM for some targeted words [4].

In the year 2011, Wei Jan Lee and Edwin Mit proposed WSD using domain knowledge, here they used wordnet. their approach is to take definition of each word and related domain using wordnet domains and gave some weight to each and based on weight assigned to each word's domain the sense of ambiguous word will be detected. By this they achieve 70% for all the context but there few drawbacks. Domain Label Factotum is an issue that has to be solved [5].

Previously when coming to supervised method they consider pos tags, word, and position of certain words and other features. But sometimes a word which contain more than one sense has characteristics context by which different senses are appear. Here they combined knowledge-based method with supervised method and develop a feature extraction algorithm, which extract different characteristics features of a word. then this feature integrates with entropy classifier to disambiguate for some specific words. This improves significant amount of accuracy for such words [6].

An unsupervised method to disambiguate words in a sentences is proposed by author, here trained words are considered as first language and its translated to other language and that considered as second language then they compared each word with translated polysemous word and calculating word similarity scores and based on that they used to disambiguate the sentences. here they used raw corpus and a bilingual dictionary to disambiguate [7].

Many supervised word sense disambiguation system have been built but it's difficult to create corpora therefore researches are made in unsupervised but in this paper author proposes to stick to supervised approach but with less annotation or corpora. Their approach is tested on limited domains thus they get good results on limited words and domains. Here authors used method is convenient middle ground between pure supervised and pure unsupervised WSD[8].

Tybots are used to extract domain related terms which produce groups of synonyms and related concepts related to domains. This is attained by using knowledge-based approach. This system is a kyoto system which is for specific domains. With lexical knowledge, this system can be applied to any language. By this they attained good result in few languages [9].

III. METHODOLOGY

The system of methods that are included under the study of particular area are considered to be the methodology of a system. Here we have the architecture and the control flow diagram that brief about the methods that have been followed to disambiguate a word to obtain its proper sense or the actual meaning of the word. When architecture of the system is considered, it is one of the main designing techniques that must be followed to compile the goal or target. Firstly, the original sentence that has an ambiguous word along with the target word whose actual sense must be find out will be provided as an input to the system. Then tokenization of the sentence will be performed where the words of the sentence will be split to individual words i.e., W1, W2, …....., Wn. The domain of the words must be detected by using dictionary. Here, the domain of the words is included in the dictionary. Then the comparison between all the words and their applicable senses i.e., S1, S2 … Sn, using dictionary can be performed. The domain information is included in the WordNet. Synsets have at least one of the domain labels. Domains can include synsets of different syntactic categories, and they may group multiple different senses of a certain word into a semantic cluster [13]. This decreases the ambiguity level in the case of specific domain disambiguation [13]. Then, a respective score will be allotted between all the pairs of words and senses. Then the one with the best score is considered to be as the actual sense or meaning of the word.

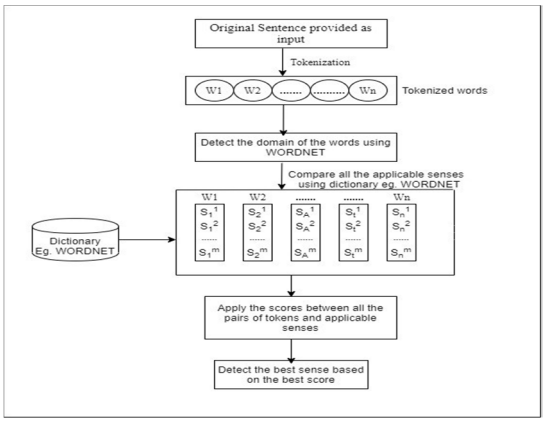


Fig (1): System Architecture

Control diagram or flow diagram gives a brief idea of steps that must be followed to obtain the desired output. Here, right from giving input to the system to obtaining output from the system a detailed idea can be expressed in the form of a diagram. The following is the Control Flow diagram of the proposed system. The sentence in which the ambiguous word is located is given as input to the system along with the target word that has to be disambiguated. Then the process of tokenization will be performed where the words in the sentence provided will be converted into tokens.

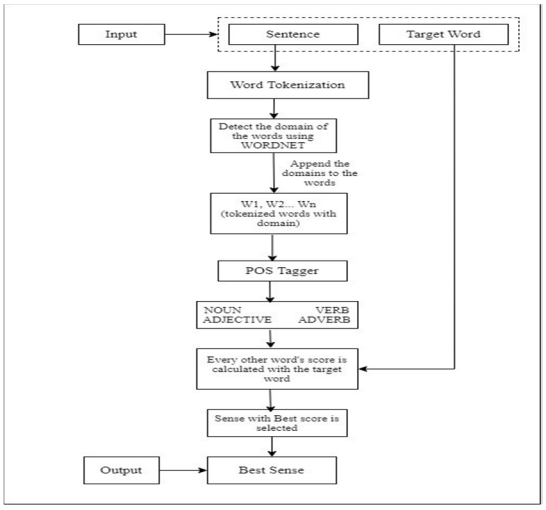


Fig (2): Flow chart

Then the domain of the words can be detected using the dictionary, here, WordNet. This is as because the domain information is included in the WordNet or dictionary. Then, by using POS tagger the main words that has more contribution in detecting the actual sense of ambiguous word can be obtained. For example, Noun, Verb, Adjective and Adverb has more contribution in finding out the actual meaning of the ambiguous word rather than the Conjunction. So, this is also one of the main steps to filter all the unnecessary words that have no use in finding out or disambiguating the word considered. Then, every other word's score is calculated with the target word. This helps in obtaining the respective scores for all the possible pairs of the word and senses. Then the sense with best score is selected as the actual sense of the word. Thus, the output can be obtained and the target word can be disambiguated.

IV.RESULT ANALYSIS

When an appropriate goal or aim is set to come up with a solution for a problem statement, then there comes the necessity to analyze the approaches that leads us to the solutions and to obtain the desired solution that will be the result of our work or procedure that has been followed. So, in our work, initially we came up with a lot of approaches that might be followed to disambiguate an ambiguous word. But, the efficiency of the result will be as important as the proper solution that should be obtained. In that regard, knowledge-based approach has been chosen as the best approach for the process. The publicly available lexicon, WordNet, has been used in our knowledge-based approach to disambiguate a word. The working of the algorithm is as follows, when we need to

disambiguate a word, here it is referred to as a target word. Firstly, the provided input sentence will undergo word tokenization. Then the domain of the words will be detected using the WordNet. These domains of the words that have been detected will be appended to the tokenized words which is followed by the parts of speech tagging process. This helps us to consider only a noun, verb, adjectives and adverb, whereas the rest of them may be ignored. Then a word family that is available for the target word will be considered. This word family refers to a set of related words which belongs to a synset. Similarly, a word family for a nearby word that has a huge impact on the target word will also be considered. The word family belonging to the target word will be kept separately from the nearby word's word family which will be clustered. Then the intersection of these two families will be performed where the more identical or common words of each sense of target will be intersected with the other word family. Then by considering the overall interaction and distance on the degree of similarity between the target and nearby words, a score will be calculated for each of the senses of a provided target word. Then the sense with the best score will be considered to be the best sense of the target word.

For example, consider an ambiguous word 'left'. As on human knowledge, it has two different senses, that is past tense of word leave or direction opposite to the right. Then if someone says 'I forgot and left my wallet home', then the target word will be 'left'. Here, the word's different senses will be considered along with the nearby supporting word. Then a score will be allotted to the intersection of the senses of the word 'left' with its nearby word that may be 'wallet' in this case. Based on this score will be allotted and the sense with the best score will be considered as a best sense. When this example is given as an input sentence and target word in our model, the sense that came up as an output was 'Leave behind unintentionally'. Similarly, for other set of example sentences. Thus, the result obtained for our work with an accuracy of 0.875.

V.CONCLUSION

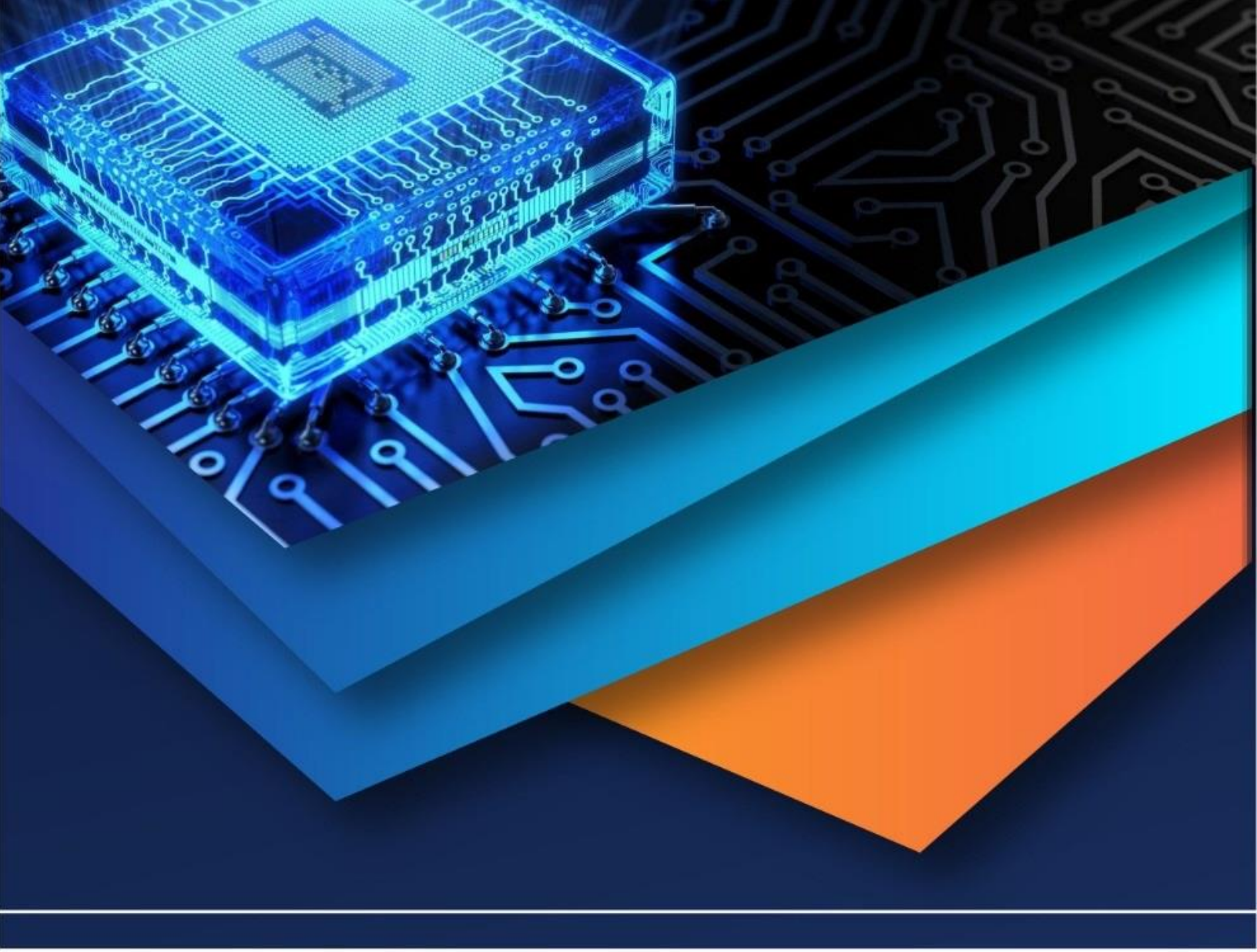
The role of domain information in disambiguation process is more advantageous. There is an underlying assumption of that the domains establishes a semantic relation among the words or its senses, which is more useful in the disambiguation process. So, the use of domain information profitably helps in the process of word sense disambiguation. This includes an adjustment of WSD algorithm from the general to a specific domain and this mainly starts from detecting the domain in the dataset and in our work, dictionary plays a vital role in finding out the domain. In our work, we have proposed the Michael Lesk work and methods to get the correct sense of the target word that has been provided along with the input sentence. Here, the algorithm calculates a value for all the possible senses of the target word. This will be done by considering the maximum similarities of each context sense with the target word. Then after allotting a specific value for each of the senses, the sense with the best score will be considered as an actual sense of ambiguate word. Thus, the disambiguation process can be performed.

VI.ACKNOWLEDGMENT

The authors would like to express their gratitude to their mentors and faculty members who have provided their constant support for them throughout their research and also helped them in achieving the desired results in a limited period of time**.**

REFERENCES

[1] S. G. Kolte , S. G. Bhirud, "Word Sense Disambiguation using WordNet Domains". 2008 IEEE.
[2] Andrew Tran, Chris Bowes , David Brown , Ping Chen , Max Choly , Wei Ding , "TreeMatch: A Fully Unsupervised WSD System Using Dependency Knowledge on a specific Domain". 2010 ACL.
[3] Eneko Agirre and Oier Lopez de Lacalle and Aitor Soroa , "Knowledge-Based WSD on Specific Domains:Performing Better than Generic Supervised WSD" .
[4] Yuhang Guo, Wanxiang Che, Ting Liu and Sheng Li , "Semi-supervised Domain Adaptation for WSD:using a Word-by-Word Model Selection Approach". 2010 IEEE.
[5] Wei Jan Lee and Edwin Mit , "Word Sense Disambiguation By Using Domain Knowledge" 2011 IEEE.
[6] Liang Wen1, Juan Li1, Yaohong Jin1, Yongjie Lu , "A Method for Word Sense Disambiguation Combining Contextual Semantic Features".
[7] Behzad Moradi,Ebrahim Ansari,Zdenek Zabokrtsky, "Unsupervised Word Sense Disambiguation Using Word Embeddings".
[8] Mitesh M. Khapra, Anup Kulkarni, Saurabh Sohoney, Pushpak Bhattacharyya, "*All Words Domain Adapted WSD:* Finding a Middle Ground between Supervision and Unsupervision".
[9] Aitor Soroa, Eneko Agirre, Oier Lopez de Lacalle, Monica Monachini,Jessie Lo, Shu-Kai Hsieh, Wauter Bosma, Piek Vossen , "Kyoto: An Integrated System for Specific Domain WSD".
[10] Domain-Specific WSD by Paul Buitelaar, Bernardo Magnini, Carlo Strapparava, Piek Vossen.
[11] https://www.researchgate.net/publication/228760009_The_role_of_domain_information_in_Word_Sense_Disambiguation
[12] Word Sense Disambiguation: A comprehensive knowledge exploitation framework by Yinglin Wang, Ming Wang, Hamido Fujita.
[13] https://www.researchgate.net/publication/303412026_Solving_Specific_Domain_Word_Sense_Disambiguation_using_the_D-Bees_Algorithm



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY