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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 12, December 2013

A Survey on Supervised Learning for Word Sense Disambiguation

Abhishek Fulmari[1], Manoj B. Chandak[2]

Student M.Tech, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India [1]

Professor, CSE Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India [2]

Abstract: Word Sense Disambiguation (WSD) is the process of determining which sense of a word is used in a given context. Due to its importance in understanding semantics it is used in many real-world applications like web information retrieval, machine translation and information extraction. The problem of WSD is mainly considered as AI-complete problem. This paper discussed supervised approach, unsupervised approach, Naïve Bayes method, Exemplar based learning method, Decision List method for WSD.

Keywords: Word Sense Disambiguation, Supervised Approach, Naïve Bayes Methods, Exemplar-based Learning Methods, Unsupervised Approach.

I. Introduction

In the present era for searching any content people depends on the web. While searching contents they never concern about ambiguity of the word. The result of such searching is either relevant or irrelevant to the query. To overcome this problem the concept of disambiguation is introduced. If a word is considered than probably it may have ambiguity (multiple meaning), and each meaning is called as sense of that word. Hence, Word Sense Disambiguation (WSD) is explained as the process of finding appropriate sense of word in the given context. WSD is mainly considered as AI-complete problem because this type of problem can only be solved by first resolving all the difficulty problem in artificial intelligence. WSD has impact on many Natural Language Processing applications like Information Retrieval, Information Extraction and Machine Translation. The main factor to solve WSD problem is the knowledge base that it is totally relies on knowledge. Knowledge sources can vary considerably from corpora of texts, either unlabelled or annotated with word senses, to more structured resources, such as machine readable dictionaries, semantic networks, etc. Another factor that has to be considered for solving WSD problem is the features selection.

In 1940, WSD was first developed due to fast research in machine translation as a distinct computational task. In following years researchers developed many different approach to solve the problem of WSD they are supervised, knowledge based, semi-supervised and unsupervised approaches. In supervised approach the training data available is labelled i.e. they are already classified. For knowledge based approach the training data is in the form of WordNet, dictionary meanings or thesaurus. In semi-supervised approach the training data are of type labelled or unlabelled, here firstly on some of the unlabelled data are to be classified into labelled group and then used for the classification of unknown sample. For the unsupervised classification approach the training data available are raw (unannotated) data.

This paper is organized as follows: Section 2 includes different approach available WSD. Section 3 includes supervised method to solve the disambiguation problem. Conclusion in Section 4.

II. Different Approach

There are four different approaches used to perform disambiguation of the word. Approach are Supervised, knowledge based, semi-supervised and unsupervised approach.

A. *Supervised Approach*

General meaning of supervised is that the training set/data is already present which help to form the classes to differentiate the new data. A supervised method includes a training phase and a testing phase. In the training phase, a sense-annotated training corpus is required, from which syntactic and semantic features are extracted to build a classifier using machine learning techniques. In the following testing phase, the classifier tries to find out the appropriate sense for the word based on surrounding words present in the sentence. This type of method will provide better results as compared to other methods. To solve the WSD problem there are different supervised methods are available and they are probabilistic methods, method based on similarity, the methods based on discriminating rule and methods based on linear classification.

In probabilistic methods it tries to estimate set of parameter such as conditional or joint probability distribution and context. This parameter then used to assign category to new sample which maximizes the conditional probability. In similarity based method the categorization is done by comparing the features of new sample with the features of trained sample and assign the sense of most similar pattern (features). In method based on discriminating rules, some rules are used which are associated with each word sense, to classify new sample one or more rules are selected that satisfy sample feature and assign sense based on their predictions.

B. Knowledge Based Approach

In Knowledge Base method different type of resources is used to gather the information to get the sense of the word. These methods don't have better results as compared with supervised method but have the advantage of a wider range of knowledge base. In natural language processing we referred these knowledge bases as a lexical knowledge base (LKB). In NLP we have different LKB such as dictionary or thesauri, WordNet [4], SemCor, etc. For example, some resources help us to provide a textual sense meaning called gloss and also in finding the relationship between the word present in the gloss and the word which has to be disambiguate (WordNet). SemCor is another LKB which will be help us to mark the sense for word. Method to solve WSD problem are discussed below:

The method based on overlapping of sense definitions. This approach is depend on the calculation of the word overlap between the senses definitions of two or more target words (Lesk algorithm). The other variant of Lesk algorithm [12] says that score are calculated by overlapping the target word and the sense of target word.

Problems with Knowledge Based Approach

- Dictionary is limited for sense of target word.
- It does not have sufficient material to create classifier.
- The solution to problem is one can expand lists of words used in classifier.

C. Semi-Supervised Approach

Semi-supervised method will make use of both type of data, that is, labelled data and unlabelled data, this is because of the problem of lack of training data. Mainly in this approach less amount of labelled data is used as compared to unlabelled data.

Bootstrapping algorithm is commonly used semi-supervised learning method for WSD. It works by iteratively classifying unlabelled examples and adding confidently classified examples into labelled dataset using a model learned from augmented labelled dataset in previous iteration. Recent research show about graph based semi-supervised learning algorithms are introduced, which can effectively combine unlabelled data with labelled in learning process by exploiting cluster structure in data. Label propagation algorithm [8] is a graph based semi-supervised learning algorithm (LP algorithm) for WSD.

D. Unsupervised Approach

These methods acquire knowledge from unannotated raw text and assuming some similarity among the words they form clusters. These approaches based on the idea that same sense of word will have similar neighbouring words. To disambiguate a word they using some measure of similarity in context to get the correct sense. Unsupervised [7] WSD performs word sense discrimination i.e. it divides the occurrence of word into a number of classes by determining for any two occurrences whether they belong to the same sense or not. Evaluation of these methods is more difficult. Main task of unsupervised approaches are identifying sense clusters.

Unsupervised methods overcome the problem of knowledge acquisition bottleneck. The performance of unsupervised method is always been lower than that of the other method used for disambiguation.

Different method in unsupervised approach [2] are word clustering method in which words are clustered according to the semantic similarity based on single feature (e.g., subject-verb, adjective-noun, etc.) i.e. they are similar to that of target word. In another context clustering method the clusters are formed by finding the co-occurrence of word (not target) with the target word and then the centroid is calculated of the vector of words occurring in the same context. In another method called graph based method in which a graph is built on some grammatical relationship, in graph weights are assign to the edge according to the relatedness. An iterative algorithm is applied to get the word with highest degree node and finally minimum spanning tree is applied to disambiguate instance of target word.

Problems with Unsupervised Approach

- The instances in training data may not be assigned the correct sense.
- Clusters are heterogeneous.
- Number of cluster may differ from the number of senses of target word to be disambiguated.

III. SUPERVISED METHODS

A. Naïve Bayes Method

This method [2] is considered under probabilistic approach. Probabilistic approach is a statistical methods usually estimate a set of probabilistic parameters that express the conditional or joint probability distributions of categories and contexts.

Naïve Bayes classifier is based on Bayes theorem and in that conditional probability is calculated for each sense (k) of a word over the features defined (x_1, x_2, …, x_m).

$$\arg\max_k P(k \mid x_1,\ldots,x_m) = \arg\max_k \frac{P(x_1,\ldots,x_m \mid k)P(k)}{P(x_1,\ldots,x_m)}$$

$$= \arg\max_k P(k)\prod_{i=1}^{m} P(x_i \mid k).$$

P(k) and P(xi/k) are the probabilistic parameters of the model and they can be estimated from the training set, using relative frequency counts.

B. Decision List Method

A decision list is an ordered set of rules for categorizing test instances. It can be seen as a list of weighted —if-then-else rules [1]. The feature considered for each word are part-of-speech, syntactic and semantic feature, collocation vector and co-occurrence vector. Firstly the classifier is trained with the training data, training data is about the importance of feature. The rules is in the form (feature-value, sense, score) is created. These rules are sorted on the basis of score in decreasing order and form decision list.

$$\text{weight}(s_k, f_i) = \log\left(P(s_k \mid f_i) \Big/ \sum_{j \neq k} P(s_j \mid \right.$$

In the testing phase, the decision list is scanned for the entries which matches input feature vector (vector of word to be disambiguate), the sense with highest score will select as the accurate sense.

C. Decision Tree Method

This algorithm is prediction based model. The knowledge source used for the decision tree [2] is a sense-tagged corpus, on which the training is done. The classification rules over here is in the form of yes-no rules. Using these rules the training data set is recursively partitioned. The characteristic of decision tree are each internal node represents a feature, each edge represents feature value and each leaf node represents sense. The feature vector here is same as the feature vector of decision list.
In testing phase, the word to be disambiguate along with feature vector is traversed through the tree (considering training information) to reach leaf node. The sense contained in the leaf node will be consider the sense for the word.

An example to disambiguate word "bank" in sentence "Sand art was made at the bank of Ganga River."

D. Neural Network Method

Neural network [5] is also an approach in supervised method which is interconnection of artificial neurons. The neural networks used for WSD purpose are Hidden Markow Model or back propagation based feed forward network.
Input feature and the expected output are the pairs of input to the learning technique. The aim of this approach is to make use of input features to partition the training contexts into non-overlapping sets according to desired responses [11]. As the new pairs of input is provided to the training set, the weights between neurons are adjusted so that the expected output is having larger values as compare with the other outputs.

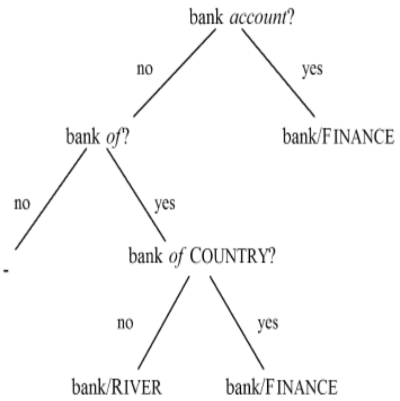

Fig. 1. An example of decision tree

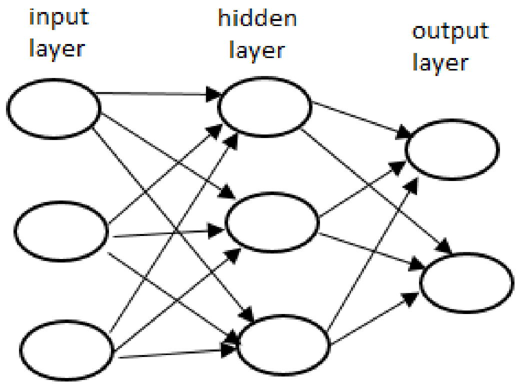

Fig. 2. A feedforward neural network WSD with 3 features and 2 responses

E. Support Vector Machine (SVM)

Support Vector Machine method is based on the idea of learning a linear hyperplane from the training set that separates positive samples from negative samples.
Basically Support Vector Machine is a binary classifier that classifies the samples into either true class or in false class. Since in WSD, one word may have more than one meaning so here SVM for WSD must be adapted to multiclass classification.

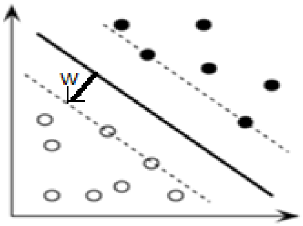

Fig. 3. An example of SVM

A simple way is to reduce the multiclass classification problem to a binary classifications problem of the kind sense Si versus all other senses.

The linear classifier [2] is based on two elements: a weight vector **w** perpendicular to the hyperplane (which accounts for the training set and whose components represent features) and a bias *b* which determines the offset of the hyperplane from the origin. An unlabeled example **x** is classified as positive if

$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \geq 0$ (negative otherwise).

The classification formula of SVM can be reduced to a function of the support vectors, which in its linear form determines the dot product of pairs of vectors. More in general, the similarity between two vectors x and y is calculated with a function called kernel which maps the original space (e.g., of the training and test instances) into a feature space such that k(x, y) = Ø(x) · Ø (y), where Ø is a transformation (k(x, y) = x . y).

F. Exemplar-based learning Method

This learning method is come under the category of memory based because the example of training data are stored in the memory [2]. For this learning method k-Nearest Neighbour (kNN) algorithm is used because the classification of testing data is based on the senses of k most similar stored example [10].

In order to obtain the set of nearest neighbour, each feature of testing data $x = (x_1, \ldots, x_m)$ is to be compared with respective feature of each training data set $x^i = (x^i_j, \ldots, x^i_m)$ and distance between them is calculated using hamming distance

$$\Delta(x, x^i) = \sum_{j=1}^{m} w_j \, \delta(x_j, x_j^i)$$

where w_j is the weight of j^{th} feature calculated using gain ratio measure and $\delta(x_j, x^i_j)$ is the distance between two values, which is 0 if $x_j = x^i_j$ and 1 otherwise.

If values of k > 1, the resulting sense is the weighted majority sense of the k nearest neighbours where each example votes its sense with a strength proportional to its closeness to the test example.

IV. Conclusion

This paper summarized the various approaches used for Word Sense Disambiguation. The hardness of WSD strictly depends on the granularity of the sense distinctions. We can say here that supervised methods perform well as compared to all other approaches. The reason behind that in supervised approach the training data is totally of domain specific and in unsupervised approach the training data is totally unannotated and to from cluster from that data set is quite difficult.

Another important fact that to be considered for WSD problem is the feature selection. The feature is selected in such a way that it will form relevant relationship with the target word and help the classifier to classify target word in correct sense.

References

[1] J. Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amjan Shaik, P. Pavan Kumar., "*Word Sense Disambiguation: An Empirical Survey*", volume 2. IJSCE, 2012.

[2] Navigli, Roberto, "*Word Sense Disambiguation: a Survey*", ACM Computing Surveys, 41(2), ACM Press, pp. 1-69, 2009..

[3] Alistair Kennedy and Stan Szpakowicz., *"A Supervised Method of Feature Weighting for Measuring Semantic Relatedness"*, 2011.

[4] C. Fellbaum., *"WordNet: An Electronic Lexical Database"*, MIT press, 1998.

[5] Agirre, Eneko and Philip Edmonds., "*Word Sense Disambiguation: Algorithms and Applications*", Springer, 2006.

[6] L. Y. Keok., "*An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation*", EMNLP, Philadelphia, pp 41–48, 2002.

[7] Ping Chen, Wei Ding, Max Choly, Chris Bowes, *"Word Sense Disambiguation with Automatically Acquired Knowledge"*, volume 24. Intelligent System, IEEE, 2012.

[8] Ankita Sati, *"Review: Semi-Supervised Learning Methods for Word Sense Disambiguation"*, volume 12, issue 4. IOSR-JCE, 2013.

[9] David Martinez, Eneko Agirre, Xinglong Wang, "*Word Relatives in Context for Word Sense Disambiguation",* ALTW, pp 42-50, 2006.

[10] Gerard Escudero, Llu´ıs M`arquez and German Rigau, "*Naïve Bayes and Exemplar-based approaches to Word Sense Disambiguation Revisited*", arXiv:CS/0007011v1, 2000.

[11] A.Azzini, C. da Costa Pereira, M. Dragoni, A. G. B. Tettamanzi, *"Evolving Neural Networks for Word Sense Disambiguation", WSPC –* Proceedings, 2008.

[12] 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.