International Journal of Computing and Applications | **ISSN: 0973-5704**
Volume 13, Number 1, (January-June 2018), pp 75-81

RESOLVING ISSUES IN WORD SENSE DISAMBIGUATION FROM HINDI-ENGLISH LANGUAGES

Shachi Mall[1] and Umesh Chandra Jaiswal[2]

[1-2]Department of Computer Science & Engineering, Madan Mohan Malaviya University of Technology, Gorakhpur, India.
Email: [1]shachimall@gmail.com, [2]ucjjaiswal@yahoo.co.in

Abstract: This paper is based on the problem of Word Sense Disambiguation for Hindi Language which is categorized as Natural Language Processing under broad area of Artificial Intelligence. Natural Language is required to establish communication either verbal or written between two or more persons. The persons involved in communication must agree upon the natural language selected for communication. Hindi is a national language in India which is not understandable by outside world. Therefore, there is a dire need of translation systems. There are various websites which provide the facility of translation such as Google Translator and Babefish Translator but these translators fail to resolve polysemy words in Hindi sentences. The paper presents the method that discusses and automatically decides the correct meaning of an ambiguous word based on the surrounding context in which it appears. A methodology is based Rule based and machine learning techniques such as supervised, unsupervised and domain specific sense with the information of Word Net tool. We modify and develop Lesk algorithm based on the Parsing. Parsing is an extension of our previous works. These combinations of features resolve the sense of a word in a context. Word Net tool use as a dictionary which contain words and their meaning semi-supervised approach use this information for disambiguation. Unsupervised approach cluster the each words with unique frequency id and match the id with domain name. The system generated result of Word Sense Disambiguation is compared with the website Google Translator.

Keywords: Word Sense Disambiguation; semi supervised; unsupervised; domain specific sense.

1. INTRODUCTION

In India mostly information is presented in English language therefore we need translation systems. There are various websites which provides the facility of translator such as Google Translator and Babefish translator but still they are fail to resolve the correct sense of a word in the given input sentence. This can be clear with the given example in Table 1 and Table 2. To remove the ambiguous word which produces multiple senses in given input sentence the paper discusses the implementation of machine translation system which produce correct translation as shown in table 3 ,section 4.6 discusses the experimental setup of word sense disambiguation and snapshot 4(b) shows the output result. Machine Translation plays a vital role it received a jolt of new activity and much more visibility in Natural language processing [1]. It is seemingly well-defined task of converting Hindi language (source language) to English language (Target language) while preserving its meaning. Machine Translation systems are divided into different modules which will give the desire result. We will discuss about each of them one by one:

- *Tokenize Sentence and Word [3] in a given sentence:* Hindi Sentence is break at word level delimited by punctuation symbol purnviram "□" or question mark "?" and words are delimited by whitespace between two words.
- *Morphological analyses:* The minimal parts of words that deliver aspects of meaning to them are called morphemes.
- *Lexemes:* Forms word is expressed in linguistic form in the given context. The concept is to set the alternatives form which can express it.

- *Parts of Speech Tagging[2]:* It tag the token or word with their related Parts of Speech like Noun, Pronoun, verb etc
- *Chunk:* Chunk determines the beginning of phrase and inside the phrase in the sentence for example NP –Noun Phrase.
- *Parsing:* Parsing means break the Hindi sentence to analyze the syntactic structure of the sentence [4].
- *Word sense disambiguation:* It is peradventure the most decisive task in the field of machine translation, either supervised and unsupervised approach is used for disambiguation.
- *Ambiguous lexemes:* Hindi word which has two separate lexemes with distinct and unrelated meaning for example: **उस कबूतर के पर क़तर दो** Here Hindi word **पर** has two synsets meaning. Synsets for adjective grammar **पर** means other and Synsets for noun grammar **पर** means wing. To find out Synsets meaning we have used English Word Net [5] and Hindi Word Net [6]. Each synset consists of a set of synonyms and ontological categories.

Table 1
Google Translator

S.No.	*Google Translator from Hindi language to English language*
Input sentence	**उस कबूतर के पर क़तर दो**
Output sentence	Two doves on the Qatar

Table 2
Babefish Translator

S. No.	*Babefish Translator from Hindi language to English language*
Input sentence	**उस कबूतर के पर क़तर दो**
Output sentence	The pigeon at Qatar two

Table 3
Our system generated translation

S.No.	*Our system generated translation from Hindi language to English language*
Input sentence	**उस कबूतर के पर क़तर दो**
Output sentence	Slice or cut or chop of the wings of that pigeon

2. RELATED WORK

A number of Indian researchers have carried out their work related to machine translation for Indian languages. Less work has been carried out in Hindi language. In Hindi language parts of speech and word sense disambiguate is an important task to produce the correct translation from Hindi to English Machine Translation A number of approaches used such as corpus based and [4] knowledge based it is based on dictionary. Various technique have been proposed through researchers to resolve the issues related to Graph based [5] the accuracy result was evaluated 10 % this method is compared to AGU and accuracy is 54.9 % in sensaval-3 and 60.2 % in sensaval-2 dataset. Another approach based on Dependency parsing [6]. The algorithm finds sense inventories using from Word Net tools. Another attempt [7] presents a multilingual joint approach for Word Sense Disambiguation on graph based. [8] Another attempt works on Specific Iterative which is based on specific domains. The accuracy result was evaluated 65%. They based on domain 23information [9]. The drawback of that algorithm was it can disambiguate a word provided it has only one sense per domain. The work on English French Cross-lingual Word Sense Disambiguation [10] where the task is to find the best French translation for a target English word depending on the context in which it is used their approach relies on identifying the nearest neighbors of the test sentence from the training data using a pairwise similarity measure. The average performance of our system was less than the baseline by around 3%, it outperformed the baseline system for 12 out of the 20 nouns. [11] modified Lesk's algorithm, context window approach the algorithm increased the context window.

A. Identification of Research Gap and Problem

There are many unresolved issues in machine translation for Indian languages such as:

- Morphological analyses faces problem in productivity and creativity in languages, word that are not licensed than it will remain unparsed. This is known as unknown word.

- Construction of electronic dictionary.
- Grammatical tagging corpora and Chunk the sentence.
- Resolve the word sense disambiguation. The major drawback is the problem of scale.

B. Solution of Research Gap and Problem

- Construction of electronic dictionary which understood the data structure and directly obtains result. This can be done by lookup operation.
- The system use shallow transfer approach which tokenize the text into the token (HTML and white space) as shown in section 4.1 parts of speech is description of Noun, Verb, Article etc. Grammatical Tagging are evaluated by comparing with gold standards test set and the accuracy is around 92.09%
- Machine learning approaches to sense disambiguation make it possible to automatically create sense disambiguation. Supervised approach is used for collections of texts annotated with their correct senses to train classifiers. Unsupervised based is used for representation of word senses from unannotated texts. To overcome with the problem of scaling the papers introduce scaling approach to deal with all ambiguous words in Hindi language. Disambiguation focused on the machine readable dictionaries. In this approach all the sense definitions of the word to be disambiguated are retrieved from the dictionary. Each of these senses is then compared to the dictionary definitions of all the remaining words in the context. The sense with the highest overlap with these context words is chosen as the correct sense. The modified Lesk's algorithm make more concrete.

3. METHODOLOGY

Ambiguity can be resolved with syntactic information is based on Parsing and Word sense disambiguation. Word sense disambiguation will automatically decide the correct meaning of a polysemy word depend on the surrounding context in a given sentence. We used machine learning techniques such as semi-supervised, unsupervised, overlap based method and domain specific sense with the information of Word Net tool. This approach and method will resolve the problem of Word sense disambiguation for Hindi language. Hindi words have multiple meanings which we call senses. Senses fetches from the Hindi Word Net. The Hindi Word Net is a tool that contains many library functions such as Synset, Gloss, Ontology, Hyponymy, Hypernymy, Meronymy and Holonymy. This all elements help to shows the relationship between lexical and semantic. For using all the functionalities of the libraries we load Pickle module is used to ahead and serialize our classifier object, so that all we need to do is load that file in real quick., once this loaded we can use all the Word Net operation. Figure 1 explains the overall process of Word Sense Disambiguation. Input Hindi sentence that contain polysemy word. Polysemy word is multiple tags with the help of parsing and overlap based method. This method used two approach supervised and semi-supervised method. Hindi words are containing polysemy words which have multiple meaning depending on the context in which they occur. Word sense disambiguation works on the following principal:

- Homonymy
- Polysemy
- Categorical ambiguity

Homonymy indicates that the words share the same spelling, but the meaning are quite disparate. Each homonymous partition however, may contain finer sense nuances that could be assigned to the word depending on the context and this phenomenon is called polysemy for example Hindi word "स 1" hold English meaning gold and स 2 hold English meaning sleep. Categorical ambiguity can be resolved with syntactic information. Word Sense ambiguities [16] disambiguate the senses of word with the meaning of multi-sense words using Distributed Domain approach

by analyzing the context in which sense the multi-sense words and produce correct output. In Hindi language contains such multi-sense words in its corpus. This is based on supervised and unsupervised approach [13]. Supervised approach [14] is used to identify the correct meanings in multi-sense words in Hindi languages. The Distributed Domain approach are used disambiguate the senses of word in the sentence context based on defined learning set this can be created manually unable to generate fixed rules for specific system. Therefore predicted meanings of an ambiguous word are in given context. Supervised learning derives partial predicted result, if the learning set does not contain sufficient information then sense the ambiguous word. It shows the result, only if there is information in the predefined database. Supervised approach consist of a machine learning classifier trained on various features extracted for words that have been manually disambiguated in a given corpus and the application of resulting models to disambiguate words in unseen test sets. Support vector classifiers are used to train word sense disambiguation models. There are following features are used:

- Lexical context
- Part of speech
- Bag of word context
- Local collocations
- Syntactic relations
- Topic features
- Voice of the sentence
- Presence of subject/object
- Sentential complement
- Preposition

Unsupervised approaches use dictionary for learning This can be done by using Word Net [15] [16] [17][18] as lexical database is an important resource to find the correct meanings in multi-sense words in Hindi languages. Progress in word sense disambiguation is stymied by the dearth of labeled training data to train a classifier for every sense of each word in a given language.

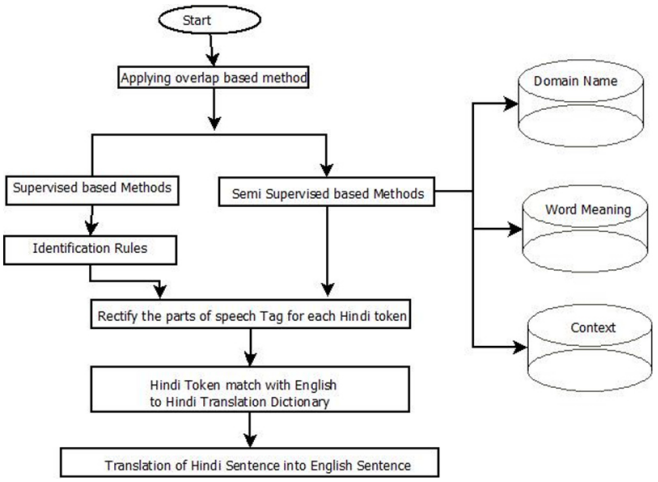


Figure 1: WSD System Model

This paper is extension of my previous paper [2] in which morphological analyzers, parts of speech tagging, chunk and parsing modules is completed. Using these modules we develop word sense disambiguation and translation of Hindi language to English language. This paper will resolved the issues in word sense disambiguation for Hindi language by using Word Net and Modified Lesk algorithm. For translation from Hindi to English language we use English Word Net.

1. Word Sense Disambiguation

Algorithm Modified Lesk Algorithm

Input: Text with only meaningful words

Output: Actual sense of ambiguous words

1. Loop Start for all dictionary definition of the ambiguous word
2. Ambiguous word is selected
3. Each word is selected from preliminary input texts.
4. Gloss of ambiguous word is obtained from typical Word Net.
5. Intersection is performed between the meaningful words from the input text and the glosses of the ambiguous word.
6. Loop End
7. If the counter value is mismatched with all other values, then associated sense is considered as the disambiguated sense.

8. Else, Bag-of-Words fails to disambiguate the sense.
9. If occurrence of an unmatched word in anticipated database having a particular sense crosses the threshold value, then the word is moved to the related bag of words database.
10. Stop.

Modified Lesk approach [17] selects a phrase from the sentence containing an ambiguous word. Now gloss of keyword is only selected in a given Hindi sentence instead of selection of all words. Number of common words is being calculated between specific sentence and each dictionary based definitions of particular keyword.

A. Precision, Recall and F-score for Word Sense Disambiguation

Precision (P) is the ratio of "matched target words based on human decision" and "number of instances responded by the system based on the particular words". Recall value (R) is the ratio of "number of target words for which the answer matches with the human decided answer" and "total number of target words in the dataset".

F-Measure is evaluated as "(2*P*R/(P+R))" based on the calculation of Precision and Recall value. Different types of datasets are being considered in our experimentation to exhibit the superiority of our proposed design.

After setting the experiments, we test the accuracy of the system generated result the experiments held using the gold data from internet. Gold data contain multi-sense words in Hind sentences and translated in English sentences with correct translation with Word Sense Disambiguation. The test data consists of 200 different Hindi sentences each containing multi-sense words. In experiment, we found that the multi-sense words in 180 sentences out of 200 are correctly disambiguated. This shows the accuracy of the system with conventional sample Hindi sentences Net found to be 90 %. Table IV shows the result of Word Sense Disambiguation.

Table 4
Result of Word Sense Disambiguation

Test No	*Number of polysemy words that are correctly Disambiguated*	*Number of polysemy words that are not correctly Disambiguated*	*Accuracy result (%)*
1	200	20	90%

4. CONCLUSION

The paper studies with the implementation of Hindi language translated in English language. The Morphological analyzer as shown in Figure 3, parts of speech and chunk as shown in Figure 4 algorithm resolve the word sense disambiguation for Hindi words. The accuracy result is shown in our previous paper [20] was evaluated for 1657 tokens result for Chunk Accuracy: 81.23% to evaluation for Parts of speech tagging is done by Conditional Random field. The confusion matrix is created to calculate the accuracy of Parts of speech tagging result is Tagging 92.09% this paper discusses some identification rules for rectification for parts of speech tagger. For Word Sense Disambiguation we modify the Lesk algorithm which produces better result. We evaluate for Word Sense Disambiguation for 200 Hindi words among that 180 sentences are correctly disambiguated. The accuracy of the system was 90%. The result is also compared by Google Translator [21] we input Hindi sentence **उस कबूतर के पर क़तर दो** as shown in Figure 2 Here word **पर** in the given Hindi sentence is polysemy word which has two meaning **wing** and **on**. We compare our output result with translating website Google translator we input same Hindi sentence as shown in Figure 9. Here Hindi word **पर** s translated as **on** in the English sentence but correct translation is **wing** for the given sentence. Here we can see that Google translator is failed to translate correctly but our system generates correct translation for the given Hindi word **पर** is translated as **wing** in the English sentence e as shown in the Figure 5. Further, we will work on word alignment

Acknowledgment

We would like to acknowledge III-Hyderabad for organizing workshop on Advance school on Natural

language Processing (ISNLP-2015) sponsored by Google. This work is supported in part of parts of speech tagging and chunk.

References

[1] Radev, D., & Lapata, M. (2008). Natural Language Processing and the Web. IEEE Intelligent Systems, (5), 16-17.

[2] Bellegarda, J. R. (2010). Part-of-Speech tagging by latent analogy. Selected Topics in Signal Processing, IEEE Journal of, 4(6), 985-993.

[3] Mall, S., & Jaiswal, U. C. (2015, October). Innovative algorithms for Parts of Speech Tagging in hindi-english machine translation language. In Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on (pp. 709-714). IEEE

[4] Agarwal, M., & Bajpai, J. (2014, August). Correlation based Word Sense Disambiguation. In Contemporary Computing (IC3), 2014 Seventh International Conference on (pp. 382-386). IEEE.

[5] Hessami, E., Mahmoudi, F., & Jadidinejad, A. H. (2011). Unsupervised weighted graph for Word Sense Disambiguation. In 2011 World Congress on Information and Communication Technologies.

[6] Li, Z., Zhang, M., Che, W., Liu, T., Chen, W., & Li, H. (2011, July). Joint models for Chinese POS tagging and dependency parsing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1180-1191). Association for Computational Linguistics

[7] Navigli, R., & Ponzetto, S. P. (2012, July). Joining forces pays off: Multilingual joint word sense disambiguation. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (pp. 1399-1410). Association for Computational Linguistics.

[8] Khapra, M., Bhattacharyya, P., Chauhan, S., Nair, S., & Sharma, A. (2008, December). Domain specific iterative word sense disambiguation in a multilingual setting. In Proceedings of International Conference on NLP (ICON 2008), Pune, India

[9] Kolte, S. G., & Bhirud, S. G. (2008, July). Word sense disambiguation using Word Net domains. In Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on (pp. 1187-1191). IEEE.

[10] Mahapatra, L., Mohan, M., Khapra, M. M., & Bhattacharyya, P. (2010, July). OWNS: Cross-lingual word sense disambiguation using weighted overlap counts and Word Net based similarity measures. In Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 138-141). Association for Computational Linguistics.

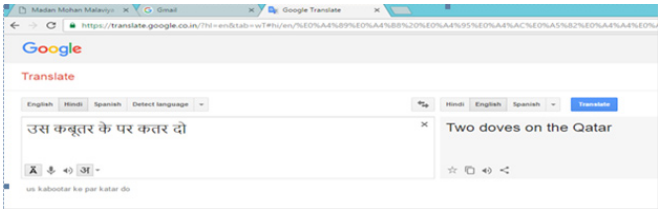
[11] Sawhney, R., & Kaur, A. (2014, September). A modified technique for Word Sense Disambiguation using Lesk algorithm in Hindi language. In Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on (pp. 2745-2749). IEEE.

[12] Gupta, R., Goyal, P., & Diwakar, S. (2010, September). Transliteration among Indian Languages using WX Notation. In KONVENS (pp. 147-150).

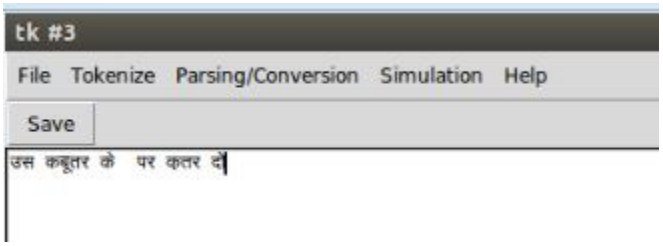
[13] Edmonds, P., & Agirre, E. (2006). Word Sense Disambiguation: Algorithms And Applications.

[14] Mallapragada, P. K., Jin, R., Jain, A. K., & Liu, Y. (2009). Semiboost: Boosting for semi-supervised learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(11), 2000-2014.

[15] Redkar, H. H., Bhingardive, S. B., Kanojia, D., & Bhattacharyya, P. (2015, March). World Word Net Database Structure: An Efficient



Snapshot 1: Output of Google Translated text from Hindi to English language



Snapshot 2: Input Hindi sentence

Snapshot 3: Parse the Hindi sentence

Snapshot 4: Output of the Hindi text with Word sense disambiguate algorithm