A PERFORMANCE OF SVM WITH MODIFIED LESK APPROACH FOR WORD SENSE DISAMBIGUATION IN HINDI LANGUAGE

Sandy Garg[1], Anand Kumar Mittal[2]

[1]*Department of Computer Science and Engineering, Guru Kashi University, Punjab, India*
[2]*Asst Prof, Department of Computer Science and Engineering, Guru Kashi University, Punjab, India*

Abstract

WSD is a Technique used to find the correct meaning of a given word in any human language. Each human language has a problem called ambiguity of a word. To finds the correct meaning of any ambiguous word is easy for human but for a machine it is great issues No of work has done on WSD but not enough in Hindi language. My objective is to provide the training to the system so that it can easily find the correct meaning of the any ambiguous word in Hindi language .For this purpose I simple used a one existing technique name modified lesk approach and give its output to the SVM to get the better result and show that SVM is better in compare to modified Lesk Approach, In this paper I simply take nine Hindi ambiguous words and three different databases to show the result.

Keywords: *Support Vector Machine, NLP, Word Sense Disambiguation, Modified Lesk approach, Comparison*

----------***----------

1. INTRODUCTION

In any language, words may convey more than one sense. The right sense of the word can be identified based on the text in which it occurs. It is the essence of communication in NLP. For instance, consider an example of the word 'volume'. The volume of the music is too high. In this sentence "the volume means the loudness of the sound instead of other possibilities like amount of space occupied by something or a series of a book". To identify the particular meaning of a word, also known as lexical disambiguation or WSD, is hardly a problem for a human, but for a machine, which has no base for knowing which meaning is suitable in a given sentence is a complex task. As a computerized problem, it is described as "AI-complete", that is, a problem whose solution exists but is not known. WSD has been considered necessary in almost every application of language technology, including machine translation. The language of human is ambiguous, so mostly words can be identified depending upon the context in which they arise. Consider an example:

1. The pencil has a sharp point.
2. It is not polite to point on people.

In the above sentences the word point is common in both the sentences. For a human it is understandable that in the:
First sentence the word "point" denotes some sharped / pointed object and the word "point" .Second line denotes to notify someone directly in a negative way. Thus, a term can only be shifted if the exact meaning of the word is assumed by the user.

In Hindi language a single word has different meaning. This is handling in machine by using WSD algorithms.WSD simply finds the correct sense of a given word. It is the essence of communication in natural language processing. For instance, consider an example of the word 'volume'. The volume of the music is too high. In this sentence "the volume means the loudness of the sound instead of other possibilities like amount of space occupied by something or a series of a book". To identify the particular meaning of a word, also known as lexical disambiguation or WSD, is hardly a problem for a human, but for a machine, which has no base for knowing which meaning is suitable in a given sentence is a complex task. But in case of human it is easy to find the correct sense of a word in Hindi language or changing the meaning of the word according to requirement. So the working of WSD is simply finding the correct meaning of ambiguous word in Hindi language in a given context.

Table-1: Example of ambiguous word "ढाल "

Context1:ढाल पर पहोंचते ही रमेश ने साइकिल का पैदल मरना बंद कर दिया |
Context2:ढाल यौधेयों को सुरक्षा प्रदान करता है |

The word "ढाल" is common in both sentences. But it has different meaning in both the sentences. For a human it is easy to find the right sense of the word "ढाल" in both the sentences. Thus, with the help of WSD techniques computer suytem will be able to find the exact sense of an ambiguous word. In table1 word "ढाल" has two senses. In first sentence the word "ढाल" refers to the plane and in second sentence the word "ढाल" refers to the Kavach.

Word sense disambiguation is the process of determining the correct sense of a word with multiple meanings. The meaning of the word depends upon the right and left words of the ambiguous word. WSD serves as an intermediary step

for many applications like speech processing, computer translation, data retrieval and hypertext navigation. Researchers have been working in this field to enhance the process of WSD, as high performance of this has an impact on various applications like search engines etc. There are many methods to perform the procedure of WSD. These are supervised approach, semi-supervised approach, unsupervised approach and Knowledge based approach or Dictionary based approach. Supervised approach is also known as "Corpus-based "approach which have high interaction between human and system.

Supervised methods are based on common sense i.e. the sentence in which ambiguous words occurrence can provide enough information by its own which is needed to disambiguate the word. SVM's and memory based learning are the most core examples of this method. Unsupervised methods are also known as word sense induction. This technique assumes that the senses which are similar usually occur in similar contexts. The disadvantage of supervised approach lead to the generation of semi- supervised method. In this type of learning both labeled and unlabelled data is allowed. Dictionary and knowledge based method works on the principle that the words with multiple meanings are correlated with each other and their relation can be obtained from the words and their senses. This approach works by overlapping the words with the greatest match of an ambiguous word with the dictionary meaning. This is a traditional approach used for disambiguation.

To determine the exact meaning of a given word in Hindi language we simply use Hindi word net. It is a standard to determine the meaning of Hindi word. It is currently available in Princeton University since the 1980s. The meaning is represented in the wordnet with the help of polysemous word. It is like a dictionary where meaning and synonyms of the word is present. To find the exact meaning of given ambiguous word to check its exact meaning from Hindi dictionary.WSD is needed in a no of application to information retrieval, correction of spelling etc .No of algorithm are present for the purpose of WSD. But in Hindi language no much work is done for remove the ambiguity of a given word. And no more work is done to compare which algorithm is best to remove the ambiguity of a given word.

Based on available literature and previous result I simply find Support vector machine is a best technique to solve the word sense disambiguation. For this purpose I compare the two approach name SVM and New Lesk approach in which first is based on knowledge or Decision based approach and second is based on supervised approach .To find the accurate result ten Hindi words are taken for experiments. Display the graph which is generated by SVM and display the table which is generated by Lesk approach.

2. LITRATURE SURVEY

Rathode et al. [2] describe, transliteration for Hindi to English and Marathi to English language pairs using SVM. In this approach, the source named entity is segmented into transliteration units. The classification of phonetic units is done by using the polynomial kernel function of SVM. Little research is carried for Indian and English languages. In Hindi to English transliteration the most of the methods incorporated are statistical in nature. The method used in this work is SVM which is the most efficient statistical supervised learning mechanisms to attain the transliteration.

Lin et al. [4] Provide the concept of preserving privacy classifier to save the important material of support vectors in the classifier. The SVM classifier if made publically available to the consumers, it reveals the private material of support vectors. This disturbs the preserving secrecy requirements for some lawful or profitable reasons. In this paper privacy violation problem is described. SVM classifier post processes the characteristic values of support vectors. The security is proved through adversarial attacks. An approach is proposed to process the Support vector classifier to convert it to a preserving privacy classifier that does not disclose the isolated satisfied of Support vector machine. Here Gaussian kernel function is used which gives training to the Support vector classifier for which Privacy Preserving Support Vector Classifier (PPSVC) is designed. Support vector machine tool gives training to the classifier. PPSVC approaches the decisive function of the Gaussian kernel. To evaluate the performance of privacy preserving SVM experiments are being conducted. The sensitive part of support vector is protected resulting PPSVC can publically be free. Future work could be the challenge of applying privacy preserving support vector classifier to high dimensional data.

Ekbal et al. [5] describe the Named Entity Recognition (NER) system for the development of Bengali language using SVM. Machine learning methods has been applied to NER in various well studied languages. For the first time the attempt is done for a Indian languages particularly Bengali. In Indian language the NER is general but for Bengali it is a difficult task. NER system is developed by using Bengali corpus obtained from an archive of a labelled Bengali newspaper. For other Indian languages particularly Hindi, Bengali, Urdu and oriya the proposed SVM system is to be trained.The information of named entity of having similar words, different kinds of features and the various lists are provided for the best-suited features for the system of NER in Bengali language. The results after experimentation with the 10-fold cross confirmation test have described reasonably good Recall, Precision and F-Score values. The comparison of system has been done with the existing three NER Bengali systems and the evaluation has been described that the SVM-based system beats other systems. Precision, Recall and F-Score values of the based system, SVM is able to hold the diverse and overlapping features of the Indian languages is the one possible reason.

LI et al. [6] provides an original method to retrieve protein-protein interaction (PPI) material from biomedical literatures based on SVM. In this work both SVM and K-Nearest Neighbour (KNN) classifiers are used. SVM is used to retrieve the interaction. The introduction of classifier is to improve the accuracy of SVM. On performing the

experiments the achievement of approach can obtain highest F-score in removing protein-protein extraction information. SVM is setup to extract the interaction. KNN method is introduced to improve the accuracy of SVM. To proper fitting of the unbalanced data, a modified SVM-KNN classifier is introduced. The extraction of information of PPI is treated as a binary classification problem.

Bhardwaj et al. [9] provide ensemble methods that use multiple classifiers to obtain better models of classifier. The analytical performance can be obtained from any of the constituent model. In ensemble classifier, SVM is used as a base learner and combines with those base learners which will have higher accuracy model. Here usage of combine learners is used for voting methods. The Voting approach is found to be efficient, as some classifiers are generated and some are selected. An efficient classifier where each base classifier is derived to ensemble with SVM, from the prior one by the exercise set. The better accuracy is shown by ensemble classifier as compared to the ensembles created using ADABOOST, Bagging. In this work, the creation of ensemble is by selecting the classifiers with greater accuracy than the average accuracy of the classifiers generated. The analysis of theoretical effect of the classifier on the diversity and the issues of convergence through which this work can be extended.

2.1 SVM with Modified Lesk Approach

According to literature view no of work done on SVM .But My approach is based on the mixture of supervised learning approach and modified lesk approach. In this I simply give the instance output of modified lesk approach to SVM and get the instance output of SVM to find better result. This algorithm is based on to provide the training to the system. And after the initial phase is completed, future data sets given to the algorithm can de classified with minimal human intervention. The Table2 show the working step of SVM with modified lesk approach.

Table-2 SVM with New Lesk Approach

Step 1:		Calculate the no word in a sentence. This includes the removal of special tokens like ',' or '|' followed by all the specialized symbols.
Step 2	senseCount←Number of senses	Calculate the number of senses of the word.
Step 3	Instance Count ←No of senses in given target Context window of size *n*, where *n* is determined dynamically	Calculate the instance output of every target word.
Step 4	Import a word dictionary and use it as a database	
Step 5	The Dictionary contains words and its meaning which are assigned or specific decimal values	
Step 6	Now match the Instance values with the words inside the dictionary and get the actual meanings of the word. And then display SVM graph.	

3. RESULT AND CONCLUSION

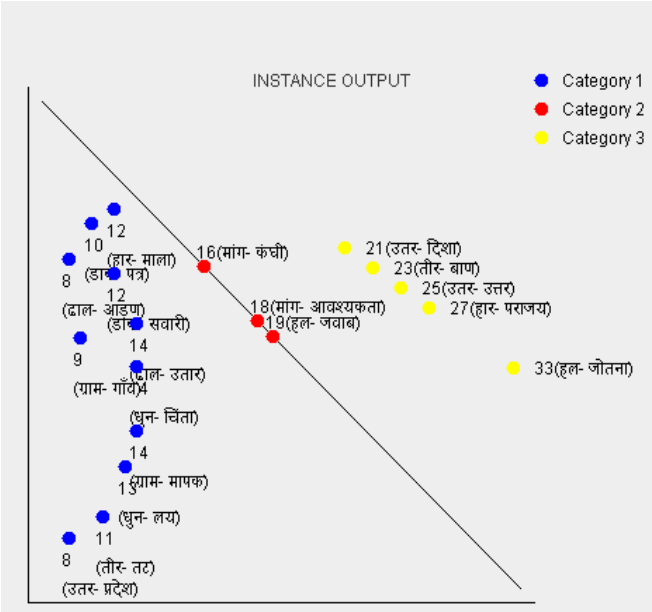

Fig -1: Instance output of SVM with new lesk Approach

Table- 3: Instance output of Modified Lesk Approach

	Number of Senses	Sense 1	Sense 2	Sense 3	Sense 4
हार	2	12	27	0	0
डाक	2	10	12	0	0
ढाल	2	14	8	0	0
धुन	2	14	13	0	0
ग्राम	2	9	14	0	0
हल	2	33	19	0	0
मांग	2	18	16	0	0
तीर	2	22	17	0	0
उतर	3	25	21	0	0

The above Fig 1 and Table 5 shows the Instance output of two different approaches; New Lesk approach and SVM with modified lesk approach.

To find the output I simply take nine hindi words to find the Instance output of SVM with modified lesk approach.According to ouput we see that Instance output of

modified lesk approach simple tell the no of senses in given context but no tell the exact meaning of the target word. So my aim is to provide the traning to the the system to with the help of SVM so that machine easily find the exact meaning of the target word. Fig 1 display the result in graphical form in compare to modified leak approach. Modified Lesk approach simply show output in a table form and simply tell the senses of the word and not the meaning of the word in the given context. The similar Data set is provided to both the approaches. The data set contains ten Nine words, these words have more than one meaning and they form ambiguity in the paragraph used in a training set. As seen from the table SVM approach has high efficiency to display the result and to disambiguate the paragraph in compare to the Modified Lesk approach.

REFERENCES

[1] Sawhney, Radhike, and Amardeep Kaur. "A modified technique for Word Sense Disambiguation using Lesk algorithm in Hindi language." Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on. IEEE, 2014.

[2] P H Rthode, M L Dhore and R M Dhore, "Hindi and Marathi to English Machine Transliteration using SVM", vol.2, August 2013

[3] Satyendr Singh and Tanveer J. Siddiqui, " Evaluating Effect of Context window size, stemming and stop word removal on Hindi word sense Disambiguation" , *Proc. IEEE* ,2012

[4] FattanehJabbari, HosseinSameti and Mohammad HadiBokaei, "Unilateral Semi – supervised learning of extended hidden vector state for Persian language understanding," Natural language processing and knowledge Engineering (NLP-KE),2011 7th International Conference on pp. 165-168, November 2011

[5] Keng-Pei Lin and Ming-Syan Chen, "On the Design and Analysis of the Privacy-Preserving SVM Classifier", *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, No, 11, pp. 1704-1717 , November 2011

[6] Asif Ekbal, Sivaji Bandyopadhyay "Bengali Named Entity Recognition using Support Vector Machine", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp. 51–58,

[7] Lishuang LI, Linmei Jing and DegenHaung, "Protein-Protein Interaction Extraction from bio medical literatures based on modified SVM-KNN", *Natural Language Processing and knowledge Engineering 2009, NLP-KE 2009, International conference on* , pp. 24-27 September 2009

[8] Roberto Navigli, "Word Sense Disambiguation: A survey", *ACM computer Survey. 41, 2, Article 10*, pages 69, DOI=10.1145/1459352.1459355 http://doi.acm.org/10.1145/1459532.1459355, February 2009

[9] Jinying Chen, DmitriyDligach and Martha Palmer, "Towards Large-scale High-Performance English Verb Sense Disambiguation by Using Linguistically Motivated Feature,"International Conference on pp. 378-388, September, 2007

[10] Manju Bhardwaj, Trasha Gupta, Tanu Grover and VasudhaBhatanagar, "An Efficient Classifier Ensemble using SVM", Methods and models in computer science, 2009, ICM2CS 2009, Proceeding of International conference on ,pp. 240-246,December 2009

[11] Yee Seng Chan, HweeTou Ng and David Chiang," Word Sense Disambiguation Improves Statistical Machine Translation", *Proc. IEEE*, pp. 33-40, June 2007

[12] Gondy Leroy and Thomas C. Rindflesch, "Effects of information and machine learning algorithm on word sense disambiguation with small datasets," *Proc. Elsevier*, pp. 573-585, March 2005

[13] Hee-CheolSeo, Hoojung Chung, Hae-Chang Rim, Sung HyonMyaeng and Soo-HongKim, "Unsupervised word sense disambiguation using word net relatives," *Proc. Elsevier* , pp.253-273 , June 2004