ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 1, Issue 5, June 2016

WORD SENSE DISAMBIGUATION USING DECISION TREE METHOD

[1] Rekha Thankappan , [2] Kala.M.T, [3] Dibin Joseph,
[1][2]CSE Department, GEC,Palakkad, Sreekrishnapuram, India
[3]Research Engineer, Cognicor Technologies Pvt Ltd , Infopark, kochi
[1] rekha.thankappan88@gmail.com , [2] kalsmol14@gmail.com , [3] dibin@cognicor.com

Abstract: ***Word sense disambiguation (WSD) is a technique to find the exact sense of an ambiguous word in a particular context.For example, an ambiguous word 'bank ',that has two senses 'institution' and 'river bank' in different contexts.Proposed system will predict the correct meaning of the ambiguous word in a particular context. The proposed work presents a supervised decision tree based learning approach to word sense disambiguation where a decision tree assigns a sense to an ambiguous word based on the set of positional and contextual features . The current approach proposes a classifier based on the REPTree supervised learning algorithm. As an initial step a set of positional and contextual features are inferred after preprocessing the collected database. Based on this features, tree is constructed. Finally, classifier will more precisely classify the new test entries. Weka machine learning tool is used for word sense disambiguation.***

Keywords: ***Word sense disambiguation; Classification; Decision tree; REPTree; Weka***

I. INTRODUCTION

An embedded control is done by a special purpose computer system designed to perform one or a few dedicated functions, often with real-time computing constraints. It is usually embedded as part of a complete device including hardware and mechanical parts. In contrast, a general-purpose computer, such as a personal computer, can do many different tasks depending on programming. Embedded systems control many of the common devices in use today. Embedded controllers are often the heart of an industrial control system or a process control application. The majority of computer systems in use today is embedded in other machinery, such as automobiles, telephones, appliances, and peripherals for computer systems. While some embedded systems are very sophisticated, many have minimal requirements for memory and program length, with no operating system, and low software complexity. Typical input and output devices include switches, relays, solenoids, LEDs, small or custom LCD displays, radio frequency devices, and sensors for data such as temperature, humidity, light level etc. Embedded systems usually have no keyboard, screen, disks, printers, or other recognizable I/O devices of a personal computer, and may lack human interaction devices of any kind.

Word sense disambiguation (WSD) is a technique to find the exact sense of an ambiguous word in a particular context. For example, an English word 'bank' may have different senses as "financial institution", "river side" etc. Such words with multiple senses are called ambiguous words and the process of finding the exact sense of an ambiguous word for a particular context is called Word Sense Disambiguation. A normal human being has an inborn capability to differentiate the multiple senses of an ambiguous word in a particular context, but the machines run only according to the instructions. So, different rules are fed to the system to execute a particular task.
WSD task is an intermediate task for many natural language processing (NLP) applications. It is useful in applications such as Information Retrieval, Information Extraction, Text summarization, Text classification, Machine translation or NLU etc. It has been recognized as one of the major problems in the field of NLP [2].

The objective of the present work is to develop a WSD system for selecting an exact sense of an ambiguous word in English using decision tree method.

The paper is organized as follows: related literatures are discussed in section 2. The details of the proposed work are mentioned in section 3 followed by the Conclusion in section 4. Finally acknowledgement and few important references are added.

II. LITERATURE SURVEY

The following subsections will give detailed knowledge about what supervised learning is, what decision tree method is, how REPTree is used for word sense disambiguation and a machine learning tool Weka for Word Sense Disambiguation.

A. Supervised Learning Method

Supervised learning is the machine learning task of inferring a function from a labeled training data. The training data consists of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a output value. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples [1].

B. Decision Tree as Classifier

Decision Trees are a non-parametric supervised learning method that can be used for both classification and regression. Decision trees essentially encode a set of if-then-else rules which can be used to predict target variable given data features. These if-then-else rules are formed using the training dataset with the aim to satisfy as many training data instances as possible. The formation of these rules from training data is called decision tree learning [3] [6].
Various decision tree learning algorithms have been developed and they work best in different situations. An advantage of decision trees is,that can model any type of function for classification or regression which other techniques cannot. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning[3].

Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making [6].
There are many specific decision-tree algorithms. Notable ones include:

- ID3(Iterative Dichotomiser 3)
- C4.5(successor of ID3)
- CART(Classification And Regression Tree)
- CHAID(CHi-squared Automatic Interaction Detector)
- MARS:extends decision trees to handle numerical data better
- REPTree(Reduced Error Pruning Tree)

A decision tree is used to denote classification rules in a tree structure that re-cursively 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). An example of a decision tree for WSD is described in the Figure . 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'. In the Figure 1, the tree is created and traversed and the selection of sense bank/RIVER is made. Empty value of leaf node says that no selection is available for that feature value.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 1, Issue 5, June 2016

Fig. 1. An Example of a Decision Tree

C. REP Tree

Reduces Error Pruning (REP) Tree Classifier is a fast decision tree learning algorithm and is based on the principle of computing the information gain with entropy and minimizing the error arising from variance RepTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree [5].

Basically Reduced Error Pruning Tree (REPT) is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances [4].

Decision tree is a tree formed data structure that verifies divide and rule approach. Decision tree is used for supervised learning. It is a tree structured model in which the local region is found recursively, with a set of division in a few steps. Decision tree consists of inner decision node and outer leaf. Every decision node m verifies an f m (x) test function, its discrete value is related to branches. Test function is performed in each node for an input and one of the branches is selected according to the result. This process starts in root and continues recursively until a leaf node is reached; the value written on the leaf produces the output .

Let Y and X be the discrete variables that have the values {y 1 , ..., y n } ve {x 1 , ..., x n }. In this case, entropy and conditional entropy of Y are calculated as shown in equation (11) and (12). After that,information gain of X is calculated as shown in equation (13).

D. Weka

Weka was developed at the University of Waikato in New Zealand; the name stands for Waikato Environment for Knowledge Analysis The system is written in Java and distributed under the terms of the GNU General Public License. It runs on almost any platforms. It provides a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset.

The Weka workbench is a collection of state-of-the-art machine learning algorithms and data pre processing tools. All algorithms take their input in the form of a single relational table in the ARFF format. The easiest way to use Weka is through a graphical user interface called Explorer. The Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [4].

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported).

The Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

Advantages of Weka include:

- Free availability under the GNU General Public License
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available [4].
The
corresponding unsupervised procedure is known as clustering or cluster analysis, and involves
grouping data into categories based on some measure of inherent similarity.

The classifiers in WEKA are designed to be trained to predict a single 'class 'attribute, which is the target for prediction. Some classifiers can only learn nominal classes; others can only learn numeric classes (regression problems); still others can learn both.

III. PROPOSED WORK

The current study attempts to find a solution to disambiguate between the possible senses of English words. A database is collected with sentences consisting of day-to-day senses. The database is then pre-processed and suggested features are collected out of it. Then, a classifier is developed with the knowledge obtained from the features using a machine learning algorithm. The classifier finally predicts the sense of the test sentences.
The word sense disambiguation system works in two phases: (i) training phase and (ii) testing phase. In the training phase, the training data is first preprocessed and features are generated. The features are used then to train the classifier based on any learning algorithm. On the contrary, in the testing phase, the test data is preprocessed and features are generated. The features of the test data are fed to the classifier and it predicts the output based on the training dataset. The current study, as it is in the preliminary attempt, uses a set of very common, widely popular and easily extractable features. A total of 11features are taken to build feature:

1) Ambiguous word
2) POS of Ambiguous word
3) Position of the Ambiguous word
4) Previous word
5)POS of Previous word
6) Previous-to-previous word
5) POS of Previous-to-previous word
8) Next word
9) POS of Next word
10)Next-to-next word
11) POS of Next-to-next word

Here, part-of-speech(POS) tagging is used for tagging the words . The process of assigning part-of-speech to each word in a sentence is called POS tagging.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 1, Issue 5, June 2016

A. Data Flow Diagram

The below figure 2 shows the working flow of the proposed work,

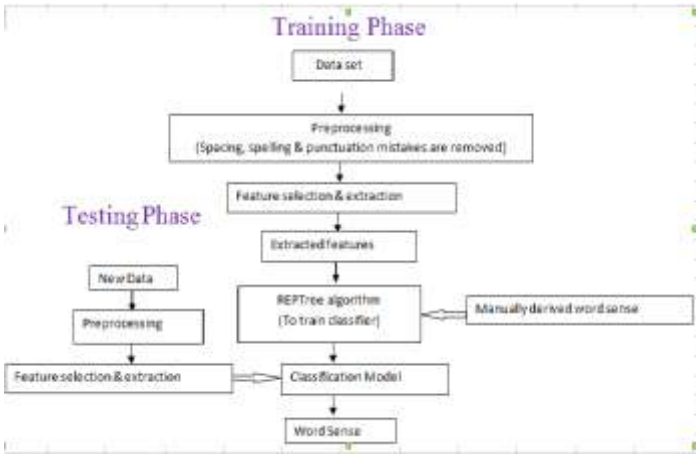

Fig. 2. Flow Diagram

B. Experimental Setup And Observations

The Language used in the experiment is python. For the experiment, python 2.7.3 has used for python coding.Weka-3-6-13 has installed for GUI Operations. Python-weka-wrapper and jdk.8.0 73 has installed for Command line Operations. In WSD system .ARFF data file format was used for training and testing .This proposed work presents decision Tree as Classifier and REPTree as machine learning algorithm which could be imported from the Weka Tool.

The performance of WSD system is evaluated using the measure precision .Precision is defined as a measure of the proportion of correct items that the system got right to the total number of items.

precision = tp / (tp +fp)
Here tp is the true positive and fp is the false positive.

- True positive : The percentage of correct items that the system got right.
- False positive : The percentage of correct items that the system got wrong.

Being trained on 574 data instances for 562 words for 25 ambiguous words and tested by 100 data instances in .ARFF data format gives about 86.00% accuracy. More accuracy can be obtained by increasing the number of training instances and update all feature values into the attribute section. The performance of the proposed model is discussed in TABLE 1.

TABLE 1: OBJECTIVE EVALUATION OF REPTREE

Predicted Sense	Correctly Predicted Sense	Accuracy
100	86	86.00%

BASED WORD SENSE DISAMBIGUATION MODEL

IV. CONCLUSION

The current study attempts to find the most appropriate meaning for an ambiguous word in English, based on the context in which it occurs.Supervised Decision Tree Machine Learning Method is used in the proposed WSD system. It is simple to interpret and understand. It requires little data preparations, and able to have value even with little hard data. It is able to process large amount of data in a short time.Weka implementation of the REPTree decision tree learner is used to generate the model files to predict word senses. A set of positional and contextual features are suggested for developing the Word Sense Disambiguation system.

The proposed work gives knowledge about the importance of a sense-tagged corpus for a supervised WSD system. It is important for a system to have sufficient training examples to attain higher WSD accuracies.

ACKNOWLEDGMENT

I am extremely grateful to Dr. Ajeesh Ramanujan, M.Tech Co-ordinator and Asst.Prof. Kala mol, Internal Guide of project, Govt. Engineering College Sreekrishnapuram, for their sincere guidance, inspiration and right direction throughout the project and for providing and availing me all the required facilities for undertaking the project in a systematic way. Gratitude is extended to all teaching and non teaching staff of Department of Computer Science and Engineering, Govt. Engineering College Sreekrishnapuram for the sincere directions imparted and the cooperation in connection with the project. Gratitude is extended to Mr.Vishal Yadav, System Design Lead (AI), CogniCor Technologies at Cochin for the constant support and the cooperation throughout the project. I am thankful to Mr.Gopalakrishnan G, NLP Engineer, Mr.Dibin Joseph Research Engineer and Jyothis K S, Indian Operations Lead CogniCor Technologies at Cochin for his encouragement and sincere guidance.

REFERENCES

[1] Alok Ranjan Pal and Diganta Saha," Word Sense Disambiguation:A Survey", International Journal of Control Theory and Computer Modeling (IJCTCM) ,Vol.5,No.3, 2015

[2] Richard Laishram Singh,Krishnendu Ghosh, Kishorjit Nongmeikapam, And Sivaji Bandyopadhyay. 2014 July. "A Decision Tree Based Word Sense Disambiguation System In Manipuri Language. Advanced Computing", An International Journal (ACIJ),Vol.5,No.4,July 2014

[3] Heng Low Wee,"Word Sense Prediction Using Decision Trees",Undergraduate Research Opportunity Program (UROP), Project Report,2010/11

[4] Sushilkumar Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News", International Journal of Innovative Science, Engineering & Technology(IJISET),Vol. 2 ,Issue 2,February 2015

[5] C. Lakshmi Devasena,"Proficiency Comparison Ofladtree and Reptree Classifiers for Credit Risk Forecast", International Journal on Computational Sciences & Applications(IJCSA),Vol.5, No.1, February 2015

[6] Saif Mohammad ,"Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation",University of Minnesota,August 2003