Word Sense Disambiguation and Classification Algorithms: *A Review*

Sunita Rawat
Department of Information Science and Engineering
The Oxford College of Engineering
Bangalore, India

Dr.Manoj Chandak
Department of Computer Science and Engineering
Ramdeobaba College of Engineering
Nagpur, India

Abstract— **Natural language is most common way to communicate with each other but it's not possible to understand all the languages. To understand different languages machine translation (MT) is required. MT is the most excellent application which helps to understand any other language in very less time and cost. Related to this context some problems are faced by researchers like words which pronounce same but having totally different meaning, few words spelled different but having identical meaning, while in some cases combination of words may change the meaning. Thus Word Sense Disambiguation is needed to resolve such kind of problems. Word Sense Disambiguation is used to understand the correct meaning of the word with respect to context in which that is used. In this paper, we will discuss about different classification algorithms, Machine Translation and Word sense disambiguation.**

***Index Terms*—Natural Language Processing, Machine Translation, Word Sense Disambiguation, Supervised Learning, Unsupervised Learning.**

I. INTRODUCTION

With the growing world and business people move from one state to another and country to country. Now a day's mostly data is computerized, lot of websites and Blogs contains the useful information. When we want to access this information the problem is of understanding the text. The concept of Natural language processing is invented to sort out this problem [1]. The natural language is the most common way to share your views with people. Various applications come under the natural language processing are: Speech Recognition, Sentiment Analysis, Text Processing, Categorization, Machine Translation, Parsing, Sentence Breaking, Information Retrieval, Word Sense Disambiguation and so on. There are many kinds of resources can be used in NLP, such as dictionaries, corpus and rule base [11]. Dictionaries describe the speech of words, meanings and other attributes statically. Whereas Corpus dynamically presents the use of polysemous words in real text situation. Rule base was formulated according to the knowledge of linguistics by linguists.

WordNet domain is used for identifying the correct sense of the word. A domain may include synsets of different syntactic categories. It groups senses of the same word into homogeneous clusters, with the effect of reducing word polysemy in WordNet [12]. WordNet domain provides semantic domain as a natural way to establish semantic relations among word senses.

In natural language many words have more than one meaning and the proper meaning is determined by the word's context. Consider an example : the English word date can be defined in common use dictionaries as:

1) the fruit of the date palm, having sweet edible flesh and a single large woody seed date.
2) a romantic or social appointment

Any resident or experienced speaker of English will not have any difficulty in understanding the correct sense of this word in contexts for instance those presented in examples a and b:

a. Her favorite fruit to eat is a date
b. Joe took Aleena out on a date

However, to accomplish tasks such as machine translation (MT) and speech recognition when computational applications have to process these examples, this distinction is not always trivial. Statistical or rule-based methods are mostly used to produce better results in language processing. Word Sense Disambiguation (WSD) algorithm is used to remove ambiguity of words and correct domain of a word to be displayed. The WSD algorithm is used to find out efficient and precise meaningful sense of a word based on domain information. It consists of automatically identifying the sense of ambiguous words in context using computational methods.

As part of MT systems first studies on WSD were carried out in the 1950's. Earlier MT systems trusted on a rule-based analysis module in concern to ambiguity at the lexical semantic level and improve the output of translation software [3]. A few applications of WSD consist of information retrieval (IR) [4] whereby words are disambiguated before being used in a search engine and speech processing systems which aim to disambiguate homographic and homophonic words. Lesk [5] proposed a method that used dictionary definitions; this was among the first move towards the WSD as an independent job. The assumption of this method was that nearby words in a sentence would tend to share the same common topic or belong to related topics. A later adaptation of the Lesk algorithm replaced dictionaries with Wordnet definitions [6]. Banerjee and Pedersen [7] has proposed the approach of Wordnet which is a large lexical database of

English rich in semantic relations. State-of-the-art methods in WSD do not rely on dictionaries for disambiguation. Ng and Li [8] invented corpora, following them, researchers started to use corpora as the main source of knowledge for disambiguation.

II. Approach

There are two approaches that are followed for Word Sense Disambiguation (WSD): Knowledge Based approach and Machine-Learning Based approach. In Knowledge based approach, it requires external lexical resources like Word Net, dictionary, thesaurus etc. In Machine learning-based approach, systems are trained to perform the task of word sense disambiguation. These two approaches are briefly discussed below:

A. *achine Learning Based Approach*

In machine learning approach, the systems are trained to carry out the task of Word Sense Disambiguation. Here the role of the classifier is to learn features and assigns senses to new unseen examples. The initial input is the target word that is the word to be disambiguated and the context that is nothing but the text in which it is embedded. Part-of-Speech tagging also known as grammatical tagging is used to find the relationship with adjacent text. Features are themselves provided by the words. Feature value is the occurrence of the word in the region surrounding the target word. The techniques available based on machine learning approaches are: supervised, semi-supervised and unsupervised, among these in this paper we will focus only on supervised and unsupervised techniques.

B. *Dictionary Based Approach*

Dictionary Based Approach provides both the means of constructing a sense tagger and target senses to be used. Machine Readable Dictionaries (MRD) are used to perform large scale disambiguation. In this approach, all the senses of a word that needs to be disambiguated are retrieved from the dictionary. These senses are then compared to the dictionary definitions of all the remaining words in context. The sense with highest overlap with these context words is chosen as the correct sense.

III. Methodology

In this paper we have discussed about the two methods. First is supervised learning and second one is unsupervised learning.

A. *Supervised Learning Method*

The learning here perform in supervision. Let us take the example of the learning process of a small child. The child doesn't know how to read/write. He/she is being taught by the parents at home and then by their teachers in school. Their each and every action is supervised by the teacher. Supervised learning is the machine learning task of inferring a function from labeled training facts. A set of training examples makes the training facts. In supervised learning, each example is a pair consisting of an input object and a desired output value [13].

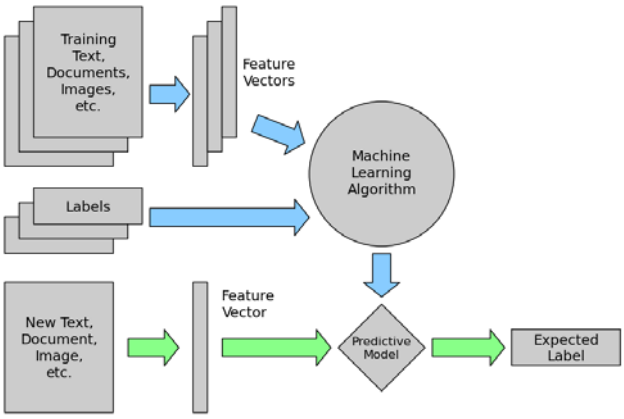

Fig. 1. Supervised Learning Model

M

In supervised learning, it is assumed that the correct (target) output values are known for each Input. So, actual output is compared with the target output, if there is a difference, an error signal should be generated by the system. This error signal helps the system to learn and reach to the desired or target output.

B. *Usupervised Learning Method*

No supervision is provided in case of unsupervised learning. Take an example of a tadpole. Child fish learns to swim without any supervision therefore its leaning process is independent. In this technique, feature vector representations of unlabeled instances are taken as input and are then grouped into clusters according to a similarity metric. These clusters are then labeled by hand with known word senses.

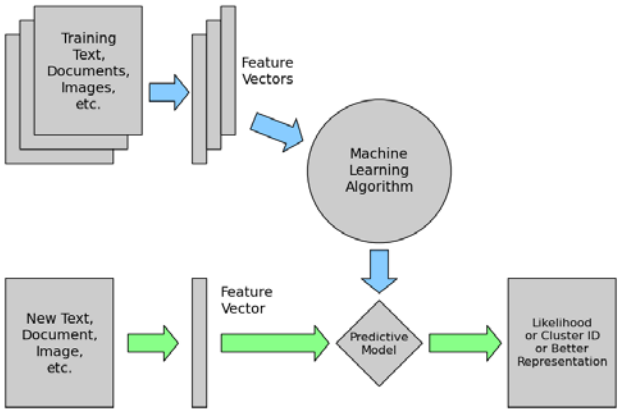

Fig. 2. Unsupervised Learning Model

In machine learning, the task of unsupervised learning is to find hidden structure in unlabeled data. Corrections to the network weights are not performed by an external agent, as in many cases we also do not know what solution network should produce. Network itself has to decide what output is best for a given input and reorganizes accordingly [14]. We will make a distinction between two classes of unsupervised learning: reinforcement learning and competitive learning. In reinforcement learning each input produces a reinforcement of the network weights in such a way as to enhance the reproduction of the desired output. In competitive learning, the elements of the network compete with each other for the "right" to provide the output associated with an input vector.

Only one element is allowed to answer the query and this element simultaneously inhibits all other competitors.

IV. ALGORITHMIC APPROACH

In this paper we have discussed 3 supervised and 3 unsupervised algorithmic approaches.

A. *lgorithms based on Supervised Learning*

Based on supervised learning 3 algorithms compared in this study (Support Vector Machines, Neural Network, Decision Trees) are generally used for WSD and differ considerably in their ways of performing classification. Three of these classifiers namely Naive Bayes, Decision Trees and Maximum Entropy are available in the Natural Language Toolkit (NLTK) [9], and rest one that is Support Vector Machines is available in WEKA Machine Learning Workbench [10].

1. Support Vector Machines (SVM) Classifier

Support Vector Machines (SVMs) is a new class of machine learning techniques which first introduced by Vapnik [15]. SVM is one of the most robust and successful classification Algorithms. It is based on the principle of structural risk minimization. SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes [19]. The key in such classifiers is to determine the optimal boundaries between the different classes and use them for the purposes of classification [19]. SVMs have been applied successfully in many text classification tasks because of their principle advantages as follow[20]: robust in high dimensional spaces, in which over fitting does not affect so much the computation of the final decision margin; robust when there is a sparsely of samples and most text categorization problems are linearly separable. Additionally, SVM method is flexible and can easily be combined with interactive user feedback methods.

2. Neural Network Classifier

Neural networks have considered as an important tool for classification. The recent vast research activities in neural network classification have found that neural networks are a good option to various conventional classification methods [22]. These classifiers consists of an input layer where patterns are presented, one or more hidden layers where actual processing is done and an output layer where answer is output [23].The basic unit in a neural network is a neuron or unit. The inputs to the network correspond to the attributes measured for each training tuple which are fed into the units making up the input layer. After weighting they are fed at the same time to a hidden layer. Usually number of hidden layers is only one but it may be arbitrary. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction [23]. Neural networks have many advantages, we can summarize some of them as [22] states as follow; first, Neural networks are able to tolerate noisy data as well as able to classify patterns on which they are not been trained. Second, inherently parallel, so parallelization techniques can be used to speed up the computational process. Third, ANN is nonlinear model that is easy to use and understand compared to statistical methods. Finally, ANN with Back propagation (BP) learning algorithm is widely used in solving various classification and forecasting problems. Although BP convergence is slow but it is guaranteed.

3. Decision Tree Classifier

*A*Decision trees are one of the most powerful used inductive learning methods. These classifiers are most commonly used particularly for data mining. Their robustness to noisy data and their capability to learn disjunctive expressions seem suitable for document classification [21]. They are designed with the use of a hierarchical division of the underlying data space with the use of different text features [19]. They are performed in two phases either tree building (top-down manner) or tree pruning (bottom-up manner). Decision tree method takes the data described by its features as input. It partitions the data of records recursively using breadth-first approach or depth first greedy approach until all the data items have assigned to a particular class.

B. *Algorithms based on Unupervised Learning*

Clustering is a type of unsupervised learning. In clustering method, objects of the dataset are grouped into clusters, like each group is different from other and the objects in the same group or cluster are very similar to each other. In clustering there are no predefined set of classes which means that resulting clusters are not known before the execution of clustering algorithm [25]. Three different clustering algorithms are chosen to investigate, study and compare them. The algorithms that are chosen are: Self-Organization Map(SOM) algorithm, k-means algorithm and hierarchical algorithm.

1. Self-Organization Maps (SOM)

Self Organization Map (SOM) uses a competition and cooperation mechanism to achieve unsupervised learning. SOM is proposed by professor T. Kohonen in1982. After adequate training the output layer of a SOM network will be separated into different regions. And different neurons will have different response to different input samples. As this process is automatic, all the input documents will be clustered.

Algorithm for SOM for text clustering can be summarized as follows:

1) Initialization: Assign some random number for all the neurons in the output layer and normalized.

2) Input the sample: Choose randomly one document from the document collection and send it to the SOM network.

3) Find the winner neuron: Calculate the similarity between the input document vector and the neuron vector, the neuron with the highest similarity will be the winner.

4) Adapt the vectors of the winner and its neighbors.

By using equation (1) adaptation can be found:

$$m_i(t+1)= m_i(t) + \alpha(t)\, h_i(t)|x(t)-m_i(t)| \quad .1$$

Where x(t) is the document vector or time t, $m_i(t)$ is the original vector of neuron I, $m_i(t+1)$ is the neuron vector after adaptation. $\alpha(t)$ and $h_i(t)$are the learning rate and neighbor rate respectively. $|\, x(t)- m_i(t)\,|$ represent the distance between neuron vector and document vector. The winner and its

neighbors are more nearer to the input document vector, after the adaptation. As a result these neurons will be more competitive if similar documents are input again.

2. Hierarchical Clustering

Hierarchical methods are well known clustering technique that can be potentially very useful for various data mining tasks. These type of clustering scheme produces a sequence of clustering in which each clustering is nested into the next clustering in the sequence [24]. Since hierarchical clustering is a greedy search algorithm based on a local search, the merging decision made early in the agglomerative process are not necessarily the right ones. Hierarchical methods are commonly used for clustering in Data Mining. To explain how hierarchical clustering algorithm works following is the pseudo code:

1. Compute the proximity matrix containing the distance between each pair of patterns (clusters).
2. Find the most similar pair of clusters using the proximity matrix and combine them into one cluster. To reflex this merge operation updates the proximity matrix.
3. If all patterns are in one cluster, stop or else, go to step 2.

3. k-means for text clustering

K-means is partition-based clustering method where items are classified as belonging to one of K-groups. The outcome of partitioning method is a set of K clusters, such that similar items falls or belongs to same cluster. Every cluster contains a centroid or a cluster representative. When the clusters are more, the centroids can be further clustered to produce hierarchy within a dataset[25]. K-means algorithm uses an iterative approach to cluster the database. The value of K that is number of clusters is defined by the user which is fixed. Euclidean Distance is used for calculating the distance of data point from the particular centroid. This algorithm consists of four steps:

1. Initialization: Initialize data set, number of clusters and the centroid for each cluster.

2. Classification: The distance is calculated for each data point from the centroid and the data point having minimum distance from the centriod of a cluster is assigned to that particular cluster.

3. Centroid Recalculation: recalculation of the centriod.

4. Convergence Conditions:

- Stop after reaching defined number of iterations.
- Stop when there is no exchange of data points between the clusters.
- Stop when a threshold value is achieved.

5. If above conditions are not satisfied, go to step 2 and repeat till the given conditions are satisfied.

COMPARATIVE ANALYSIS

In this paper we have done comparative analysis for 3 supervised and 3 unsupervised algorithms.

A. Analysis of Supervised classifires

Supervised classification is one of the tasks most frequently carried out by intelligent techniques. The large number of techniques have been developed, some of which have been discussed in the previous sections. The table I shows the comparative studies of some commonly used classification techniques from the existing evidence and theoretical studies [16] [17] [18]. This comparison shows that not a single learning algorithm outperform other algorithm all over the other datasets.

TABLE I. COMPARATIVE STUDY OF COMMONLY USED SUPERVISED CLASSIFIER

Classifier	Accuracy	Speed of Learning	Speed of Classification
Decision Trees	Good	V. Good	Excellent
Neural Network	V. Good	Average	Excellent
SVM	Excellent	Average	Excellent

B. Analysis of Unsupervised classifires

According to the number of cluster, k (Table II), except for hierarchical clustering, all clustering algorithms compared here require setting k in advance. Here, the performance of different k's is compared in order to test the performances that are related to k. To simplify the situation and to make the comparisons easier, k is chosen equal to 8, 16, 32 and 64 and the lattices for SOM are the square of them.

To compare hierarchical clustering with other algorithms, the hierarchical tree is cut at two different levels to obtain corresponding numbers of clusters (8, 16, 32 and 64). As a result, as the value of k becomes greater the performance of SOM algorithm becomes lower. However the performance of k-means algorithms becomes better than hierarchical clustering algorithm.

TABLE II. THE RELATIONSHIP BETWEEN NUMBER OF CLUSTERS AND THE PERFORMANCE OF THE ALGORITHMS

Number of clusters (K)	Performance		
	SOM	K-Means	HCA
8	59	63	65
16	67	71	74
24	78	84	87
32	85	89	92

The performance of hierarchical algorithm goes decreasing and time for execution increased as the number of records increases.

TABLE III. THE RELATIONSHIP BETWEEN NUMBER OF CLUSTERS AND THE ACCURACY OF THE ALGORITHMS

Number of clusters (K)	Accuracy		
	SOM	K-Means	HCA
8	1001	1112	1090
16	920	1089	960
24	830	910	850
32	750	840	760

According to the accuracy(Table III), SOM shows more accuracy in classifying most of the objects to their clusters than other algorithms. But as the number of k becomes greater the accuracy of hierarchical clustering becomes better until it reaches the accuracy of SOM algorithm. K-means algorithm

have less accuracy than the others. As a general conclusion, k-mean algorithm is good for large dataset and hierarchical is good for small datasets.

V. CONCLUSION

In this paper we discussed about the machine translation and word sense disambiguation in natural language processing. Also comparison of the most well known classification algorithms like decision trees, neural network, SVMs, self organizing feature maps, hierarchical clustering and k-means has been done. The aim behind this study was to learn their key ideas. Both supervised and unsupervised methods have advantages and disadvantages: on one hand, it is possible to apply simple supervised methods to disambiguate a small pre-defined set of words. Whereas, for more robust applications, unsupervised methods seems to be more suitable as they can deal with a bigger portion of the lexicon.

REFERENCES

[1] M Ozaki, Y. Adachi, Y. Iwahori, Fabio Ciravegna, Sanda Harabagiu , IEEE Computer Society 2003. Recent Advances in Natural Language Processing.

[2] http://en.wikipedia.org/wiki/Word-sense_disambiguation

[3] M. Stevenson and Y. Wilks ”Word Sense Disambiguation” in R. Mitkov Oxford Handbook of Computational Linguistics, Oxford University Press,

[4] A. Kulkarni, M. Heilman, M. Eskenazi and J. Callan ”Word Sense Disambiguation for Vocabulary Learning” Ninth International Conference on Intelligent Tutoring Systems, 2008.K. Elissa,

[5] F M. Lesk ”Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone.” In Proceedings of ACM SIGDOC Conference, Toronto, Canada, 1986 p. 25-26

[6] G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller “Introduction to Wordnet: an Online Lexical Database” in International Journal of Lexicography, 1993 p. 234 244.

[7] S. Banerjee and T. Pedersen” An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet.” in Lecture Notes In Computer Science, Springer, 2002.

[8] H. Ng and H. Lee” Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach” Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL-96), Washington, DC, 1996, pp. 40-47.

[9] S. Bird, E. Klein and E. Loper “Natural Language Processing with Python Analyzing Text with the Natural Language Toolkit”, OReilly Media. 2009.

[10] I. Witten and E. Frank ”Data Mining: Practical Machine Learning Tools and Techniques” (2nd Edition), Morgan Kaufmann, San Francisco. 2005.

[11] Ling Che and Yangsen Zhang “Study on Word Sense Disambiguation Knowledge Base Based on Multi-sources,” Published in: Intelligent Systems and Applications (ISA), (IEEE), Wuhan , 2011, PP. 1-4.

[12] Ping Chen and Wei Ding (2010), "Word Sense Disambiguation with Automatically Acquired Knowledge”, IEEE Intelligent Systems.

[13] Reza Soltanpoor, Mehran Mohsenzadeh and Morteza Mohaqeqi (2010), “A New Approach for Better Document Retrieval and Classification Performance Using Supervised WSD and Concept Graph", First International Conference on Integrated Intelligent Computing, IEEE

[14] Roberto Navigli and Mirella Lapata (2010), "An Experimental Study of graph connectivity for unsupervised word sense disambiguation", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 32, No. 4.

[15] Vapnik V. N., "The nature of statistical learning theory," Springer Verlag, Heidelberg, DE, 1995.

[16] Cover and T. Hart, “Nearest Neighbor Pattern Classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[17] D. E. Rumelhart, G. E. Hinton and R. I. Williams, “Learning internal representation by error propagation,” published in Parrallel Distrubuted Processing, 1986.

[18] J. Han and M. Kamber, Data Mining Concepts and Techniques, Elevier, 2011.

[19] Aggarwal, Charu C and Zhai Chang "A survey of text classification algorithms," In:Mining Text Data, pp. 163–213. Springer (2012).

[20] Thorsten Joachims, "Text categorization with support vector machines: learning with many relevant features, "In Proceedings of the 10th European Conference on Machine Learning ECML-98, Chemnitz, Germany. Pages 137–142. 1998.

[21] Li Y.H. and Jain A.K, "Classification of Text Documents," The Computer Jour., vol. 41, no. 8, 1998, pp. 537-546.

[22] Zhang, G.P., "Neural networks for classification: a survey," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on , vol.30, no.4, pp.451,462, Nov 2000.

[23] Navadiya Darshna and Patel Roshni, "Web Content Mining Techniques-a Comperhensive Survey, " International Journal of Engineering Research & Technology (IJERT), 2278-0181 Vol. 1 Issue 10, December- 2012

[24] S Navjot Kaur, Jaspreet Kaur Sahiwal and Navneet Kaur, “Efficent K-means Clustring Algorithm Using Ranking Method In Data Mining” ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.

[25] Kehar Singh, Dimple Malik and Naveen Sharma, “ Evolving limitations in K-means algorithm in data mining” IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011.