Volume 6, Issue 2, February 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Word Sense Disambiguation using Maxnet Approach for Hindi Language

[1]Madhuri Bansal*, [2]Dr. Pratistha Mathur

[1] M.tech Scholar, Department of Computer Science and Engneering., Apaji Institute Banasthali Vidyapith Jaipur, Rajasthan, India

[2] Professor of Computer Sciences, Apaji Institute Banasthali Vidyapith Jaipur, Rajasthan, India

***Abstract*— Word sense Disambiguation is a central topic in the study of natural language processing and has long been object of research in a wide range of disciplines. At present, Disambiguation Of ambiguous words is one of the most challenging and complex task to be handled. This paper completely emphasis on word sense Disambiguation for Hindi Language. There are various methodologies for disambiguation ambiguous words but as far as my knowledge most of them have been implemented for English language. Maxnet is one of such algorithm which has not been utilized yet for Hindi language. So, based on the concept of Maxnet Approach for Word sense Disambiguation we have developed a model that performs Word sense Disambiguation for Hindi for Hindi Language. In this paper we use Maxnet Classifier to disambiguate ambiguous Hindi words with part-of-speech 'noun'. The system also uses Sense Annotated Hindi Corpus as training data. This paper demonstrates an experiment on 5 to 10 sentences of different ambiguous words.***

***Keywords*— Natural Language processing, Hindi Language, Maxnet Classifier, Sense annotated Hindi Corpus, Supervised Approach.***

## I. INTRODUCTION

Word Sense Disambiguation (WSD) is the task of finding the appropriate sense of a word used in a given sentence, when the word may have more than one sense [1]. For eg. **जंगल मे शेर** Here **शेर** word is ambiguous word contains two meaning as ' बिल्ली की जाति का एक बहुत बड़ा और भयंकर, हिंसक पशु' (or) गज़ल के दो चरण. In this sentence mean of **शेर** word is **'**बिल्ली की जाति का एक बहुत बड़ा और भयंकर, हिंसक पशु'

Another Eg. **शेर सुनाओ** In this sentence mean of **शेर** word is **'**गज़ल के दो चरण'.

Word sense disambiguation is not an isolated system by itself, but can be fitted with other important tasks such as, information retrieval, machine translation, speech processing, parts-of-speech tagging and text processing [3,4]. WSD is a task of classification: word senses are the classes, context provides the evidence, and each occurrence is assigned to one or more of its possible classes based on evidence [1]. Sense Disambiguation [2] is an 'intermediate task' which is not an end itself, but rather is necessary at one level or another to accomplish most NLP tasks. Sense Disambiguation involves Sense Knowledge. Sense Knowledge can be represented by a vector, called a sense knowledge vector (sense ID, features), where features can be either symbolic or empirical. The word to be sense tagged always appears in a context. Context can be represented by a vector, called a context vector (word, features). Thus, we can disambiguate word sense by matching a sense knowledge vector and a context vector.

The goal of the present work is to develop a system which can disambiguate between the different senses of a polysemous word in Hindi language. In European languages, efficient and automatic WSD systems are present. In Indian languages, such systems are mostly rule-based due to lack of standardized database and presence of the proper knowledge acquisition tools. So far, negligible amount of work has been reported in Hindi language. That motivates the present attempt to collect a suitable database consisting of every possible contexts and senses, used in day-to-day life. We have used the Maxnet Classifier to mark the sentence structure. Besides, we have an standard Sense Annotated Hindi Corpus to capture the actual sense a word generates in a normal Hindi sentence.

 

## II. RELATED WORK

At yet for word Sense Disambiguation Maxnet Classifier have not been used But for Another work Maxnet is used.

- Lachlan L. H. Andrew, Krister Jacobsson, Steven H. Low, Martin Suchara, Ryan Witt, Bartek P. Wydrowski The MaxNet TCP network congestion control protocol [5].

## III. AMBIGUITY FOR HINDI LANGUAGE

Hindi Language is official language of India. It is written from left to right and spaces between words. It is syllabic alphabet and writ-ten in circular shape. It has sentence boundary mark. It is a free-word-order language, which usually follows the subject-object-verb (SOV) order However; English Language has a rigid subject-verb-object (SVO) order. In table 1 show some examples of Hindi ambiguous nouns and their senses.

Table I Some Ambiguous Nouns and Their Senses

| Ambiguous words | No. of Senses | Sense 1 | Sense 2 | Sense 3 |
|---|---|---|---|---|
| कदम | 3 | उपाय | चरण | पग |
| कमान | 3 | धनुष | लगाम | एक विशेस आदेश |
| चारा | 2 | घास-भूसा | उपाय | - |
| कुंभ | 3 | मिट्टी का घड़ा | राशि | महा कुंभ का मेला |
| फल | 3 | परिणाम | खाने का फल | तीर का भाग |

## IV. SENSE ANNOTATED HINDI CORPUS

The Sense Annotated Hindi corpus that is used in this work is developed under the TDIL (Technology Development for the Indian Languages) project, Govt. of India. The Sense Annotated Hindi Corpus for lexical sample Word Sense Disambiguation task consisting of 60 polysemous Hindi nouns. The total number of instances in the corpus are 7506 and total words in the corpus are 381875. For Example हार word contains two senses 1st असफलता and 2nd माला then there each sense related individually sentences are in one text file. This corpus is exhaustively used to extract sentences of a particular word required for our system as well as for validating the senses evoked by the word used in the sentences.

## V. MAXNET FOR CLASSIFICATION

A classifier model based on maximum entropy modelling framework. This framework considers all of the probability distributions that are empirically consistent with the training data; and chooses the distribution with the highest entropy. A probability distribution is "empirically consistent" with a set of training data if its estimated frequency with which a class and a feature vector value co-occur is equal to the actual frequency in the data.

A maximum entropy classifier (also known as a "conditional exponential classifier").This classifier is parameterized by a set of "weights", which are used to combine the joint-features

that are generated from a feature set by an "encoding". In particular, the encoding maps each ``(feature set, label)`` pair to a vector. The probability of each label is then computed using the following equation::

$$prob(fs|label) = \frac{dotprod(weights, encode(fs,label))}{sum(dotprod(weights, encode(fs,l)) \text{ for } l \text{ in labels})}$$

Where ``dotprod`` is the dot product::

dotprod(a,b) = sum(x*y for (x,y) in zip(a,b))

Train a new Maxnet classifier based on the given corpus of training samples. This classifier will have its weights chosen to maximize entropy while remaining empirically consistent with the training corpus.

## VI. OVERVIEW OF PROPOSED ALGORITHM

Step 1: User enter the Query and splitting of complete sentence will be in this phase
Step 2: If user enter query is single ambiguous word then return all senses form sense Annotated corpus
Step 3: If user enter query is sentence then after splitting Ambiguous word will detect by system.
Step 4: Retrieve and collect data corresponding ambiguous words senses from sense annotated corpus
Step 5: Tokenization Stop words Removal and Labelling of data will be occur in this step
Step 6: Feature are extracted because Processing on complete data will be very time consuming so extract the features
Step 7: Training of data done by Maxnet Approach.

 

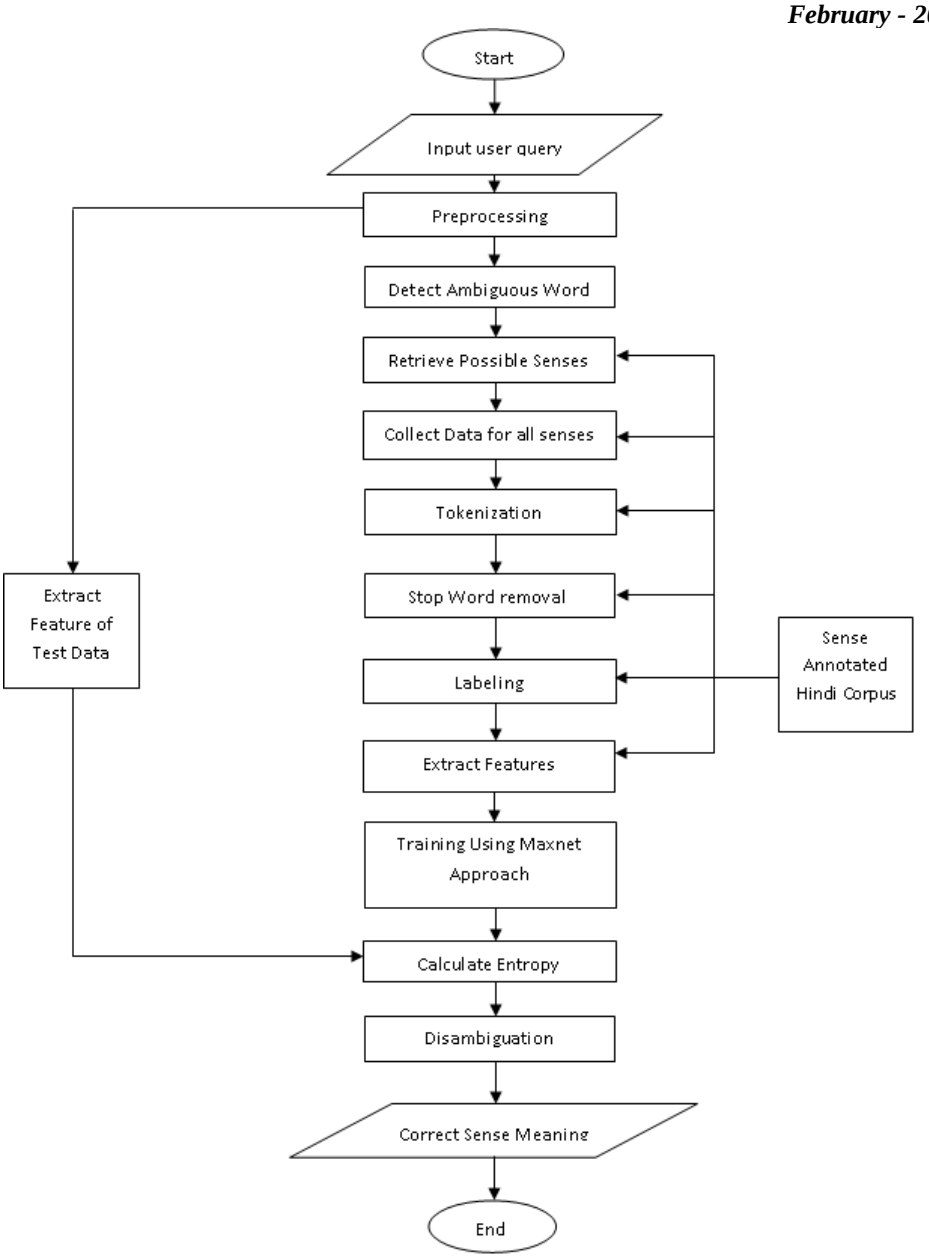


Figure 2 shows complete Word sense Disambiguation using Maxnet Approach

Step 8: Classification of sense will be based on extract feature of test data and extract feature of training data highest matching of training sense will be result sense. Our proposed algorithm is shown in the following figure 3.

1) Pre-processing
   a. Remove stop words from input sentence
   b. Feature extract after input sentence
2) Retrieve Possible Senses
   -Lookup possible sense meanings of the ambiguous word from the Sense Annotated corpus
3) Train a new Maxnet classifier based on the given corpus of training samples. This classifier will have its weights chosen to maximize entropy while remaining empirically consistent with the training corpus.
4) Choose s' = argmax score(si)

Figure 3: Maxnet Algorithm for Hindi WSD

## VII. IMPLEMENTATION OF THE SYSTEM

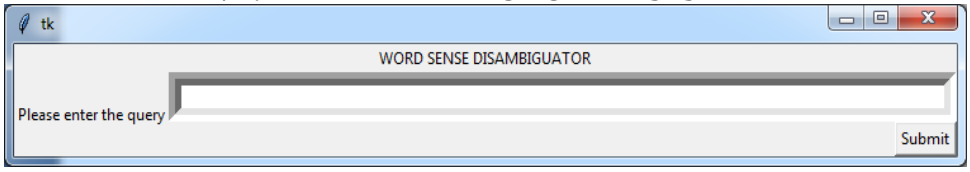


Figure 4: User Interface

 

**Main Screen:** The main screen consists of Textbox and one submit button. User enter the query in Textbox clicks on submit labelled button, then sense of ambiguous word will come according neighbouring words.

**User enter the query:** To find the sense of ambiguous word

**And System Detect the ambiguous word and give the sense**: Click on Submit Labelled button to find the sense using Decision Tree Approach for —"**कुंभ**" word

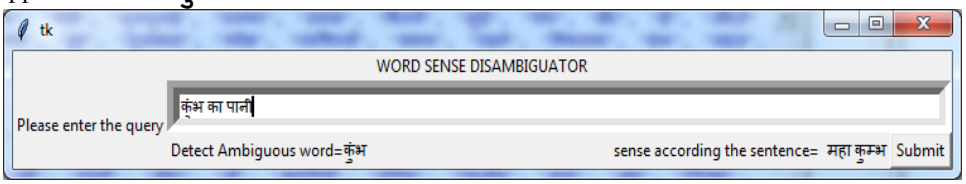


Figure 5: User enter Ambiguous Query and get Result Using Decision Tree Approach

## VIII. EVALUATION OF THE SYSTEM

$$Accuracy = \frac{Correct\ Answers\ Provided}{Answers\ Provided}$$

Table II. Data Set of Words and Results of Maxnet Classifier

| Ambiguous Words | No. of Sentences | Correct Result using feature [-1,1] | Correct Result using Feature [-2,-1,1,2] | Accuracy from [-1,1] | Accuracy From[-2,-1,1,2] |
|---|---|---|---|---|---|
| कुंभ | 12 | 8 | 10 | 66% | 83% |
| हार | 5 | 4 | 5 | 80% | 100% |
| अशोक | 5 | 3 | 4 | 60% | 80% |
| उत्तर | 10 | 7 | 8 | 70% | 80% |
| जेठ | 5 | 3 | 4 | 60% | 80% |
| दाम | 8 | 5 | 7 | 62% | 87% |
| फल | 8 | 3 | 5 | 37% | 62% |
| शेर | 11 | 8 | 8 | 72% | 72% |
| सोना | 6 | 3 | 4 | 50% | 66% |
| हल | 8 | 5 | 7 | 62% | 87% |

## IX. CONCLUSION

Table II. Below shows the final results of accuracy for Maxnet approach. By looking at the results that we came across there are few words which are providing accurate results.

## REFERENCES

[1] Kumari Sabnam and Singh Paramjit, "Genetic Algorithm based Hindi Word Sense Disambiguation", International Journal of Computer Science and Mobile Computing (IJCSMC), Volume 2, Issue 5, May 2013.

[2] Hephaestus Books, "Articles on Word Sense Disambiguation", 29-Aug-2011, online, Available:books. google.co.in/books?isbn=1242967184.

[3] M.Sinha, M.K.Reddy, P.Bhattacharyya, P.Pandey and L.Kashyap, "Word Sense Disambiguation," in proceedings of International Journal of Computer Applications (IJCA),2010, vol. 5.9, pp. 25-32.

[4] R.Mihalcea and D. Moldovan, "A Method for Disambiguating Word Senses in a Large Corpus," in proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing, Computers and the Humanities, 1992, volume 26, no. 5-6,pp. 415-439.

[5] Lachlan L. H. Andrew, Krister Jacobsson, Steven H. Low, Martin Suchara, Ryan Witt, Bartek P. Wydrowski, "MaxNet: Theory and Implementation," .

[6] S Kumari "Optimized Word Sense Disambiguation in Hindi using Genetic Algorithm", International Journal of Research in Computer and Communication Technology(IJRCCT),vol2,Issue 7,July-2013

[7] http://www.nltk.org/nltk_data/

 