Word Sense Disambiguation: Enhanced Lesk Approach in Punjabi Language

Jagbir Singh
Centre for Development of advance Computing
Mohali

Iqbal Singh
Centre for Development of advance Computing
Mohali

ABSTRACT

India is a country with diverse languages. Many a times, words are spoken and used which have more than one meaning. While interaction between humans, the determination of the correct meaning of the ambiguous word (word with multiple meaning) can easily be judged by referring the context of the communication. But for a computer to judge the best and the correct meaning of the word, training needs to be provided to the system. Word Sense Disambiguation (WSD) is technique used to disambiguate the ambiguous words (single word with multiple meaning). Our work deals with analyzing the correct meaning of the ambiguous word(s) in Punjabi language. Not much work has been done in this field which deals with the Punjabi language. A text with multiple senses in natural language is open problems of Natural Language Processing (NLP) which can be resolving using WSD. The Supervised learning methodology is used for this purpose which is the conventional approaches to WSD. The semantic lexicon for the various languages of India i.e., Indo WordNet has been used to obtain the sense definition of the Punjabi language. An enhanced Lesk approach is used to analyze the correct sense of the ambiguous word which uses the concept of dynamic context window. The proposed algorithm works on one assumption that the words on the left and the right of the context window should be of the same theme in its neighborhood. Finally, instance and precision is obtained which shows that larger the size of context window, more appropriately the correct sense of the word can be determined.

Keywords

Word sense disambiguation, Lesk, Punjabi Language, Natural language processing

1. INTRODUCTION

In Natural languages some words have different sense in different context. This is known as the ambiguity in human languages. Humans are master in identify the complexity of spoken language; this is the ability which separates them from animals. But for computers it is hard to understand the natural languages. This problem defined as Natural Language Processing (NLP) which is the area under Artificial Intelligence and Linguistics that deals with the connections between computers and natural (human) languages.

For example, the English noun "crane" has two meaning: one is "a large, tall machine used for moving heavy objects by suspending them from a projecting arm or beam" and second is "a tall, long-legged, long-necked bird".

In Hindi language "डाक" word has multiple senses: one is "नीलाम के समय चीज़ का चिल्लाकर दाम लगाने की क्रिया" and another one is "डाक व्यवस्था द्वारा लाई या ले जाई जाने वाली चिट्ठियाँ आदि" .

Similarly, Punjabi language is also widely used in India. If we take example of Punjabi ambiguous word, ਉੱਤਰ i.e. "answer" or "North direction".

Table 1. Example of ambiguity in Punjabi language

Context 1	ਤੁਸੀ ਮੇਰੇ ਪ੍ਰਸ਼ਨ ਦਾ ਉੱਤਰ ਨਹੀਂ ਦਿੱਤਾ / ਪੈਪਰ ਵਿਚ ਪੁਛੇ ਗਏ ਸਾਰੇ ਸਵਾਲਾਂ ਦੇ ਉੱਤਰ ਗੱਲਤ ਸੀ
Context 2	ਭਾਰਤ ਦੇ ਉੱਤਰ ਵਿਚ ਹਿਮਾਲਿਆ ਪਰਬਤ ਵਿਰਾਜਮਾਨ ਹੈ

These ambiguities in the natural languages are the biggest barrier in natural language processing. The task of determining the exact sense in the given context is known as Word Sense Disambiguation (WSD) in NLP.

In Table 1, the word ਉੱਤਰ has two different meaning in both the contexts. In the first context, a word ਉੱਤਰ refers to the English word "Answer". In the second context, a word ਉੱਤਰ refers to the north direction. WSD is the method which is used to identify the correct meaning of a particular ambiguous word in the given context. Most of the work is done which focuses on English or Hindi language. Our efforts consider Punjabi language WSD.

There are 4 conventional approaches to WSD: knowledge based, supervised WSD, semi-supervised or minimally supervised and unsupervised method. Knowledge based method is rely on the lexical knowledge base and dictionaries without any corpora evidence. In supervised WSD method, training is provided to make use of sense-annotated corpora. Semi-supervised or minimally-supervised methods make use of a secondary source of knowledge such as a small annotated corpus as seed data in a bootstrapping process, or a word-aligned bilingual corpus. In Unsupervised WSD approach work directly from raw annotated corpora. In [1], lesk approach is used to detect the correct meaning of a given ambiguous word. It uses a fixed size context window and uses the words in Hindi language for disambiguation. In [2], author has disambiguated Hindi language using the lesk approach. Hindi word net has chosen from where the ambiguous words have been selected. The results shows that the precision of the algorithm increase if the size of the context window is large.

Table 2: Different senses of word “ਹਾਰ”

1.	ਹਾਰ, ਹਾਰ_ਜਾਣਾ, ਅਸਫਲਤਾ, ਮਾਤ, ਫੇਲ, ਭਾਜ_ਖਾਣਾ, ਸ਼ਿਕਸਤ, ਪਿੱਛੇ_ਹੱਟਣਾ, ਪਿੱਠ_ਲੱਗਣਾ, ਝੁਕ_ਜਾਣਾ, ਗੌਡੇ_ਟੇਕਣਾ; ਹਾਰਨ ਦੀ ਅਵਸਥਾ ਜਾਂ ਭਾਵ ; "ਇਸ ਚੌਣ ਵਿੱਚ ਉਸਦੀ ਹਾਰ ਨਿਸਚਿਤ ਹੈ /ਚੌਣ ਵਿੱਚ ਉਸ ਦੇ ਹਾਰ ਹੱਥ ਲੱਗੀ"
2.	ਹਾਰ, ਮਾਲਾ, ਗਾਨੀ, ਚੈਨ, ਚੈਨੀ, ਤਵੀਤ, ਕੈਂਠਾ; ਸੂਤ ਆਦਿ ਵਿਚ ਗੋਲਾਆਕਾਰ ਪਰੋਈ ਹੋਈ ਵਸਤੂ ਜਿਵੇਂ ਮਣਕਾ,ਫੁੱਲ ਆਦਿ ਜੋ ਗਲੇ ਵਿਚ ਪਹਿਨੇ ਜਾਂਦੇ ਹਨ ; "ਉਸ ਦੇ ਗਲ੍ਹੇ ਵਿੱਚ ਮੋਤੀਆਂ ਦੀ ਮਾਲਾ ਸ਼ੁਸੋਭਿਤ ਹੋ ਰਹੀ ਸੀ"
3.	ਹਾਰ, ਕੁੱਟ, ਥੱਪੜ, ਪੀਟਾਈ ; ਮਾਰਨ ਦੀ ਕਿਰਿਆ; "ਮਾਰ ਸਹਿੰਦੇ ਸਹਿੰਦੇ ਉਹ ਢੀਠ ਹੋ ਗਿਆ ਹੈ/ਅੱਜ ਉਸਦੀ ਖ਼ੂਬ ਮਰੰਮਤ ਕੀਤੀ ਜਾਵੇਗੀ"

Table 2 shows the word “ਹਾਰ” with 3 different senses. For the computer system, to identify the correct meaning of the target word in the given context is very difficult task to achieve. For this proper training is provided as per the use of supervised learning methodology. For instance, the correct sense of the word “ਹਾਰ” in the sentence “ਪਿਛਲੇ 5 ਸਾਲਾ ਚੋ ਭਾਰਤ ਦੇ ਹਥ ਕੋਈ ਹਾਰ ਨਹੀ ਲਗੀ ਸੀ | “is sense 1 from the Table 2. The rest of the paper is organized as follows. In section II we present the related work for word sense disambiguation using lesk algorithm. Details of the suggested algorithm are presented in section III. We conclude the paper with a discussion in the final section.

2. REVIEW OF LITERATURE

Agarwal et al. [6] describe the increased demand of Hindi language in today's world. With the increase in Hindi Language demand, the amount of polysemous words is also increasing. It is important to judge the correct sense of the polysemous words for the proper processing of the language. The author derives the target word from Hindi Word net which is developed at IIT Bombay. In this paper the correlation analysis of context using the target word in collation to the vector of definition of the target word is computed. The method in this paper assigns weights to the different senses of the target words. The weights are assigned using collation and co-occurrence information of the target word. In this paper, knowledge based learning is used. The words which are located at the left and the right of the target words; their information is encoded using collation features. The meaning of the target word is determined by using the correlation matrix. The dataset consisting of 60 polysemous Hindi words is used in this paper. With the average window size 7, 8 and 9, the precision achieved was 88.92. Nouns were used in this paper and their collation and co-occurrence information was obtained from Hindi word net.

Jain et al. [7] describe the detection of ambiguity in Hindi language by using graph connectivity measure and Hindi Word Net. Node neighbor and graph clustering measures are used to determine the importance of notes. All open words that are noun, verb, adjective and adverb are tackled in this paper. The author explains word sense disambiguation as the method of judging the correct sense of the word which has multiple meanings. Hindi which is a common language in India has numerous words which can be interpreted in different ways. For a machine to judge these dual meaning words accurately, some kind of training has to be provided to the system. The author in this paper explains the importance of constructed graph. Constructed graphs are used to weigh the vertices of the graph. This is done by calculating the connectivity measures. These are based on the node neighboring and graph clustering. The correct sense of the word is determined by examining the most important vertex of the graph. In the proposed method the vertices of the target word is created. These vertices are created with the help of the most appropriate meaning of the target word as described in the Hindi Word Net which has been developed as a standard for Hindi Language. These are vertices are added by using depth breadth search and are then measured. The vertices with highest rating are the correct meaning of that particular ambiguous word. Various graph connectivity measures such as connectivity measures based on node neighbors, degree centrality, edge betweeness centrality etc. are used in this paper for determining the vertices of the graph and eventually determining the correct meaning of ambiguous words.

Haroon [8] describes the problem of word sense disambiguation in Malayam language. The author describes WSD a common problem in computational linguistics. WSD is defined as detecting the correct sense of the word from the given text. The target words are those which have multiple meanings. It is very easy for a human being to detect the correct sense of the word, but for a computer machine it is not easy to judge the correct sense of the ambiguous word. The author describes that WSD can be done by two different approached that is machine based approach and knowledge based approach. Knowledge based approach are easy as compared to machine based approach as it involves only to look up into a source from where the knowledge can be retrieved such as dictionary. In other approach that is machine learning approach the systems are trained to detect the proper meaning of the words. This is a long process as the system needs to be given human capabilities. This paper uses knowledge based techniques to detect the ambiguity in Malayam language. The author given input to the system as dataset of Malayam words which contains ambiguous words. The author gives two inputs to the system that is dataset and knowledge source. Dataset has contexts and knowledge has senses from which the system will judge the meaning of the words. The algorithm works by overlapping the ambiguous with the senses. Based on the overlap, a score is generated and the higher score is the correct sense of the ambiguous word.

Hwang et al. [9] describe the automatically creation of lexical dictionary for unknown word. The unknown word is generated by new technology, new product, and new style and so on. The approach in this work calculates the relationships and distance between the terms using word net; and this is known as relation structure. Based on this relation structure, the algorithm is able to detect the correct meaning of the words. This approach is based on the structural semantic interconnection which is the best known technique for WSD. The author further describes that most of the work in the area of WSD is based on semantically matching of the meanings of various words with the target word. These are basically subject near words of the target words. The main objective of this work is creation of unknown word lexical dictionary in an autonomous way. Structural semantic interconnections is the best way of WSD done till current date. The main focus of this work is to automatically create dictionary of unknown words and extract nouns from it in an autonomous way to detect the correct meaning of those nouns. The work first focus on creating the correct nouns in a document, then detecting the type of nouns that is whether it is a single noun or a compound noun. Then based on the noun list a relation structure is applied which will detect the relationship between nouns based on word net. The relational structure approach for WSD which is used in this paper detects the highest relationship between the words.

Zampieri [10] describe a case study on knowledge poor and knowledge rich features in automatic classification of words. This work focus on performance evaluation of five machine learning approaches of automatic disambiguation. These five approaches are naïve bayes, Support Vector Machines, Decision Trees, KStar and Maximum Entropy. Various methods are used such as statistical methods and rule based method for automatic detection of words which have more than one meaning in any given language. The disambiguation of words is necessary to increase the performance of automatic translation. The disambiguation of the words plays a vital role in the task of information retrieval. The ambiguous word having more than one meaning needs to disambiguate before they can be retrieved by the system. A very famous approach that is lesk approach uses dictionary method for disambiguation of the words. The lesk approach uses the dictionary. It matches the dictionary definition of the target words with the most closely related left and the right words of the target words. These target words are also known as neighboring words. The correct sense of the word is detected by measuring the most common words and counting it. This work as described above focus on comparing the results of the five techniques which are used for disambiguation of ambiguous words. The work first creates a data set of words from the Portuguese language and then tries to determine the correct sense of the word from the dictionary. Three classes for each word has been formed in this work namely S1, S2 and S3. Knowledge poor features were also used to disambiguate the ambiguous words. These features were divided into three groups that is text domain, neighboring words and key words. The work provides an overview of each classifier used to disambiguate the words. The parameters used to measure the results are precision, recall, f-measure and accuracy. The results showed that the knowledge rich features are somewhat less efficient as compared to knowledge poor features. Support Vector Machines, which is a knowledge rich feature yielded better results as compared to others.

3. MODIFIED LESK ALGORITHM

WSD is a technique by which single word having more than one meaning is clubbed together. The system is then trained to find the correct meaning of that word. Our algorithm uses the concept of dynamic context window. The context window is comprised of left and the right words of the target ambiguous word. The removal of special tokens like “|” & “,” is the first step of the proposed algorithm.

At every step we increase the size of the context window, since greater the size of the context window better will be the precision of the algorithm. Instance and Precision of the algorithm are calculated after determining the number of an ambiguous word.

3.1 Proposed Algorithm

Modified Lesk Approach

```
$\boldsymbol{Array\ A} < -Ambiguous\ words$    { “ ” , “ਹਾਰ”, “ਮੂਲ” , “ਵੰਡ” , “ਭਾਗ” ,
                                  “ਉੱਤਰ” , “ਗ੍ਰਾਮ” , “ਹਲ” , “ਸੰਬੰਧ” ,
                                  “ਧੁਨ” , “ਬੋਲੀ” , “ਚਾਲ” , “ਖਾਨ” ,
                                  “ਟਿੱਕਾ” , “ਫਲ” , “ਵਿਚਾਰ” }

$\boldsymbol{Precision\ P < -\left\{\frac{number\ of\ instances}{number\ of\ average\ words}\right\}}$

$\boldsymbol{Step\ 1}$:   $Words\ < -\ no.of\ words\ in\ a\ sense\ after\ removal\ of\ special\ tokens$
$\boldsymbol{Step\ 2}$:   $Sense\ computation < -\ no.of\ senses\ of\ a\ word$
$\boldsymbol{Step\ 3}$:   $Determine\ instance\ count$
$\boldsymbol{Step\ 4}$:   $If\ word\ sense\ overlap\ target\ word\ sense$
    $\boldsymbol{Step\ 5}$:   $Instance\ count\ =\ +1$
    $\boldsymbol{Step\ 6}$:   $Calculate\ precision\ (P) for\ every\ target\ word$
    $\boldsymbol{Step\ 7}$:   $if\ P < threshold\ value$
    $\boldsymbol{Step\ 8}$:   $increase\ context\ window\ size$
        $\boldsymbol{Step\ 9}$:   $calculate\ precison\ again$
```

4. RESULTS AND DISCUSSIONS

Table 3: Instance Output

Word	No. of Senses	Sense 1	Sense 2	Sense 3	Sense 4
ਹਾਰ	4	144	64	56	0
ਮੂਲ	4	21	36	37	32
ਵੰਡ	4	42	39	32	28
ਭਾਗ	3	42	64	34	0
ਉੱਤਰ	4	25	48	10	14
ਗ੍ਰਾਮ	2	50	100	0	0
ਹਲ	2	102	24	0	0
ਸੰਬੰਧ	3	24	56	8	0
ਧੁਨ	2	24	32	0	0
ਬੋਲੀ	3	8	56	24	0
ਚਾਲ	4	27	43	16	14
ਖਾਨ	3	15	9	8	0
ਟਿੱਕਾ	3	18	32	21	0
ਫਲ	3	22	9	11	0
ਵਿਚਾਰ	2	16	12	0	0

Table 3 shows the instance output of the various ambiguous words. The number of senses denotes the number of meanings a particular word has. Every instance value gives the usage of the particular meaning of an ambiguous word in a given context. The system needs to be trained before it can judge the correct instance of the meaning of the word.

Table 4: Precision Output

Word	Precision				
	n=5	n=10	n=15	n=20	n=25
ਹਾਰ	0.1626	0.3434	0.5	0.6509	0.7711
ਮੂਲ	0.1154	0.249	0.3896	0.5564	0.6582
ਵੰਡ	0.1785	0.2882	0.4393	0.545	0.646
ਭਾਗ	0.1958	0.3282	0.445	0.5962	0.7009
ਉੱਤਰ	0.134	0.2716	0.3928	0.5286	0.709
ਗ੍ਰਾਮ	0.1214	0.2957	0.4547	0.6139	0.7452
ਹਲ	0.1071	0.2793	0.4329	0.5678	0.689
ਸੰਬੰਧ	0.1975	0.359	0.4882	0.6266	0.7548
ਧੁਨ	0.1129	0.2577	0.3897	0.5729	0.7082
ਬੋਲੀ	0.1103	0.2812	0.399	0.5621	0.7371
ਚਾਲ	0.1157	0.2833	0.3988	0.5015	0.6866
ਖਾਨ	0.1882	0.31	0.4882	0.6096	0.7209
ਟਿੱਕਾ	0.1043	0.2634	0.367	0.5417	0.7305
ਫਲ	0.1102	0.2402	0.4083	0.5679	0.7458
ਵਿਚਾਰ	0.1972	0.3457	0.4677	0.5953	0.7231

Table 4 shows the precision output of the various words. The output helps us to see that if the context window size is increased the precision of the algorithm will be increased since the system will then be able to judge the correct meaning of the word more accurately.

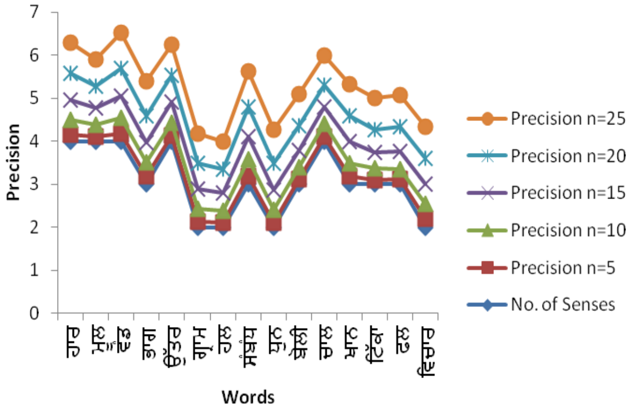

Fig. 1: Precision Graph

Fig. 1 shows the graphical output of the precision of the algorithm. It can be analyzed from the graph that the precision for the context window size 25 is the maximum. Thus it can be concluded that more the size of the context window, more is the accuracy of the system to judge the correct sense of the word.

5. CONCLUSION

In this paper, the lesk approach has been modified which now uses a dynamic context window. The context window which forms the left and the right words of the target word is chosen dynamically. The results shows that greater the size of the context window more accurate the results will be. The output has been formed in two ways. Table 1 show the instance output and the highest value of instance is treated as the correct meaning of the word. Table 2 shows the precision of the target words. Precision basically shows the accuracy of the algorithm for every target word for various sizes of context window. The whole work has been carried out considering Punjabi language. The current work can be extended with the help of Support vector machine which creates a graph and arranges the words on the left and the right side of the graph.

6. REFERENCES

[1] Satyendr Singh and Tanveer J. Siddiqui," Evaluating Effect of Context Window Size, Stemming and Stop Word Removal on Hindi Word Sense Disambiguation," in IEEE, pp. 1-5, 2012

[2] Radhike Sawhney and Arvinder Kaur," A Modified Technique for Word Sense Disambiguation Using Lesk Algorithm in Hindi Language," in International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, pp. 2745-2749, 2014

[3] Priti Saktel and Urmila Shrawankar," Context Based Meaning Extraction for HCI Using WSD Algorithm: A Review," in International Conference On Advances In Engineering, Science And Management (ICAESM), IEEE, pp. 208-212, 2012

[4] Miguel Ángel Ríos Gaona, Alexander Gelbukh and Sivaji Bandyopadhyay," Web-based Variant of the Lesk Approach to Word Sense Disambiguation," in Eighth Mexican International Conference on Artificial Intelligence, IEEE, pp. 103-107, 2009

[5] Sudip Kumar Naskar and Sivaji Bandyopadhyay,” Word Sense Disambiguation Using Extended WordNet,” in Proceedings of the International Conference on Computing: Theory and Applications (ICCTA'07), IEEE, 2007

[6] Madhavi Agarwal and Jyoti Bajpai, “Correlation based Word Sense Disambihuation,” in IEEE, 2014

[7] Amita Jain, Sudesh Yadav and Devendra Tyal,” Measuring Context-Meaning for Open Class Words in Hindi Language,” in IEEE, pp. 118-123, 2013

[8] Rosna P Haroon,” Malayalam Word Sense Disambiguation,” in IEEE, 2010

[9] Myunggwon Hwang, Chang Choi, Byungsu Youn and Pankoo Kim,” in International Conference on Advanced Language Processing and Web Information Technology, IEEE, pp. 15-20, 2008

[10] Marcos Zampieri,” Evaluating Knowledge-poor and Knowledge-rich Features in Automatic Classification: A Case Study in WSD,” in 13th International Symposium on Computational Intelligence and Informatics (CINTI), IEEE, 20-22 November, pp. 359-363, 2012