International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Supervised Word Sense Disambiguation using Decision Tree

Sunita Rawat

Abstract*: Semantic processing is an essential task in natural language processing. In semantic processing it has observed that some words have more than one meaning. Multiple meanings of a word create serious problems to linguists which produces ambiguity in sentence. Word Sense Disambiguation is one of the main challenges in natural language processing which is present in almost all the languages. By existing knowledge and experience human can certainly disambiguate the words but for machine it is difficult task. In the proposed work, we are resolving the ambiguity of all open class word in English sentence and translating it to the Hindi sentence. We have used decision tree as a classifier. For improving the speed of translation we have used the concept of translation memory.

Index Terms: Word sense disambiguation (WSD), Classification and regression tree (CART), polysemous word.

I. INTRODUCTION

According to current scenario for searching any type of information people are mostly dependent on the web. While searching information on the web if query consist of ambiguous words then chances are less that we will get all the relevant information. And this problem arises in all the languages. The major problem created by these polysemous words is in translation task. One of the solutions towards this problem is the recognition of text similarity that means to find out how close is the interpretation of two specified texts [1].
For example, the word "bank" has more than one meaning. We can use it in terms of river bank also and we can use it in terms of commercial bank also. Consider the statements

- Shipra will withdraw money from bank
- Mayank is taking walk beside river bank

As a human being we know the word bank is used in which sense but system can't predict in which sense we have used that word in the particular context. When a word has numerous interpretations then it is treated as ambiguous word. Therefore, in natural language processing Word sense disambiguation (WSD) becomes a major problem [2].
Instead of providing solution for this problem in terms of computer programming, natural language processing has started to make use of corpus for the task of WSD. These methods make use of statistics and machine learning. From long time in the field of computational linguistics main goal is automatically disambiguate the ambiguous words. But because of different reasons researchers are not able to give the 100% accuracy in their work [3]. Flow of the rest of the paper is as follows: Section 2 covers the literature survey in which work done by the various researchers on technique of word sense disambiguation is discussed. In section 3 we have discussed the concept of decision tree in detail. Section 4 discusses the architecture of the proposed system and finally we have concluded the paper.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Sunita Rawat*, Assitant Professor, Computer Science and Engineering Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

II. LITERATURE REVIEW

Machine translation task is started long back. When machine deals with different languages the problem comes into picture is of ambiguity. To disambiguating this ambiguity is the big challenge for the researchers. So many researchers have worked on this topic of word sense disambiguation and reached to different results. For removing the ambiguity WSD task is the oldest task. We will discuss work done by some of the researchers on Word Sense Disambiguation.

In [4], author has designed a machine translation system using rule based approach which translates English language to Marathi language. Grammatical structure of the target language is the main focus for the translation. Maximum rules are constructed for the target language for getting better translation. Still some exceptions are there in the language which has no effect of these rules. These exceptions can be solved at the cost of increase in complexity and size of the knowledge base.

In [5], author has worked on unsupervised Graph based WSD which has the capability to eliminate the accuracy gap from the supervised learning techniques. They have discussed the work done by different researchers on WSD for Indian languages. Then there is a discussion on work done on WSD for Asian languages. Problem they are facing while working on Asian languages is lack of resources like proper corpus and availability of WordNet. As size of corpus plays important role in accuracy of the WSD. In [6] author has surveyed on different methods used by researchers for solving the problem of WSD in Indian as well as international languages.
All those techniques of WSD are discussed along with their applications and comparison among them in the paper. Author has also discussed the good progress in Indian languages because of availability of morphological inflections in large amount.

In [7], author has done the comparison between Random Forest which belongs to supervised techniques and combination of Random Forest and K-Means cluster which belongs to unsupervised techniques.

Both the ideas are explained in detail and came to the conclusion that Random Forest classifier with K-Means cluster has higher accuracy as compared with the other considered method.

In [8], author has make use of cosine similarity on supervised learning algorithm for performing the task of word sense disambiguation in Hindi language. For performing the experiment database having 90 ambiguous words were considered. Sense was assigned to the ambiguous words by using cosine similarity. Performance of the system is calculated in terms of precision and recall.

In [9], author has used unsupervised method for disambiguating the ambiguity in the Hindi language. Training is done by using Hyperspace Analogue to Language (HAL). For forming the clusters Fuzzy C–means method is used. At last testing is done by mapping the test data with high dimensional space. Author has also concluded that this technique is not limited to particular language.

In [10], author has considered Assamese language for word sense disambiguation and used Walker algorithm for disambiguation. To check the performance of the system algorithm has been tested on manually designed sentences as well as randomly taken sentences. Author has concluded that if we increase the window size accuracy will be more.

In [11], author has considered Malayalam language for word sense disambiguation and used Naïve Bayes Classifier for disambiguation. The main problem solved by this technique is of Lack of good Corpus. Author found that the quality of word sense disambiguation system is dependent on the quality of corpora used for the research. For this work they have considered corpus which contains 1 lakh words.

In [12], author has make use of knowledge based technique with some restrictions. In this WSD system use of Hindi WordNet is done which was built by using collocation and co-occurrence. Accuracy of the system was average and reason given for less accuracy was some of the considered words were not tagged properly with part of speech tagger.

In [13], author has considered Manipuri language for word sense disambiguation and used supervised approach for disambiguation. The classifier used in this paper is based on the decision tree. Here dataset is consists of Manipuri language having 672 sentences. Total 2K polysemous words are present in this dataset. To find out the sense of the polysemous words techniques used are Context based and conventional positional features. But dataset considered for this work was of very limited size.

In [14], author has worked on Lesk algorithm which comes under the knowledge based method. Author has extended the Lesk algorithm by including lexicon network. For creating lexicon network semantically tagged glosses are used. A method is developed which makes use of lexicon network for creating semantic trees for each sense. Proposed method decreases the computational volume for finding the the best variant detection.

In [15], author has done the survey on different methods used for word sense disambiguation. Detail explanation is given about how to find the ambiguity and how to remove it. Three machine learning techniques i.e. supervised, unsupervised, knowledge based are discussed along with their applications and comparison among themselves.

III. OVERVIEW OF DECISION TREES

The decision tree applies the prediction based strategy. In machine learning algorithms Decision Tree is the very popular algorithm to be used. Searching mechanism perform by this algorithm is from general to specific of a feature space. Training of decision tree is performed on sense-tagged corpus which is provided as knowledge source to it.

Training data set is partitioned recursively in decision tree by applying classification rules which are in the format of yes or no. Aim is to choose a smallest set of features that precisely partitions the feature space into modules of observations which will collect them into a tree. Here, in our work, sense-tagged sentences having ambiguous words are the observations and different available senses are termed as partitions.

During searching each chosen feature is shown by a node in the learning decision tree. Among different feature values each node shows a selection point. Till all the training examples are not considered learning process continues. Actually, this kind of a tree will be very precise to the training data and not that much generalize to unknown data or new examples. Hence, after learning process there is a pruning step by which a new tree is produced by reorganizing or eliminating some nodes which tends to new circumstances.

Disambiguation of test instances is done by finding the path in the learned decision tree from root node to a leaf node that matches with the observed features. Each test inquires if a given bigram arises in the existing window of context. Sense of an ambiguous word is found by going through such series of test.

The characteristics of decision tree are as follows:

• A feature is represented by an internal node, where we can perform the test.

• Outcome of the test on the internal node or a feature value is represented by a branch.

• Sense is represented by each leaf node.

Each word w has the following feature vector associated with it:

• Part-Of-Speech

• Syntactic and Semantic features

• Collocation vector - normally forms by next word (+1), previous word (-1), next-to-next word (+2),previous to previous word (-2).

• Co-occurrence vector.

A. The WSD algorithm on Decision Tree

When word w is considered for disambiguation, together with its feature vector, the decision tree is traversed till leaf node by considering the previously collected training information. The sense available in the leaf node provides the correct sense.

1. Get data set and perform part of speech calculation
2. Get context having different senses of word
3. Find frequency at context
 (i.e. - p- (Negative) and +P+ (Positive))
4. Find information gain for finding entropy
 (S) = -P+log2P+-P-log2P-
5. Gain (S,A)= Entropy(S) - $\sum_{v \varepsilon DA} \frac{|SV|}{|S|} Entropy(SV)$
6. Select maximum (Entropy, Attribute ratio)

Retrieval Number: B3323078219/19©BEIESP
DOI: 10.35940/ijrte.B3323.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

IV. PROPOSED ARCHITECTURE

For this work we have taken the Health database provided by IIT Powai. Very first step carried out on this database is preprocessing on data and recommended features are extracted from it. Second step is to develop a classifier that holds the knowledge gained from the features using decision tree algorithm. In last step classifier predicts the sense of the unknown sentences or test sentences.

The process of word sense disambiguation is carried out in two parts:

(a) training part
(b) testing part

In the training part, preprocessing is done on the considered training data and features are produced. These generated features are then utilized for training the classifier based on some learning algorithm. Here we have considered decision tree algorithm. On the other hand, in testing part, preprocessing is done on testing data and features are produced. The features generated by the test data are given as input to the classifier (decision tree). Performance evaluation of the system is done by comparing the output predicted by the classifier. Therefore, the proposed architecture consists of five modules:

- Preprocessing of considered data
- Generation and Selection of feature
- Training on given features
- Testing using features
- Performance evaluation of the system

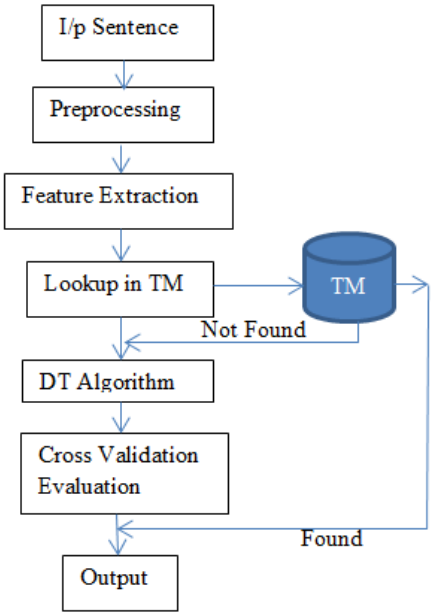

Fig.1 Architecture of the proposed system

A. Preprocessing

The process of efficiently generating features from the training data and test data by processing raw data is comes under preprocessing. After removing of present discrepancies like punctuations, irrelevant sentences, extra spacing data is converted to equivalent ASCII format. The chosen features are created semi-automatically once the data is transformed to ASCII format.

- **Data Cleaning**

The collected 15000 sentences from corpus, consists of raw data or noisy data. This kind of data contains unstructured data, some incomplete sentences, some meaningless sentences, punctuation marks which is irrelevant or not understandable by the system. So, removal of extra spaces between words, characters from other languages, different punctuation symbols from the corpus is done in the process of data cleaning. Small sentences which carry two or three words are removed from the corpus as it will give very less hint in disambiguation work. After cleaning the dataset overall 11500 sentences are available in the corpus.

- **Stop-word removal**

The most recurrently seen words in a manuscript are stop words which gives very less information for removing the ambiguity of the word. Hence, all stop words are deleted from the given context. Consider some examples of English stop words : what, or, and, am, but.

- **Lemmatization**

It is the process of finding the root word. By using morphological analysis of words as well as vocabulary lemmatization tries to do things correctly. Like in stemming we just remove the extra appended thing in the word like played will become play, forwarded will become forward to know the root word. Whereas, in lemmatization we try to find out the lemma for the given word. So, it can also find root word for “went” is “go”.

B. Feature Selection and Generation

ASCII formats created during the preprocessing phase are used to generate the feature in this phase. In this work we have used very common, extensively popular and simply extractable features. To create feature total 6 features are considered:

(i) Target word whose sense we want to find
(ii) Word’s normalized position in the given sentence
(iii) Previous word of the targeted word
(iv) Next word of the targeted word
(v) Previous to previous word of the targeted word
(vi) Next to next word of the targeted word

Targeted word and context word together forms a 5-gram window which gives the information about the context. Depending on the context targeted word may have various senses. Therefore, to find out the correct sense of the targeted word, contextual information plays the vital role [16]. Due to lack of other significant morphological features, positional feature is considered in our work. From the considered database, above mentioned features are taken automatically to produce the final input feature vector.

C. Training

In this work decision tree algorithm is considered which comes under the supervised learning algorithm category. The classifier which comes under the supervised learning algorithm build with two things: one is output labels of data and other is input feature vector.

Therefore, the resultant feature vector is created by using the output sense of the targeted word and the mentioned six features in previous section. Entries are given to the classifier after finding the sense of the targeted word. Then with the help of decision tree algorithm classifier is being trained. The approach used by decision tree is very simple and based on linear decisions.

In this work the algorithm which we are going to use to train the classifier is classification and regression tree (CART) based. By the entries of the targeted word available in the training data binary decision tree is built automatically. Based on features binary questions are asked by the CART which utilizes all the available instances of the training data.

Specialty of CART algorithm is: from available feature set it selects the most predictive feature and the top probable question to classify the training accurately. Training data is divided into two parts depending on the feature and distribution values of the features. Depending on each of the feature segmentation is done and output class is gained lastly at the leaves.

D. Testing

In testing phase CART performs the prediction by comparing features of the test case with the trained decision tree. The equivalent features are created while predicting the sense of a target word and those features get compare with the trained decision tree. It starts from the root node by making a question on each feature. Features of the targeted word derive to a leaf which gives approximate output. Features were previously produced in the feature generation phase. These features are given to trained CART and tested. Whatever equivalent approximations are given by it is noted. To evaluating the system these predictions are compared with the correct sense tags.

E. Role of Translation memory

Translation memory will store translated sentences in pair form. Like here we want to translate entered English statement into Hindi. So all entered English statement along with their translated Hindi statement will get stored in translated memory. So that when next input query will come instead of going directly to decision tree first it will check whether that statement is already available in the translation memory or not. If it is available in translation memory then no need to translate it again and we can directly select the translated sentence and can show as output. If no match is found for input query then it will go to decision tree and carry out the remaining process of translation.

F. Performance Evaluation

The performance of our system is calculated in terms of Precision, Recall and F-Score. Accuracy of the system is determined by the percentage of the prediction of the correct sense of the test sentence.

V. Result Analysis

For performing experiments we have considered 2000 sentences in corpus. Total four experiments are performed by varying the size of training and testing data.

As shown in Table 1, In Experiment 1 we have trained the system on 400 sentences and tested on 1600 sentences. As training was less accuracy or performance of the system is less. In Experiment 2 we have trained the system on 800 sentences and tested on 1200 sentences. We observe that result of second experiment is good as compare to the first one. In Experiment 3 we have trained the system on 1200 sentences and tested on 800 sentences. Result of third experiment is better as compare to the second experiment as system is getting more trained as compare to the previous one. In Experiment 4 we have trained the system on 1600 sentences and tested on 400 sentences. In this case now system is over trained and not generated the output that much correctly.

Table.1 Performance Analysis of Experiments

Exp. No.	Training	Testing	Precision	Recall	F-Score
1	400	1600	0.192	0.814	0.293
2	800	1200	0.242	0.793	0.215
3	1200	800	0.265	0.700	0.281
4	1600	400	0.145	0.853	0.167

VI. Conclusion and Future Scope

The method which gives 100% accuracy cannot exist. It totally depends upon algorithm we have considered for word sense disambiguation and quality of dataset we are using. Here we have make use of decision tree algorithm as a classifier. And for improving the speeds of the translation we have used translation memory. Performance of system is average. We can again improve the performance of the system by combining the logic of decision tree along with naïve bayes.

REFERENCES

1. Wael H. Gomaa and Aly A. Fahmy ," A Survey of Text Similarity Approaches",International Journal of Computer Applications (0975 – 8887) Volume 68–No.13, April 2013.
2. Subha Mahajan1, Rakesh Kumar2, Vibhakar Mansotra3 "Comparative Analysis of Supervised Approachesfor Word Sense Disambiguation Using TextSimilarity" International Journal for Research in Applied Science & EngineeringTechnology (IJRASET), Volume 5 Issue VI, June 2017.
3. Jumi Sarmah, Shikhar Kr. Sarma, "Decision Tree based Supervised Word SenseDisambiguation for Assamese", International Journal of Computer Applications (0975 – 8887)Volume 141 – No.1, May 2016
4. G V Garje, G K Kharate, Harshad Kulkarni, " *Transmuter: An Approach to Rule-based English to Marathi Machine Translation*", International Journal of Computer Applications (0975 – 8887) Volume 98 – No.21, July 2014.
5. Mulkalapalli Srinivas and B. Padmaja Ran, "Word *Sense Disambiguation Techniques for Indian and other Asian Languages: A Survey*", International Journal of Computer Applications (0975 – 8887), Volume 156 – No 8,p.1-7, 2016.
6. Alok Ranjan Pal and Diganta Saha, "*WORD SENSE DISAMBIGUATION: A SURVEY*", International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3, p.1-16, 2015.
7. Neetu Sharma, Dr. S. Niranjan, "OPTIMIZATION *OF WORD SENSE DISAMBIGUATION USING CLUSTERING IN WEKA*", Int.J.Computer Technology & Applications,Vol 3 (4), 1598-1604, p.1-7 July-August 2012.
8. Sarika and Dilip Kumar Sharma, "Hindi *Word Sense Disambiguation Using Cosine Similarity*", Proceedings of International Conference on ICT for Sustainable Development, p.801-808, 2016

9. Devendra K. Tayal et al, "Word *Sense Disambiguation in Hindi Language Using Hyperspace Analogue to Language and Fuzzy C-Means Clustering*", International Conference on Natural Language Processing, p.247-256, 2013.
10. Purabi Kalita and Anup Kumar Barman, "*Implementation of Walker Algorithm In Word Sense Disambiguation for Assamese language*", International Symposium on Advanced Computing and Communication (lSACC), pg.1-5, 2015.
11. Sreelakshmi Gopal and Rosna P Haroon, "Malayalam *Word Sense Disambiguation using Naïve Bayes Classifier*", IEEE,International Conference on Advances in Human Machine Interaction (HMI - 2016), R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India, Page no.1-4, March 2016.
12. Prity Bala, "Word Sense Disambiguation Using Selectional Restriction", International Journal of Scientific and Research Publications, Volume-3,Issue 4,p.1-5, 2013.
13. Singh and R. L. Ghosh, K. and Nongmeikapam, K. and Bandyopadhyay, S., "*A decision tree based word sense disambiguation system in manipuri language*", Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, page no- 17-22, July 2014.[4] Prity Bala, "Word Sense Disambiguation Using Selectional Restriction", International Journal of Scientific and Research Publications, Volume-3,Issue 4,p.1-5, 2013.
14. Andrei Mincă, Ştefan Diaconescu, "An Approach *to Knowledge-Based Word Sense Disambiguation Using Semantic Trees Built on a WordNet Lexicon Network*", 6th Conference on Speech Technology and Human-Computer Dialogue(SpeD), IEEE, p.1-6, 2011.
15. R.Navigli, "*Word sense disambiguation:a survey"*, ACM Computing Surveys(CSUR), Vol. 41 , no .2, p.-10, 2009.
16. Richard Laishram Singh1, Krishnendu Ghosh1, Kishorjit Nongmeikapam2 and Sivaji Bandyopadhyay3 "A Decision Tree Based Word Sense Disambiguation System In Manipuri Language", Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, July 2014

AUTHORS PROFILE

Sunita Rawat is working as a Assistant professor in Computer Science and Engineering department and pursuing her PhD in the field of NLP. She has publications in National and International Conferences. and have 3 publications in Scopus Indexed journal. Her research areas are Natural Language Processing, Information Retrieval and Machine learning. She has ISTE membership.