Word Sense Disambiguation : Methods and Algorithms

Rakesh kumar

Abstract : This paper discuss various technique of word sense disambiguation. In WSD we disambiguate the correct sense of target word present in the text. WSD is a challenging field in the natural language processing, it helps in information retrieval, information extraction, machine learning. There are two approaches for WSD machine learning approach and knowledge based approach. In Knowledge based approach a external resource is used to help in disambiguation process, but in Machine learning approach a corpus is used whether it is annotated, un-annotated or both.

Keywords: Natural Language processing, machine learning, knowledge based learning, word sense disambiguation.

I. INTRODUCTION

Almost in all human languages there are words, which have different own separate sense in different context. These types of words are called polysemous word, which are spelled same but keep separate sense in different context. To, get the proper sense of a ambiguous word and allocate the actual meaning to the word is called WSD. This is also a important approach in natural language processing(NLP).

Word sense disambiguation is important machine Tranlation(ML), Semantic mapping(SM), ontology Learning(OL), Semantic Annotation(SA), it also helps in Information Retrieval(IR), Information Extraction(IE).

We the human being are with intelligence, we can easily get the proper sense of the polysemous word in particular context. We can do it with the help of surrounding words i.e. the local or global context. Local context means few word nearer to the target word. Global context, lines or paragraph nearer to the ambiguous word. But if we want to do with the system we have to give some rules to the system and train the system with the help of rules. Then after the system can disambiguate the proper sense of the ambiguous word in the particular context.

Sense of a word means that the meaning accepted by the language community for the word. There are words which are spelled same own separate sense in different surrounding. To find the correct sense of a ambiguous we tends to look up the dictionary. So, for research community we needed require resources as evolution material. For English WORDNET.

The sentence is spoken by speaker so, we are only concern with the sentence not the emotion. We as a human being can easily get the emotion attached with the sentence. With the machine learning we can only process the sentence and if any polysemous word occurs then the surrounding of the ambiguous context of the target word is analyzed. The surrounding of the ambiguous word depend upon the dimension of context.

Zero context—the word itself.
Local context—phrase and clause.
Global context—sentence, discourse, topic and domain.
Mainly there are two types of word sense disambiguation approaches:-
1) Machine Learning Approach.
2) Dictionary Based Approach.

II. DICTONARY BASE APPROACH

In this approaches an external resource is used to help the disambiguation process. External resource can be machine readable dictionary, thesauri, etc. Wordnet is the mostly used MRD in this research area.
A MRD provides the following things:
1) All meaning of given word.

2) A description of all word meaning.

3) One or more example sentences for most of the meaning.

A. Lesk Algorithm:

This algorithm is the first algorithm for word sense disambiguation problem with knowledge based approach. In this the surrounding of the ambiguous word is taken to consideration and the sense for whole context is retrieved from the MRD. Then the words in the surroundings in neighbourhood are used to measure the similarity with the dictionary senses of the ambiguous word.
Algorithm:
1) Retrieve all senses of target word from the MRD.
2) Determine the similarity among all the sense combination.
3) Select senses that tends to maximum overlap.

B. Selection Preferences

The surrounding of the ambiguous word is limited for getting the proper sense of ambiguous word. It finds knowledge of the word type relation and propse common meaning using dictionary. In this the improper relation between words are omitted or not considered only those are considered own proper relation with sense rules.

Semantic Similarity:

It is the human behaviour that we use the word which effectively communicate our dialogue. This behaviour of words is exploited by WSD. The discourse is logical and consistent if the word in the discourse are related. Means that the words in the discourse are related by meaning. The proper sense of the words are identified by this approach with finding the semantic difference among the meaning of the ambiguous word, which have less distance could be the

Dr. Rakesh Kumar. Assistant professor at Department of computer science in Asssam University, silchar, Assam.

appropriate sense of ambiguous word within the particular context. Context can be local or global. Local means few words in the surrounding. Global means more than that of a sentence.

III. MACHINE LEARNING APPROACH:

In this approach the system is trained to find out the proper sense .A classifier is applied to learn the features and assign senses to unseen examples. In this technique the ambiguous word and the text surroundings of the ambiguous word is given as the input to the system.

A. Supervised Word Sense Disambiguation

The supervised approach applied to WSD system use machine learning technique from manually created sense annotated data. Classifier is trained by the data set which contains also example related to the target word. The classifier take the target word as input and look up for the manually sense data set for appropriate sense . the training data set is used to train the classifier which also contains the target word.

Methods in supervised WSD are as follows:

Decision List,

The decision list is a set of "if –then-else" rules. To get the features of word, the training set is used in definition list. With the help of these rules. Based on the score the final list is created. The maximum score for a vector represents the sense.

Decision Tree

A decision tree divides the training data in the recursive manner and represents the rules for classification in a tree structure. The internal node of the tree represents test on the features , branch represents how decision is made and leaf node represents the result of the prediction or sense of the word.

Neural Network

In the neural network the neuron is the basic processing unit, the input to the neuron is given and the output is generated after processing with in the neuron. Neural network is a many layered network where each layer contain one or more than one neuron. There are input layer, hidden and output layer, there can be one or more than one hidden layer be present in the neural network. The input layer collect the input from outside world then process the input and propagate towards out layer. If the output of output layer dosen't match with the desired output the error is calculated and then the error is back propagated and again the parameters are changed. We change the value of parameters according to the error, then output is calculate. This process is repeated until the error is minimized.

Examplar Based Learning:

This learning technique build classification model from examples. The model stores the points in feature space whenever a example comes, this technique calculate the distance between new example and the information in feature space. The more closer to already stored particular class, chances to that class is high.

B. Unsupervised Machine Learning Approach:

These methods does not depend on the external resource to disambiguate word. In this learning algorithm they discriminate the word meaning based on the information in un-annotated corpora.

Main approaches of unsupervised learning are as follows:

Conext Clustering

In this method the context vector are created and they are grouped in the cluster to identify the meaning of word. The vector space is used as word space and the dimension are the word itself. The word which is going to disambiguate work as a vector and how many times it occurs in the context is counted.

Word Clustering

This approach was based on the intuition that different words, occurring in similar local context are likely to have similar meaning.

Lin Algorithm:

1) The input sentence having ambiguous word w is passed to take out the local context of target word.
2) Search is made through local context database to find the words S which has identical context as the w.
3) Choose the sense that maximises the similarity between S and w and assign the selected sense to all occurrence of the target word.

Co-Occurrence Graph

With this method we can create the co-occurrence graph with vertices V and edge E. V is the set of words in text and E is added if the words co-occur with some syntactic relation within paragraph or text. First of all graph is created then Adjacency matrix is created from the graph. After that Markov clustering technique is used to find the meaning of ambiguous word.

C. Semi-Supervised Machine Learning Approach

This approach use both technique of supervised and un supervised machine learning concepts.

In this the information may be given as in the supervised learning but that has less information, there only critic information is available not the exact information. It starts with small amount of sense labelled data a large amount of unlabelled data and one or more classifiers.

IV. RESULTS

Algorithm Category	Strengths	Weaknesses	Predominant Algorithm(s)
AI Methods	Some ideas formed the basis of all further work on the subject e.g. word window	Very domain specific	Expert Systems, as described by Small (1981) Semantic Networks, as described by Dahlgren (1988)
Knowledge Based	Accuracy	Rely on precompiled lexical knowledge resources	The Lesk algorithm, as described in Lesk (1986) Yarowsky's algorithm, as described in Yarowsky (1992)
Supervised	Accuracy	Dependent on pre-annotated corpora for training data	Naïve Bayesian Classifier, as described by Gale et al. (1993)
Unsupervised	No pre-training necessary Works on multiple languages with no modification to the algorithm	Merely discriminates between word senses; not disambiguate word senses	Yarowsky's algorithm, as described in Yarowsky (1995)

Summary of Word Sense Disambiguation Approaches

V. CONCLUSION

In this paper, we have done proper survey of different approaches for word sense disambiguation. Mainly in machine learning method a large corpora is used to help the process of learning and disambiguation in this method a training is required for the system which is used for disambiguation process. These machine learning methods perform better but the main disadvantage is they require large corpora.

REFRENCES

1. Kilgarriff, A., Evans, M. (eds): Special issue on SENSEVAL. Computer and the Humanities, 34(1-2) (2000)
2. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Five Papers on WordNet. Special Issue of International Journal of Lexicography, 3(4), (1990)
3. Miller, G. A., Leacock, C., Tengi, R., Bunker, R.T.: A Semantic Concordance. Proceedings of the ARPA Workshop on Human Language Technology (1993)
4. Hirst, G.: Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press. Cambridge, England (1987)
5. McRoy, S.: Using Multiple Knowledge Sources for Word Sense Discrimination. Computational Linguistics, 18(1) (1992)
6. Ide N., Veronis J.: Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. Computational Linguistics, 24(1) (1998)
7. Bruce, R., Wilks, Y., Guthrie, L., Slator, B., Dunning, T.: NounSense - A Disambiguated Noun Taxonomy with a Sense of Humour. Research Report MCCS-92-246. Computing Research Laboratory, New Mexico State University (1992)
8. Rigau, G., Atserias, J., Agirre, E.: Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. Proceedings of ACL-EACL, Madrid, Spain. (1997)

AUTHOR PROFILE

Dr. Rakesh Kumar. Assistant professor at department of computer science in Asssam University, silchar, Assam. Done B.Tech(CSE) from VBS Purvanchal Univetsity, Jaunpur (U.P.), M.Tech (CS) from DAVV, Indore(M.P.) and PhD from Assam University, Silchar.,Assam