IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Approaches of Resolving the Ambiguities of Word in Sentences

Roshan Karwa

Assistant Professor, CSE Department, PRMIT&R, Badnera, India

Abstract: There is multiple meaning of single word, for example, the word "Cold". One meaning of Cold is Weather and other meaning is viral infectious disease. Identification of correct meaning of ambiguous word with respect to particular context is nothing but Word Sense Disambiguation (WSD) which is required in every field of Natural language processing like in Machine Translation for lexical choice for words that have dissimilar versions for different senses. In Information Retrieval, WSD is for Resolve ambiguity in questions and in Information Extraction for discriminate among precise occurrences of concepts. WSD is one of the demanding problems in Natural Language Processing (NLP). NLP is ability of computer program being able to processes human like language like Hindi, English, and French etc. This document presents an analysis on methods for WSD and proposed one method which is based on Existing method.

Keywords: Natural language processing, Word Sense Disambiguation, Dictionary, Corpus.

I. INTRODUCTION

Even if words with several senses give English a linguistic affluence, but they can as well generate ambiguity. For example, putting money in the bank could mean depositing it in a financial institution or putting in the ground by the riverside. To know accurate meaning of particular word in a given context is largely unconscious and usual in human but it's quite tough for computers as it lack real world knowledge necessary between word meanings i.e. Computer program has no basis for knowing which one meaning is appropriate. Hence, determining correct meaning for words in context is important and called as Word Sense Disambiguation [1, 9, 11]. Important step in Word Sense Disambiguation are as follows: given a set of word, a classifier is applied which makes use of one or more sources of Knowledge to find out the most appropriate senses with words in context. Sources from, which knowledge about word will get is of two types, one is corpus based which is either unlabelled i.e. unannotated or annotated with word senses, and other is dictionaries related machine readable dictionaries, dictionaries, thesauruses etc. Without knowledge sources, it is difficult for both humans and machines to identify the correct sense i.e. meaning. A number of WSD techniques have been proposed in the past such as knowledge based, supervised or unsupervised methods. Supervised and unsupervised is depend on corpus. Knowledge based WSD is rely on knowledge resources like Machine Readable dictionaries, dictionaries, thesauruses [8].

This document is structured as follows: Section 2 discusses WSD task and its basic elements, section 3 deals with dissimilar methods of WSD task and next section is Conclusion is in section 4.

II. WORD SENSE DISAMBIGUATION: TASK

WSD can be summarized as a classification task: word senses are the classes, and a self regulating classification method is used to assign each occurrence of a word to one or more classes based on the evidence from the context and from external knowledge sources [1].

A word sense is a correct meaning of a word. Consider the following two sentences, One is "I like cold Weather" and other is "I have suffered from cold since two days." The word COLD is used in two senses. One is of type of Weather related and other is related to Viral Disease. Selection of appropriate word sense is one of the elements. As without knowledge, it is impossible both for human being and computer to identify correct meaning so for that Knowledge sources are created by researchers which provide data which is essential to associate senses with words. This source is of two types, one is corpus which is either unlabelled or annotated with word senses, and other is dictionaries like machine readable dictionaries [1].

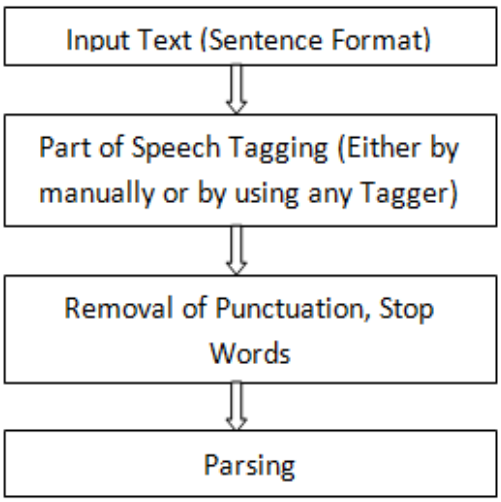

Fig1: Pre-Processing of Input Text

As text is in unstructured form, Preprocessing of the input text is usually performed, which includes the following steps: Sentence is entered, then tokenization is done. Afterwards Part of Speech tagging either manually or by

using any Tagger. Doing POS speech tagging manually is very time consuming. Also require manpower. So best way to do it is by using any available tagger but also that should be efficient. Then by removing Stop words like a, an, the etc and also by removing punctuation, Getting content words from sentence is objective. At the end, unstructured sentence summarizes as structured content words.

III. SELECTION OF METHOD FOR EXTRACTING THE ACCURATE SENSE

A. *KNOWLEDGE BASED LEARNING METHODS*

Knowledge based methods, often refer as dictionary based methods uses lexical knowledge bases such as dictionaries like WordNet [6], thesauri, ontologies etc and acquire information related to word from word definition and relations present the respective knowledge base [1,11]. In this section, we review the various knowledge based approaches proposed by researchers.

Agirre, Eneko & German Rigau (1996) [2] proposed Word Sense Disambiguation with Conceptual Density method which uses lexical knowledge base. This method opts for a sense based on the how close the concept characterize by the word and the concept characterized by its neighboring words. This is Conceptual distance. First find the noun in context then its senses and relations majorly the hypernym. So for finding the hypernym, here researcher proposed to use the dictionary.

Agirre, Eneko & David Martínez (2001) suggested Selectional preferences [3]. Basic idea they framed is that look for argument frame of verb, wherever the partocular property dictates particular sense, pick up that sense. SERVE_EDIBLE, SERVE_SECTOR are some examples of such semantic constraints, here if context is about serving food, it will take SERVE_EDIBLE sense and if the context is about serving a regions then it will take sense of SERVE_SECTOR i.e. based on the preference property.

Lesk (1986) [1], Satanjeev Banerjee and Ted Pedersen (2002) [14] suggested Overlap based approaches which purely matching based approaches i.e. finding the match between ambiguous word and feature word i.e. neighbor context words.

The above approaches do not need enormous training but the problem with knowledge based approach is that the lexical knowledge base such as dictionary, thesaurus is restrained for sense of target word. Only the lexical information is there which is insufficient for acquiring the accurate sense.

B. *SUPERVISED LEARNING METHODS*

Machine learning **S**upervised WSD [1,13] is the method which depend on the external knowledge source i.e. corpus evidence which is tagged one. Machine learning requires a training of corpus and testing of unknown samples. Training module requires a sense training corpus which is annotated one from which syntactic and semantic features are picked up using machine learning techniques such as Naïve baye‟s probabilistic learning, Support vector machine learning, and Decision list log likelihood learning etc. In testing, based on training date, it extracts the winner i.e. best sense for a word on the basis of its surrounding words [15].

A Naive Baye‟s supervised approach suggested by Gerard Escudero et al. (2000) [7] is a simple probabilistic approach based on the application of mathematical Baye‟s theorem. Basic idea is to consider the feature vector[10] consisting of POS of an ambiguous word; Collocation feature i.e. neighboring words of fixed window size say +2,-2 and co-occurrence feature [15,16] then calculate the prior probability and final score which depends on the conditional probabilities of each feature which is independent. Winner sense will be one which will have higher probability.

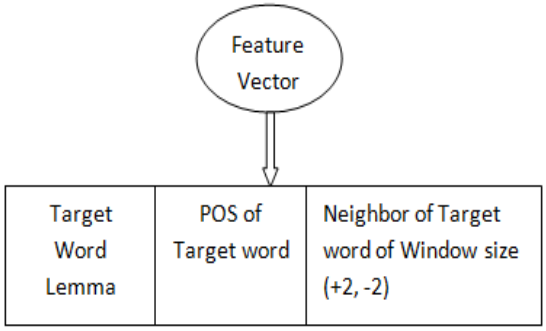

Fig 2: Naïve Baye's Feature Vector

Agirre, E. and Martinez, d. (2000) [4] proposed the decision list supervised method which is also mathematical approach, and based on the log likelihood ratio. Higher the ratio, that will be the best sense. In this, Feature vector to be considered is relied on „One sense per collocation‟ property as nearby words providing strong and uniform hint as to the sense of a target word.

Gerard Escudero et al. (2000) [7] proposed Exemplar-based supervised in which the classification model is built from examples. The model preserves examples in memory as points in the feature space and, as new examples are subjected to classification, they are gradually added to the model.

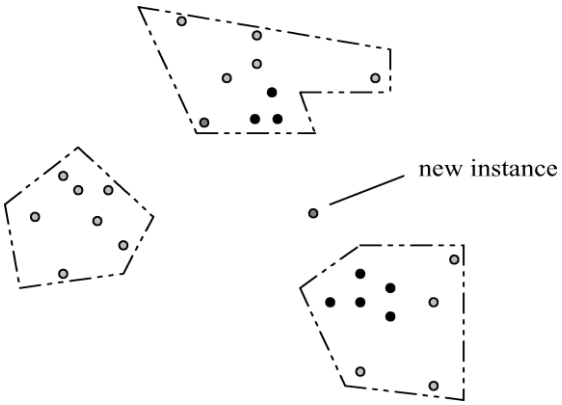

Fig 3: Exemplar Based Learning

Support Vector Machines (SVM) is the method introduced by Boser et al. (1992). Basic idea is in training phase, SVM is trained using POS, collocation, co-occurrence and

syntactic relation and in testing phase, give a test sentence, a test example is constructed using above features and fed as input to each binary classifier [1].

Performance of supervised approach is comparatively better than the rest approaches but effort of creating the training corpus- annotated sense marked corpus is a major issue faced by natural language processing community. Also the knowledge acquisition problem is serious issue of these approaches.

C. *SEMI SUPERVISED LEARNING APPROACHES*

This approach is proposed designed for the alternative to Knowledge based and Supervised. Motivation behind this approach is Annotated data is expensive and difficult to create where as unlabelled data is cheaper but annotation is needed somewhat so this approach uses minimal annotated data[1]. Basic idea is to have seed training data, then train a system using seed data, after that tag unseen data, henceforth manually correct tags then retrain using the larger data and repeat it until satisfactory accuracy level is reached.

D. UNSUPERVISED LEARNING

As supervised face the problem of knowledge acquisition, Unsupervised methods have the potential to overcome this problem by acquiring sense unannotated i.e. untagged corpus which is based on the idea that the same sense of a word will have similar neighbouring words, they are able to induce word senses from input text by clustering word occurrences, and then classifying new occurrences into the induced clusters.

Researchers R. Navigli and M. Lapata (2010) [12] have proposed several Graph based methods in which they builds a graph with senses as nodes, and relations among words and senses as edges, with the relations usually acquired from an LKB such as wordnet. Then, the researcher conducts a ranking algorithm over the graph, and assigns senses that are ranked the highest to the corresponding words.

Researchers using these methods have experimented with different relations and ranking algorithms, such as the E. Agirre and A. Soroa (2009) Personalized pagerank algorithm[5]. These approaches are based on the notion of a *cooccurrence graph*, that is, a graph $G = (V, E)$ whose vertices V correspond to words in a text and edges E connect pairs of words which co occur in a syntactic relation, in the same paragraph, or in a larger context.

Hyperlex, first, a cooccurrence graph is built such that nodes are words occurring in the paragraphs of a text corpus in which a target word occurs, and an edge between a pair of words is added to the graph if they cooccur in the same paragraph. Each edge is assigned a weight according to the relative cooccurrence frequency of the two words connected by the edge.

Researchers R. Navigli and M. Lapata (2010) [12] proposed Similarity-based algorithms which assign a sense to an ambiguous word by comparing each of its senses with those of the words in the surrounding context. The sense whose definition has the highest similarity is assumed to be the correct one.

As unsupervised approach mainly deals with the clustering, number of cluster may differ from the number of senses of target word to be disambiguated which is a major issue. Also unsupervised approach is that the instances in training data may not be assigned the correct sense, clusters are heterogeneous [16].

IV. CONCLUSION

Based on study of WSD scenarios, I make the following conclusions:

1. Considering the disadvantages of all existing approaches i.e. knowledge based requires exhaustive enumeration search and knowledge resources, supervised has a problem of data sparseness, also huge number of parameters require to be trained and the unsupervised algorithm fails to distinguish between finer sense of a ambiguous word so effort should made to resolve the issue by suggesting the hybrid approach.

2. Integration of various knowledge resources for a feature set such as Part of speech, morphological form(Lemma) of word, Neighboring words(in form of collocation vector), verb noun syntactic relation are helping us to obtain a good accuracy for classification [15].

3. System will work with high accuracy when the inappropriate information is detached from the sentences and also when the training data is increased.

REFERENCES

[1] Navigli, roberto, "word sense disambiguation: a survey", ACM computing surveys, 41(2), ACM press, pp. 1-69, 2009.
[2] Agirre, Eneko & German Rigau. 1996. "Word sense disambiguation using conceptual density", in Proceedings of the 16th International Conference on Computational Linguistics (COLING), Copenhagen, Denmark, 1996.
[3] Agirre, Eneko & David Martínez. 2001. "Learning class-to-class selectional preferences" in Proceedings of the Conference on Natural language Learning, Toulouse, France, 15–22.
[4] Agirre, E. and Martinez, d. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In Proceedings of the 18th International Conference on Computational Linguistics (COLING, Saarbr ¨ucken, Germany). 11–19.
[5] E. Agirre and A. Soroa, "Personalizing PageRank for Word Sense Disambiguation," Proc. 12th Conf. EuropeanChapter of the Assoc. for Computational Linguistics (EACL 09), Assoc. for Computational Linguistics, 2009, pp. 33–4.
[6] Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts, 1998.
[7] Gerard Escudero, Llu´ıs M`arquez and German Rigau, "Naïve Bayes and Exemplar-based approaches to WordSense Disambiguation Revisited", arXiv:CS/0007011v1, 2000.
[8] Mitesh M. Khapra, Anup Kulkarni, Saurabh Sohoney, and Pushpak Bhattacharyya. 2010. All words domain adapted WSD: Finding middle ground between Supervision and unsupervision. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, ACL, pages 1532-1541.
[9] Mitesh Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2010. Domain-specific word sense disambiguation combining corpus based and wordnet based parameters. In *5th International Conference on Global Wordnet (GWC2010)*.
[10] Mihalcea and D.I. Moldovan. Pattern Learning and Automatic Feature Selection for Word Sense Disambiguation. In Proceedings of the Second international Workshop on Evaluating Word Sense Disambiguation Systems(SENSEVAL-2), 2001.

[11] Ping Chen and Chris Bowes, University of Houston-Downtown and Wei Ding and Max Choly, University of Massachusetts, Boston Word Sense Disambiguation with Automatically Acquired Knowledge, 2012 IEEE INTELLIGENT SYSTEMS published by the IEEE Computer Society.

[12] R. Navigli and M. Lapata, "An Experimental Study of Graph Connectivity for Unsupervised Word SenseDisambiguation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32,no. 4, 2010, pp. 678–692.

[13] Roshan R. Karwa , M.B.Chandak "Word Sense Disambiguation: Hybrid Approach with Annotation Up To Certain Level – A Review", International Journal of Engineering Trends and Technology (IJETT), V18(7),328-330 Dec 2014. ISSN:2231-5381. www.ijettjournal.org. published by seventh sense research group.

[14] Satanjeev Banerjee, Ted Pedersen, "An adaptive Lesk Algorithm for Word Sense Disambiguation Using WordNet", Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, page no: 136-145, 2002.

[15] Roshan Karwa & Manoj Chandak "Hybrid approach to word-sense disambiguation with and without learned knowledge" Published in: International journal of Natural language computing Vol. 4, No.2, April 2015.

[16] Roshan R Karwa, Dr. M.B.Chandak, Deepak Pande, "A Knowledge based Approach to Resolve Word Level Ambiguity for Machine Translation", In International Journal of Computer Systems, Volume 2, Issue 5, May, 2015, pages: 206-211.

BIOGRAPHY

Roshan Karwa received the M.Tech degree in Computer science & Engineering from Shri Ramdeobaba College of Engineering & Management, Nagpur. He is currenlty an assistant professor in Department of Computer Science at Prof Ram Meghe Institute of Technology & Research, Badnera. His research focuses on natural language processing (specifically, word sense disambiguation).