International Journal of Science and Engineering Applications
Volume 5– No.4, 2016 ISSN:2319-7560 (Online)

A Review on Strategies of Word Sense Disambiguation

Roshan G. Belsare
PRMIT&R, Badnera, Amravati

Prof. Shrikant P. Akarte
PRMIT&R, Badnera, Amravati

Abstract: Word sense disambiguation is an important and challenging task in natural language processing. Its goal is to find the correct sense in which a word occurs in a sentence or a query when it can have multiple meanings. It is used in various applications of NLP like machine learning, text summarization, information retrieval etc. In this paper, we made a survey of supervised, unsupervised, knowledge based and corpus based approaches of word sense disambiguation. In this paper, study of various word sense disambiguation strategies has been done.

Keywords: WordSense Disambiguation, supervised, unsupervised, knowledge based

1. INTRODUCTION

Most natural languages contain *polysemous* words that is, the words that have the same spelling but have different possible meanings or *senses.* For example, in English the word *bark* can refer to the sound made by a dog or the covering of a tree. Humans are naturally good at identifying which sense of the word is used in a particular sentence. For instance, take the sentence *The old lady got scared hearing the bark of the dog* , we immediately come to know that *bark* is used to refer to the sound made by the dog whereas given the line *the dog scratched its back on the bark of the tree* we know that bark here means covering of the tree.

However, this knowledge of human beings is based vast experience of the world as well as languages which lacks in computer programs, hence determining the correct sense applied according to the context is a difficult task. The process of differentiating between the different meanings of a polysemous word and assigning the correct meaning or sense to it is Word Sense Disambiguation. WSD is considered as AI complete problem [5].

Difficulty in WSD is due to two aspects. First, Dictionary based glosses tend to be ambiguous. Different lexicographers may tag different senses to the same instance. Second, WSD involves much world knowledge or common sense, which is difficult to verbalize in dictionaries [6].

The task description of WSD can be formulated as a method of assigning the appropriate sense to all or some words in the text T where T is a sequence of words ($w_{0,}$ w_1, ...,w_{n-1}) to find the mapping M from words to senses such that M(k) ⊆ Senses J(w_k) where M(k) is the subset of senses of w_k which are appropriate in the text T and Senses J(w_k) is the set of senses in dictionary J for word w_k.

The mapping M can assign more than one sense to w_k belonging to T but eventually the most appropriate sense is selected. Thus WSD is a classification task where word senses are the classes and the classification method classifies each occurrence of the word to more than one class based on external knowledge sources and context.

The paper has been further divided into six sections. In section II a brief discussion of history of the research done in WSD has been given. Section III gives a brief discussion of the knowledge based approaches. In section IV supervised disambiguation approach has been highlighted followed by unsupervised disambiguation approach in section V. Conclusion has been discussed in section VI.

2. LITERATURE SURVEY

In 1940s WSD was first formulated as separate computational task during the early days of machine translation. This makes it one of the oldest problems in computational linguistics. In 1949, Warren Weaver first introduced the concept in computational context.

Till 1970s WSD was a subtask of semantic interpretation systems which were developed within the field of artificial intelligence. However, since WSD systems were at the time largely rule-based and hand-coded they were prone to a knowledge acquisition bottleneck. In the 1990s, the statistical revolution swept through computational linguistics, and WSD became a paradigm problem on which to apply supervised machine learning techniques. Since then, supervised techniques have reached a plateau in accuracy, and so attention has shifted to coarser-grained senses, domain adaptation, semi-supervised and unsupervised corpus-based systems, combinations of different methods, and the return of knowledge-based systems via graph-based methods. Still, supervised systems continue to perform best.

3. KNOWLEDGE BASED APPROACH

The idea behind the knowledge based approach is to make extensive use of knowledge sources to decide upon the senses of words in a particular context. It was found that although alternate supervised approaches were more efficient than knowledge based approaches but their advantages also covered a wide range. Collocations, thesauri, dictionaries etc are the most commonly used resources in this approach. Initially knowledge based approaches started in limited domains in 1979 and 1980 [8]. There are three Knowledge based approaches which are discussed as follows:

3.1 Lesk Algorithm

M. Lesk proposed a approach to determine the overlap between words in the sense definitions of ambiguous words and the definitions of context words surrounding these ambiguous words in a given text. The biggest drawback of this algorithm is that dictionary definitions are often very short (as lexico and do not have enough words for this algorithm to work well. A modification has been proposed by Banerjee and Pedersen [9] that deals with this problem by adapting this algorithm to the semantically organized lexical database called WordNet. Besides storing words and their meaning like a normal dictionary, WordNet also ”connects” related words together. They overcome the problem of short definitions by looking for common words not only between the definitions of the words being disambiguated, but also between the definitions of words that are closely related to them in WordNet. Their algorithm achieves an 83% improvement in accuracy over the standard Lesk algorithm, and that it compares favorably with other systems evaluated on the same data..

3.2 WSD using conceptual density

The conceptual density is the measure of how the concept that the word represents is related to the concept of the words in its context. Conceptual density is related to conceptual distance inversely. The conceptual distance is determined from the WordNet.

3.3 Walker’s Algorithm

Walker proposed a simple algorithm by incorporating subject codes. His algorithm is based on the assumption that the subject codes assigned to a word reflects the sense of the word. If a word has more than one subject code then it will have more than one sense. For example: Longman’s Dictionary of Contemporary English includes subject code EC (economic) for the “financial” sense of “bank". This subject code helps us in knowing that “deposit” is related to the “financial” sense of bank [4] [10].

His algorithm is based on the assumption that the subject codes assigned to a word reflects the sense of the word. If a word has more than one subject code then it will have more than one sense. For example: Longman’s Dictionary of Contemporary English includes subject code EC (economic) for the “financial” sense of “bank". This subject code helps us in knowing that “deposit” is related to the “financial” sense of bank [4] [10].

4. UNSUPERVISED APPROACH

Unsupervised methods of WSD eliminate the need for sense tagged training data and therefore, they overcome the knowledge acquisition bottleneck [1].

Strictly speaking, using a completely unsupervised sense disambiguation task, we can only discriminate word senses. That is, we can group together instances of a word used in different senses without knowing what those senses are.

However, Yarowsky [3] proposed an unsupervised algorithm that can accurately disambiguate word senses in a large completely untagged corpus. He exploited two powerful properties of human language in an iterative bootstrapping setup to avoid the need of manually tagged training data (adapted from Yarowsky 1995)[4]:

1. One sense per discourse: The sense of a target word is highly consistent within any given document or discourse.
2. One sense per collocation: Nearby words provide strong and consistent clues to the sense of a target word, conditional on relative distance, order and syntactic relationship.

This approach has two types of distributional approaches; first one is monolingual corpora and other one is translation equivalence based on parallel corpora. And these techniques are further categorized into two types; type-based and token-based approach. The type-based approach disambiguates by clustering instances of a target word and token-based approach disambiguates by clustering context of a target word.

A. Context Clustering

In Context Clustering method [2], first context vectors are created and then they are grouped into clusters to identify the meaning of the word. This method uses vector space as word space and its dimensions are words only. Also in this method, a word which is in a corpus will be denoted as vector and the no of times it occurs will be counted within its context. After that, co-occurrence matrix is created and similarity measures are applied. Then discrimination is performed using any clustering technique.

B. Word Clustering

This technique is similar to context clustering in terms of finding sense but it clusters those words which are semantically identical. For clustering, this approach uses Lin’s method. It checks identical words which are similar to target word. And similarity among those words is calculated from the features they are sharing. This can be obtained from the corpus. As words are similar they share same kind of dependency in corpus. After that, clustering algorithm is applied to discrimination among senses. If a list of words is taken, first the similarity among them is found and then those words are ordered according to that similarity and a similarity tree is created. At the first stage, only one node is there and for each word available in the list, iteration is applied to add the most similar word to the initial node in the tree. Finally, pruning is applied to the tree. As a result, it generates sub-trees. The sub-tree for which the root is the initial word that we have taken to find sense, gives the senses of that word. Another method to this approach is clustering by committee. As mentioned earlier, the word clustering is approach is clustering by committee. As mentioned earlier, the word clustering is a kind of context clustering, this clustering by committee follows similar step, first the similarity matrix is created, so that, matrix contains pair-wise similar information about the words. And in the next step, average-link clustering is applied to the words. The discrimination among words is performed using the similarity of centroids. For each committee, one centroid exists. So, according to the similarity of the centroid, the target word gives the respective committee. In the next step, features between the committee and the word are removed from the original word set, so in next iteration, identification of senses for same word which are less frequent, is allowed.

C. Co-occurrence Graph

This method creates co-occurrence graph with vertex V and edge E, where V represents the words in text and E is added if the words co-occur in the relation according to syntax in the same paragraph or text. For a given target word, first, the graph is created and the adjacency matrix for the graph is created. After that, the Markov clustering method is applied to find the meaning of the word.

Each edge of graph is assigned a weight which is the co-occurring frequency of those words.

Weight for edge {m,n} is given by the formula:

$$\mathbf{w_{mn} = 1- max\{P(w_m \mid w_n), P(w_n \mid w_m)\}}$$

Where P(Where P($w_m|w_n$) is the $freq_{mn}/freq_n$ where $freq_{mn}$ is the co-occurrence frequency of words w_m and w_n, $freq_n$ is the occurrence frequency of w_n. Word with high frequency is assigned the weight 0, and the words which are rarely co-occurring, assigned the weight 1. Edges, whose weights exceed certain threshold, are omitted. Then an iterative algorithm is applied to graph and the node having highest relative degree, is selected as hub. Algorithm comes to an end, when frequency of a word to its hub reaches to below threshold. At last, whole hub is denoted as sense of the given target word. The hubs of the target word which have zero weight are linked and the minimum spanning tree is created from the graph. This spanning tree is used to disambiguate the actual sense of the target word.

5. SUPERVISED APPROACH

Approaches relying on sense tagged corpora for disambiguation are known as supervised. WSD approaches. They yield very high accuracy in the domain of the training corpus. But this accuracy comes at the cost of sense tagged corpora which is a costly resource in terms of the time and the manual efforts involved. Creating such corpora for all languages in all domains will be impracticable. Hence these approaches cannot be easily ported to different languages or domains. Some good supervised approaches are mentioned below [8].

A. Decision Tree

A decision tree [11-12] is used to denote classification rules in a tree structure that it recursively divides the training data set. Internal node of a decision tree denotes a test which is going to be applied on a feature value and each branch denotes an output of the test. When a leaf node is reached, the sense of the word is represented (if possible). For example, The noun sense of the ambiguous word “bank” is classified in the sentence, “I will be at the bank of Narmada River in the afternoon”.

B. Neural Networks

In the Neural Network based computational model , artificial neurons are used for data processing using connectionist approach. The input includes the input features and the target output and goal is to partition the training context into non-overlapping sets. The training dataset is divided into sets which are non-overlapping based on desired responses. When the network encounters new input pairs the weights are adjusted so that the output unit giving the target output has the larger activation. The network can have weights both positive and negative corresponding to correct or wrong sense choice. Neural networks can be used to represent words as nodes and these words will activate the ideas to which they are semantically related. The inputs are propagated from the input layer to the output layer through the all intermediate layers. The input can easily be propagated through the network and manipulated to arrive at an output. It is difficult to compute a clear output from a network where the connections are spread in all directions and form loops.

C. Naïve Bayes

Naive Bayes classifier is the classifier based on Bayes theorem and assumes that every feature is class conditionally independent of every other feature. This approach classifies text documents using two parameters: the conditional probability of each sense (Si) of a word (w) and the features (fj) in the context.

WSD is very tough problem and needs large number of lexical and knowledge resources like sense tagged corpora, machine readable dictionaries *etc*. It is evident that use of such resources improves the performance of WSD. Hence one might think that, if such resources are available, and then why not use them? Or why not spend sufficient time in creating high quality resources and perform great in terms of accuracy. The main reason is that, even if we have all possible resources to build a great supervised approach, it cannot be ported to other language easily. The resources have to be replicated for all possible languages. Another disadvantage of using the supervised approaches is, by using fixed sense repositories; we constrain our self to the fixed number of senses present in that repository. We cannot discover new senses of words, which are not present in the sense repository. Hence only considering the accuracy of the approach is not a good idea, but considering its versatility and portability to other languages and domains is also equally important. This is the reason we see many unsupervised approaches being tried by many researchers in WSD [7].

6. CONCLUSION

WSD is a very complex task in Natural language processing as it has to deal with complexities found in a language. In this paper we have put forwarded a survey of comparison of different approaches available in word sense disambiguation with primarily focusing on the knowledge based, supervised and unsupervised approaches. We concluded that supervised approach is found to perform better but one of its disadvantage is the requirement of a large corpora without which training is impossible which can be overcame in unsupervised approach as it does not rely on any such large scale resource for the disambiguation. Knowledge based approach on the other hand makes use of knowledge sources to decide upon the senses of words in a particular context provided machine readable knowledge base is available to apply.

7. REFERENCES

[1] Gale, W. A.,Church, K., and Yarowsky, D. 1992b. A method for disambiguating word senses in a corpus. Comput Human. 26, 415–439.

[2] Niu, C., Li, W., Srihari, R. K., Li, H., Crist, L.,(2004) “Context Clustering for Word Sense Disambiguation Based on Modeling Pairwise Context Similarities”, SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July 2004.

[3] Yarowsky, D.,1995,’Unsupervised word sense disambiguation rivaling supervised methods,’ *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, pp. 189-96.

[4] U. S. Tiwary, Tanveer Siddiqui,2008, Natural Language Processing and Information Retrieval.

[5] Samit Kumar, Neetu Sharma, Dr. S. Niranjan, “Word Sense Disambiguation Using Association Rules: A Survey”,

International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 2, Issue 2, 2012

[6] Veronis, J. 2000. Sense Tagging: Don't Look for the Meaning But for the Use, *Workshop on Computational Lexicography and Multimedia Dictionaries*, 1-9. Patras, Greece.

[7]Alok Ranjan Pal, Diganta Saha, 2015, Word Sense Disambiguation: A Survey.

[8] J. Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amjan Shaik, P. Pavan Kumar, “Word Sense Disambiguation: An Empirical Survey”, International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-2, May,2012.

[9]Satanjeev Banerjee, Ted Pedersen, 2002, An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet.

[10]Walker, Donald, 1987, ‘Knowledge resource tools for accessing large text files,’ Machine Translation: Theoretical and Methodological Issues, Sergei Nirenberg(Eds.), Machine Translatin of Languages, John Wiley & Sons, New York, pp.

[11] Singh, R. L., Ghosh, K. , Nongmeikapam, K. and Bandyopadhyay, S.,(2014) “A DECISION TREE BASED WORD SENSE DISAMBIGUATION SYSTEM IN MANIPURI LANGUAGE”, Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, July 2014, pp 17-22.

[12]http://www.d.umn.edu/~tpederse/Pubs/naacl01.pdf date: 14/05/2015.

[13]Devendra Singh Chaplot, 2014, Literature Survey on Unsupervised Word Sense Disambiguation.

[14] Pranjal Protim Borah, Gitimoni Talukdar, Arup Baruah, , 2014, Approaches for Word Sense Disambiguation **:** A Survey, International Journal of Recent Technology and Engineering (IJRTE),ISSN: 2277-3878, Volume-3, Issue-1, March 2014