A Walk Through the Approaches of Word Sense Disambiguation

Dhanya Sreenivasan
Department of Computer Science & Engineering
Vidya Academy of Science & Technology Thrissur, India

Vidya M
Department of Computer Science & Engineering
Vidya Academy of Science & Technology Thrissur, India

Abstract

Word Sense Disambiguation (WSD) is an important and challenging technique in the area of natural language processing (NLP). A particular word may have different meaning in different context. So the main task of word sense disambiguation is to determine the correct sense of a word used in a particular context. Ambiguities provide great difficulty in the use of Natural languages. Words that occur in a particular context can be interpreted in more than one way depending up on the context. Natural languages like Malayalam also have many ambiguity words to be resolved. WSD plays a major role to improve the quality of the system. Here, we put forward a survey of various methods available in word sense disambiguation (WSD) and Malayalam WSD approaches.

Keywords: Word Sense Disambiguation, Natural Languages Processing

I. INTRODUCTION

Word Sense Disambiguation is a task that determines the correct sense, selected from a set of different senses of a polysemic word in a particular context [1] [2]. Polysemic words are the same word with different senses or meaning. Many polysemic words are there in natural languages. WSD system gives the exact sense. It is an important and challenging technique of natural language processing (NLP). The main applications of WSD are machine translation (MT), semantic mapping (SM), semantic annotation (SA), ontology learning (OL), information retrieval (IR), information extraction (IE), and speech recognition (SR).

Words with more than one sense are called ambiguous words and the process of determine the exact sense among them in that context is called Word Sense Disambiguation. A normal human being has the ability to differentiate the diffrent senses of an ambiguous word in a particular context. Malayalam is a Dravidian language used by around 36 million people in the state of Kerala. Malayalam WSD system disambiguates the polysemic word from Malayalam sentence. WSD approaches are categorized mainly into two types. Knowledge-based and machine learning methods. Knowledge based method uses external lexical resources such as dictionaries, thesaurus, WordNet etc. In machine learning approaches, systems are trained to perform word sense disambiguation. This approach again classified into Supervised and Unsupervised learning method. In Supervised Learning method, training set contains feature encoded inputs along with their appropriate category, or label. In unsupervised learning the classification of the data in the training sample is unknown. it is a clustering task.

This paper is divided into sections. Different WSD Approaches are discussed in section two; Section three depicts various Malayalam word sense disambiguation techniques.

II. DIFFERENT WSD APPROACHES

Word Sense Disambiguation Approaches are mainly classified into three main categories - a) Knowledge based approach, b) Supervised approach and c) Unsupervised approach.

A. Knowledge based approaches:

Knowledge-based approaches are based on knowledge sources such as machine readable dictionaries, thesauri, or sense inventories etc. WorldNet is the mostly used machine readable dictionaries in this research area.

1) LESK Algorithm:

This algorithm mainly based on the overlapping of the sense bag and context bag. Sense bag contains different senses of the ambiguous word. In the Context bag, the words in the definition of each sense of context word is included[3][4]. Then calculate the overlapping of these two bags. The maximum number of overlaps represents the correct sense of the ambiguous word.

2) Walkers Algorithm:

It is a thesaurus based approach. For each sense of the target word find the thesaurus category to which that sense belongs. then calculate the score for each sense by using the context words. Context words will add 1 to the score of the sense if the thesaurus category of the word matches that of the sense. The higher score sense will be the output.

3) Semantic Similarity:
The words that are similar share common context [5]. So the exact sense is chosen with the smallest semantic distance. similarity measures are used to determine how much two words are semantically related, When more than two words are there. This approach is computationally intensive.
4) Selectional Preferences:
This method find similar relations of word types, and define common sense between them using the knowledge source[6]. For example, Modeling-dress, Walk-shoes are the words with semantic relationship. In this approach improper word senses are not taken into account. The basic idea is to count how many times this kind of word pair occurs in the knowledge source with syntactic relation. From measure, correct senses of words will be identified.
5) Heuristic Method:
This approach based on heuristics. The heuristics are evaluated from different linguistic properties to determine the exact sense. Three types of heuristics are used for WSD system, Most Frequent Sense, One Sense per Discourse and One Sense per Collocation. The Most Frequent Sense fined all similar senses that a word can have. a word will preserve its meaning among all its occurrences in a given text in the one sense per discourse category. One Sense per Collocation is same as One Sense per Discourse except that words that are nearer provide strong and consistent signals to the sense of a word.

B. Supervised approaches

The supervised approaches use machine-learning technique from manually created sense-annotated data. Training set consists of examples related to target word. Each occurrence of an ambiguous word is annotated with semantic label. The main task is to build a classifier which correctly classifies new cases based on their context of use.
1) Decision List:
It is based on set of "if-then-else" rules. Training sets consists of set of features for a given word. Using the rules, determine the parameters like feature-value, sense, score. First the given word occurrence is calculated and then create the decision list based on feature vector. Then the score is calculated from that [7]. The maximum score represents the correct sense.
2) Decision Tree:
A decision tree is a tree structure. it use classification rules in a tree structure that recursively divides the training data set. Parent node of a decision tree denotes a test which is going to be applied on a feature value [8]. Each branch denotes an output of the test. The exact sense of the word is represented in the leaf node.
3) Naïve Bayes:
It is a probabilistic classifier which is based on Bayes Theorem. Two parameters are used for the classification of text document. The conditional probability of each sense (Si) of a word (w) and the features (fj) in the context [9][10]. The maximum value evaluated from the Bayes formula represents the most accurate sense in the context.
4) Neural Networks:
Here artificial neurons are used for data processing using connectionist approach. Learning program input is the input features, and goal is partitioning the training context into non overlapping sets. Newly formed pairs and link weights are gradually adjusted to produce a larger activation. Neural networks can be used to represent words by nodes and these words will activate the ideas in which they are semantically related. The inputs propagated from the input layer to the output layer through all the intermediate layers. The input easily can be propagated through the network and manipulated to get an output. It is difficult to compute a clear output from the network where the connections spread in all directions and form loops.
5) Exemplar-Based or Instance-Based Learning:
This model uses examples as point in feature space. Gradually added new examples will be considered for classification. The k-nearest neighbor algorithm is used here. In this procedure, first, all certain number of examples is collected; after that the Hamming distance of an example is calculated by using k–NN algorithm [11]. This distance calculates the closeness of the input with respect to the stored examples. If k >1, that represents the majority sense of the output sense among the k-nearest neighbors.
6) Support Vector Machine:
The goal of this approach is to separate positive examples from negative examples with maximum margin. Margin is the distance of hyper plane to the nearest of the positive and negative examples [12]. The positive and negative examples which are closest to the hyper plane are called support vector. This algorithm finds a hyper plane in between these two examples, so that, the separation margin between these two classes becomes maximum. It finds a hyperplane between two classes.
7) Majority Voting:
In this method, one vote is given to a particular sense of the word. Sense which having maximum majority votes will be selected as final sense of the word. If tie occurs, then random choice is done.
8) Probability Mixture:
In this strategy, target word is evaluated by the first order classifiers and then normalization is applied. As a result the probability distribution on the senses of the word is obtained. Next, these probabilities are added, and score is calculated. The sense with highest score, considered as the exact sense.

9) Rank-Based Combination:
Here First order classifier is used to rank the senses for a given input target word. Sense with the maximum value among the summations of its rank will be the output sense.
10) AdaBoost:
This method creates strong classifiers by the linear combination for several weak classifiers. The method used here finds the misclassified instances from previous classifier, so that it can be used for further upcoming classifier. The classifiers are learns from weighted training set and at the beginning, all the weights are equal. At every step, it performs certain iteration for each classifier. In every iteration, weight for the classifier which is incorrect is increased. So the upcoming classifiers can focus on those incorrect examples.

C. Unsupervised approaches

Unsupervised WSD [13] methods do not depends on external knowledge sources or sense inventories, machine readable dictionaries or sense-annotated data set. This approach has two types of distributional approaches; first one is monolingual corpora and other one is parallel corpora based on translation equivalence. These techniques are again categorized into type-based and token-based approach. In the type-based approach disambiguation is done by clustering instances of a target word and in the token-based approach disambiguation is done by clustering context of a target word.
1) Context Clustering:
This method is based on clustering techniques. Clustering depends upon the context of words. Here first, context vectors are created for context words and then they will be grouped into clusters to identify the meaning of the word. vector space is used as word space and its dimensions are words. A word which is in a corpus will be denoted as vector and how many times it occurs will be counted within its context [14]. Then co-occurrence matrix is created and similarity measures are applied in that matrix. Then discrimination is performed using any clustering technique.
2) Word Clustering:
Word clustering is similar to context clustering in terms of finding sense. but here, clusters those words which are semantically identical. This approach uses Lin's method. It identified the identical words which are similar to target word. And similarity among those words is calculated using the features they are sharing. This can be done from the corpus. then clustering algorithm is applied to discrimination among senses. If a collection of words is taken, first the similarity among them is identified by using measures. Then words are arranged in an order according to the similarity and create similarity tree. At the starting stage, only one node is there and for each word available in the list, iteration is applied. Finally, pruning is applied to the tree. As a result, it generates sub-trees. The sub-tree where the root is the initial word that we have taken to find ambiguity, gives the senses of that word.
3) Co-occurrence Graph:
This method creates co-occurrence graph with edge E and vertex V , where E is added if the words co-occur in the relation according to syntax in the same text or paragraph and V represents the words in text and. For a given input target word, first, the graph is created and then adjacency matrix for the graph is determined. After that, the Markov clustering method is applied to the graph to find the meaning of the word. Each edge of graph is assigned a weight which represents the co-occurring frequency of those words. Weight for edge {m,n} is given by the formula: wmn = 1- max{P(wm | wn), P(wn | wm)} Where P(wm|wn) is the freqmn/freqn where freqmn is the co-occurrence frequency of words wm and wn, freqn is the occurrence frequency of wn. Word having high frequency is assigned the weight zero, and assigned the weight one for the words which are rarely co-occurring,. Edges, whose weights exceed certain value threshold, are omitted. Then an iterative algorithm is applied and then the node having highest relative degree is selected as hub. Algorithm stops or come to an end, when the frequency of a word to its hub reaches to below threshold value. At last, whole hub is represented as sense of the given target word. The hubs of the target word which have weight zero are linked and from the graph, the minimum spanning tree is created. This spanning tree is used to disambiguate the correct sense of the target word [15].
4) Spanning tree based approach:
The idea of this method is that a given word carries a specific sense in a particular context when it co-occurs with the same neighboring words. In this approach, first a co-occurrence graph (Gq) is constructed. Then all the nodes whose degree is 1 are eliminated from Gq. The maximum spanning tree (MST) TGq of the graph is determined. Then, the minimum weight edge e_TGq is removed from the graph one by one, until the N connected components that are the word clusters are formed or until there remains no more edges to eliminate.

III. MALAYALAM WSD APPROACHES

Malayalam is a Dravidian language commonly used in the state of Kerala, in southern India. It is one of the 22 official languages of India, and it is used around 36 million people in the world. There are so many people in our state who prefer their native language for interacting with the computer system. Internet plays an important role in our day to day life. Now a day, if anyone who is not at all comfortable in English language can also use Internet activities in their own native language. Here comes the need of implementing the system

Malayalam document is taken as input, and polysemic words in the documents are detected, and if any they are disambiguated. A knowledge based approach is used here for disambiguation. Absence of training corpora in Indian languages like Malayalam

prevents us to use the machine learning methods. System is implemented in two ways. One approach used is based on a hand devised knowledge source and the other is using the concept of conceptual density, by using Malayalam WordNet as the lexical resource [16]. Various Malayalam WSD approaches are follows:

1) The Lesk and Walkers approach:

In this approach, lesk and walkers algorithm is used. The collection of the contextual words is taken as as context bag. Next, sense bag, containing words with all the diffrent senses are generated from the Knowledge source [16]. After that, the overlap between the contextual words and the sense bags are measured. A score of 1 is added to each overlap occurrences, if any overlap is there. Highest score for a sense is selected as the winner.

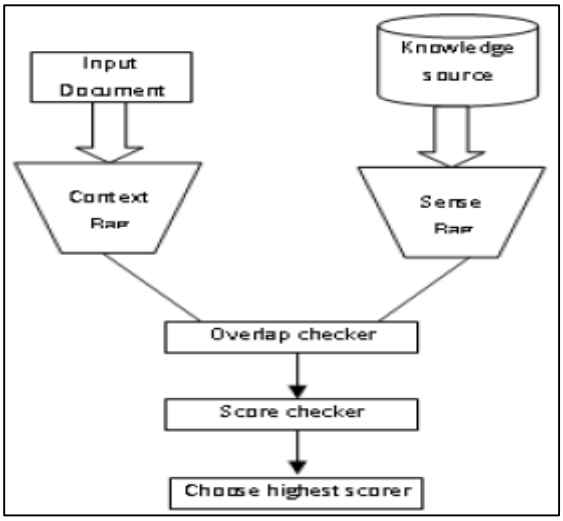

Fig. 1: Lesk and Walkers system design

This algorithm is designed based on Lesk Algorithm and walkers Algorithm which are proposed by Michel M Lesk and Walker respectively. System Architecture is shown in Fig1.

2) The Conceptual Density based Algorithms:

It find the semantic relatedness between the words in the input. It is measured in many ways. One way is to considering the Depth, Path, and Information content of words in the WordNet. This algorithm, depth is the main measurement criteria. For each sentence, the sentence is tokenized, then next, in a sequence of steps, the stop words are removed and stemming is performed[16]. Then, the ambiguous word in the input sentence is detected. If an ambiguous word is detected, that word is stored into one document and sense lookup is performed. After that, the nouns are extracted from the sentence and saved it as a document. For each sense in the sense lookup, the depth with each noun is calculated. If there are more than one nouns, depth of each noun is added and taken as the depth. The sense, which having lower depth that is the highest conceptual density is selected as the correct sense .

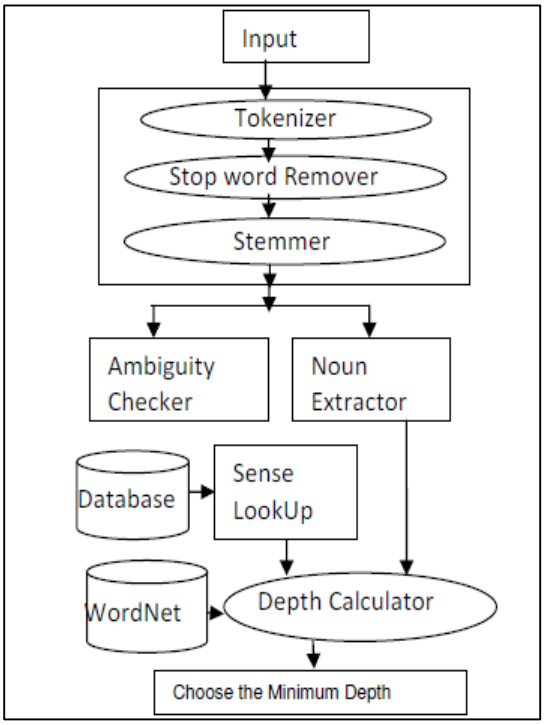

Fig. 2: System Design using conceptual density

System design using conceptual density based algorithm is shown in fig 2.For each sentence, Tokenized the sentence, Remove the stop words, perform stemming, and Check for ambiguous words. If ambiguous word occurs, shift that word into one document and sense lookup is performed. then Extract the nouns from the sentence and save it as a document. For each sense in the sense lookup, calculate the depth with each noun. If there are multiple nouns, depth of each will be added and taken as depth. The sense which results in lower Depth (highest conceptual density)is selected as the correct sense.

3) Memory Based Approach:

This approach solving WSD using memory based approach. Memory based approach is a classification-based, supervised machine learning approach. It keeps all training data in memory and abstract the data from the similar items in memory at classification time. The machine learns how to associate a word sense in a particular context using manually collected annotated corpus. Tokenization, POS tagging, sense tagging and Training the model are the major tasks in the system. For POS tagging hierarchical BIS tag set is used. Sense tag is the combination of BIS tag of the word along with its sense. Using TiMBL model is generated from the training corpus. The system is then with a sample untagged Malayalam text. The output of the system is a sense tagged text. the system design is shown in fig 3.

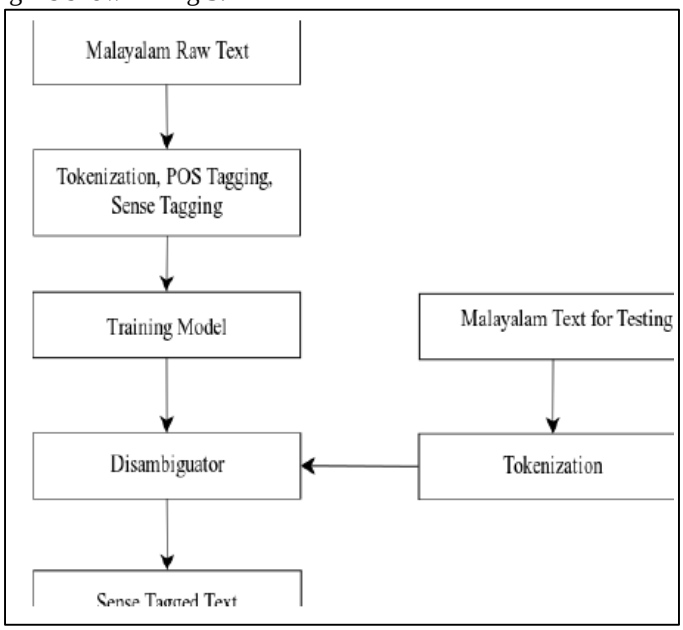

Fig. 3: System Design of memory based approach

Corpus based WSD methods by using memory based approach result in higher accuracy.. But for an Indian language like Malayalam, availability of corpus for training is very less. We can improve the accuracy of the system by increasing the size of the corpus. Manual sense tagging is quite a time consuming and difficult process. The accuracy of the WSD and performance of the system depends on size of the corpus [17]. As a future work, Corpus creation and sense tagging can be automated. Future work can also include, improving the performance of the system by using large training corpus and handling morphology exhaustively.

4) Support Vector Machine approach:

It is a corpus based approach to malayalam word sense tagging,where machine learning technique called support vector machines(SVM).it make use of contextual feature information along with the part-of-speech tag feature inorder to predict the various WSD classes. Training set contains limited number of ambiguous words has been manually annotated with 16 WSD classes.it also handling morphology exhaustively.

5) Language Model Approach:

This is a new method of word sense disambiguation for malayalam languages. it is a supervised learning system. Training is done by an annotated corpus of 10,000 words. This model checks the trigram possibility in the training corpus in terms of tag occurences. For better tagging results, Morphological Analyzer and Named Entity Recognizer are used with the languages model. accuracy can be improved by increasing the size of annotated corpus[18]. The system architecure is shown in fig4.

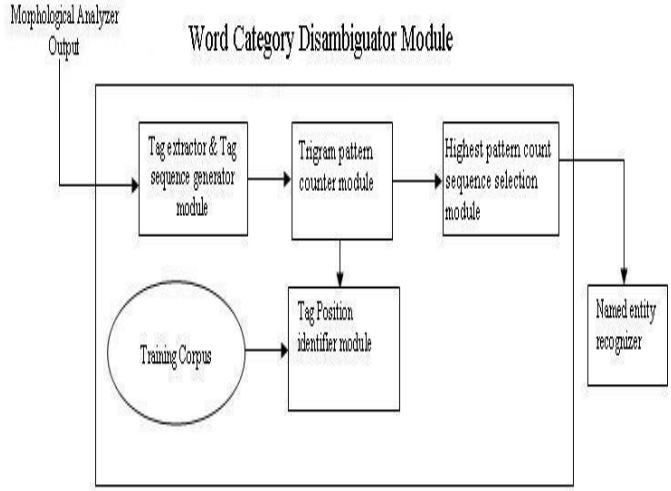

Fig. 4: System architecture of language based approach

IV. CONCLUSION

This paper focused on various word sense disambiguation methods and Malayalam word sense disambiguation approaches. The research work in WSD has been preceded up to different extents according to the availability of different resources like corpus, WordNet, thesauri tagged data set etc. In Asian languages, due to large scale of morphological inflections, development of corpus, WordNet and other resources are is under progress. Language like Malayalam, availability of corpus for training is very less. The accuracy of the WSD and performance of the system depends on size of the corpus. accuracy of word sense disambiguation techniques can be improved by large training corpus.

REFERENCES

[1] Ide, N., Véronis, J., (1998) "Word Sense Disambiguation: The State of the Art", Computational Linguistics, Vol. 24, No. 1, Pp. 1-40.
[2] Cucerzan, R.S., C. Schafer, and D. Yarowsky, (2002) "Combining classifiers for word sense disambiguation", Natural Language Engineering, Vol. 8, No. 4, Cambridge University Press, Pp. 327- 341.
[3] Banerjee, S., Pedersen, T.,(2002) "An adapted Lesk algorithm for word sense disambiguation using WordNet", In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February.
[4] Lesk, M.,(1986) "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", Proceedings of SIGDOC.
[5] Mittal, K. and Jain, A.,(2015)"word sense disambiguation method using semantic similarity measures and owa operator", ictact journal on soft computing: special issue on soft computing theory, application and implications in engineering and technology, january, 2015, volume: 05, issue: 02.
[6] Patrick, Y. and Timothy, B.,(2006) "Verb Sense Disambiguation Using Selectional Preferences Extracted with a State-of-the-art Semantic Role Labeler", Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), pages 139–148.
[7] Parameswarappa, S. and Narayana V.N,(2013) "Kannada Word Sense Disambiguation Using Decision List", Volume 2, Issue 3, May – June 2013, pp. 272-278.
[8] 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.
[9] Le, C. and Shimazu, A.,(2004)"High WSD accuracy using Naive Bayesian classifier with rich features", PACLIC December 8th-10th, 2004, Waseda University, Tokyo, pp. 105-114.
[10] Aung, N. T. T., Soe, K. M., Thein, N. L.,(2011)"A Word Sense Disambiguation System Using Naïve Bayesian Algorithm for Myanmar Language", International Journal of Scientific & Engineering Research Volume 2, Issue 9, September-2011, pp. 1-7.

[11] Brody, S., Navigli, R., Lapata, M.,(2006) "Ensemble Methods for Unsupervised WSD", Proceedings the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 97–104, Sydney, July 2006.
[12] Buscaldi, D., Rosso, P., Pla, F., Segarra, E. and Arnal, E. S.,(2006)"Verb Sense Disambiguation Using Support Vector Machines: Impact of WordNet Extracted Features", A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 192–195.
[13] Martín-Wanton, T. , Berlanga-Llavori, R.,(2012)"A clustering-based Approach for Unsupervised Word Sense Disambiguation", Procesamiento del Lenguaje Natural, Revista no 49 septiembre de 2012, pp 49-56.
[14] 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.
[15] Navigli, R. (2009) "Word Sense Disambiguation: a Survey", ACM Computing Surveys, Vol. 41, No.2, ACM Press, Pp. 1-69.
[16] Rosna P Harron, "Malayalam Word Sense Disambiguation", IEEE International Conference, 2010.
[17] Robert Jesuraj K and P. C. Reghu Raj, "MBLP approach applied to POS tagging in Malayalam Language", NCILC, 2013.
[18] T Dinesh, V Jayan, V K Bharan "Word category Disambiguation for Malayalam: a language model approach" proceedings of the second international conference on computer science, engineering