International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 2, Issue 5, May 2015, PP 24-28
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
www.arcjournals.org

A Comparative Study of Svm and New Lesk Algorithm for Word Sense Disambiguation in Hindi Language

Mr. Sandy Garg
Department of Computer Science and Engineering, Guru Kashi University, Punjab, India

Er. Anand Kumar Mittal
Asst Prof, Department of Computer Science and Engineering, Guru Kashi University Punjab, India

Abstract: *This paper is simply based on comparative study of the two algorithms which is based on Word Sense Disambiguation in Hindi language. Hindi is an human language which is used for people to express its views, In India the most used language for people is Hindi. The* ***third most spoken language*** *in all over the world is Hindi by the people. In Hindi language a word have more than one meaning, which is known as Ambiguity. For a human it is not difficult to find the meaning of a word but for a machine it is a big issue. Word Sense Disambiguation (WSD) is used for to find the correct meaning of a given word. There are no of algorithm purposed for Word Sense Disambiguation(WSD).The main problem is what algorithm used to get accurate or suitable result for a user to find correct meaning of a given Word .There are basically four methods for WSD namely supervised, unsupervised, semi supervised and knowledge based. There are various algorithms that come under supervised learning approach. . In this paper compare two supervised approach used to resolve the problem of ambiguity. Support Vector Machine (SVM) and new lesk algorithm is one of the algorithms that come under supervised learning approach. So the main need of this paper is to study the comparative analyst of above two or more WSD algorithm and examine its effects on Hindi language. In my paper for find the result I simply take 10 Hindi words to compare the algorithms.*

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

1. INTRODUCTION

In Hindi language a single word has different meaning. This is Handel 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.

Table1. *Example of ambiguous word “मूल”*

Context1: बरगद के वृक्ष कि मूल सबसे मोटी होती है
Context2:गाओं के लो बहुत बीमार हो रहे थे जिसका मूल वहाँ की फैली गंदगी था
Context3: हिंदुस्तान में महंगाई होने का कारन हर वस्तु का मूल बढ़ गया है

In Table1 the word “मूल” is common in three sentences. In Hindi language it represents the different sense in different context. In Table1 the word “मूल” represent three senses. In first context the “मूल” refers to the stem of the tree. In second context “मूल” refers to the reason. In third context “मूल” refers to the price of things. For a human it is easy to find the correct sense of a given word in no of contexts. But for a machine it is a great issue. This problem is solving by machine with help of ambiguity.

For a system it is easy to find the correct meaning of a given word in a given context if we provide a traning to the system .From table1 the correct sense of word “मूल” in context2 is sense 1 from table 2.

Table2. *Different senses of word "मूल"*

1Basic: बुनियादी,आधार भूत, आधार की, आधार संबंधी , मूल 2Basis:आधार,नींव,जड,मूल 3Constituent: निर्वाचक निर्वाचकसंघका सदस्य मूल सार,त्वअवयव 4Content: तृप्ति , सन्तोष, सन्तुष्टता, रजामन्दी समाई, तत्व, मूल copy: नकल उतारना, मूल देख कर प्रतिलिपि बनाना , अनुकरण करना 6Decompose: पृथक पृथक करना, मूल अवयवों मे परिणत करना, गलाना, विच्छिन्न करना, विश्लेषण करना . सडाना . बिगड जाना या सड जाना 7Derivative: दूसरी वस्तु से लिया हुआ , व्युत्पतिलब्ध , व्युत्पन्न , मूल धातु से उत्पन्न शब्द,यौगिक पद.प्रसूत 8Element: तत्व, मूल वस्तु, मूल, जड 9Fundamental: मूलभूत, आवश्यक, मौलिक, बुनियादी

No of WSD techniques are present to find the correct sense of a given word in a given context in Hindi language. 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 labelled 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 which approach is better to find the meaning of ambiguis word in Hindi language. 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. For each approach a table is drawn to show the result. Algorithms are present in section 2 and their result and conclusion find in section 3.

2. Different WSD Approaches

2.1. Support Vector Machine [1]

This approach is based on supervised approach. This algorithm is based on the to provide the training to the system. And after the initial phase is completed, future data sets given to the algorithm can be classified with no or minimal human intervention. The Table3 show the working step of SVM algorithm.

Table3. *SVM Algorithm*

Step 1:	Import a word dictionary and use it as a database
Step 2	The Dictionary contains words and its meaning which are assigned or specific decimal values
Step 3	Now match the Precision values with the words inside the dictionary and get the actual meanings of the word
Step 4	Display SVM graph

2.2. New Lesk Approach [2]

This approach is based on Knowledge based or decision based. The Table 3 show the working step of new lesk approach.

Table4. *New Lesk Algorithm*

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 ←Context window of size n, where n is determined dynamically	Calculate the instance output of every target word. The context window is dynamic. SVM approach is applied and graphical user interface is developed
	if given word sense overlaps the target word sense, then, instanceCount += 1	Context window are the number of left and the right words of the target words.
Step 4	for each target word in context vector	
Step 5	Determine precision Output	

3. Result and Conclusions

Table5. *Precision output of SVM Approach*

Precision					
	n = 5	n = 10	n = 15	n = 20	n = 25
हार	0.307	0.32	0.335	0.349	0.364
डाक	0.023	0.39	0.457	0.523	0.59
ढाल	0.06	0.09	0.19	0.29	0.29
धुन	0.10	0.14	0.19	0.14	0.19
ग्राम	0.562	0.591	0.62	0.649	0.677
हल	0.197	0.207	0.217	0.227	0.237
मांग	0.0736	0.796	0.080	0.0916	0.0956
मूल	0.07	0.09	0.19	0.29	0.39
तीर	0.10	0.24	0.32	0.56	0.68
उतर	0.662	0.695	0.728	0.762	0.795

Table6. *Precision output of Lesk Approach*

Precision					
	n = 5	n = 10	n = 15	n = 20	n = 25
हार	0.411	0.429	0.448	0.468	0.488
डाक	0.433	0.523	0.612	0.701	0.791
ढाल	1.327	0.121	0.255	0.389	0.523
धुन	1.327	0.054	0.121	0.188	0.255
ग्राम	0.754	0.793	0.831	0.869	0.907
हल	0.264	0.277	0.291	0.304	0.318
मांग	1.067	1.12	1.174	1.227	1.281
मूल	1.327	0.121	0.255	0.389	0.523
तीर	1.327	0.255	0.523	0.791	1.059
उतर	0.887	0.931	0.976	1.021	1.065

The above Table 5 and Table 6 shows the precision values of two different approaches; New Lesk approach and SVM approach.

Here "n" denotes the size of context window and in both the approaches the size of context windows are equal. The precision is calculated for Hindi words over window size (5, 10, 15, 20, 25) is shown in both the tables. The similar Data set is provided to both the approaches. The data set contains ten Hindi 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 precision values which indicates that SVM approach has good accuracy to disambiguate the paragraph in compare to the New Lesk approach.

References

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

[2] 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

[3] 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

[4] 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

[5] 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,

[6] 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

[7] 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

[8] 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

[9] Manju Bhardwaj, Trasha Gupta, Tanu Grover and Vasudha Bhatanagar, "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

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

[11] 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

[12] 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

[13] YoongKeok Lee, HweeTou Ng and Tee Kiah Chia, "Supervised word sense disambiguation with support vector machine and multiple Knowledge sources", Senseval-3: Third International workshop on the evaluation of systems for the semantic analysis of text, 2004

[14] Philip Resnik and David Yarowsky,"A Perspective on Word Sense Disambiguation and Their Evaluation", P*roc. IEEE*

[15] Saeed Mozaffari, Karim Faez and HamidrezaRashidyKanan, "Feature Comparison Fractal Codes and Wavelet Transform in Handwritten Alphanumeric Recognition using SVM Classifier",Proceedings of the 17th International Conference, pp. 23-26, August 2004.

[16] W.M.Campbell, F.Richardson and D.A.Reynolds, "Language Recoginition with Word Lattices and Support Vector Machines", *Proc. IEEE,* 2007.

[17] Gorgevik, D, Cakmakov, D., "Handwritten Digit Recognition by Combining SVM Classifiers", International Conference on, pp.21-24, November 2005.

[18] Feng Cai and Vladimir Cherkassky, "SVM + Regression And Multi – Task Learning", *Proc. IEEE,* 2009.

[19] Joachims, Thorsten, "Text categorization with support vector machines: Learning with Many Relevant Features", Springer Berlin Heidelberg, 1998.

[20] K. Sundarakantham, S. Mercy Shalinie, "Natural Language Grammatical Inference with Support Vector Machines*", proc. IEEE, 2005.*

[21] Satyendr Singh and Tanveer J. Siddiqui, "A Supervised Algorithm for Hindi Word Sense Disambiguation", in University of Allahabad international Journal of Systems Engineering Publications, Hyderabad 2013

[22] Vasilescu, Florentina, Philippe Langlais, and Guy Lapalme, “Evaluating Variants of the Lesk Approach for Disambiguating Words,” *in LREC*, 2004

[23] Satanjeev Banerjee and Ted Pederson, ”An Adapted Lesk Algorithm for Word Sense Disambiguation using Word Net, “in University of Minnesota, Duluth MN55812 USA, 2002

[24] Timothy Baldwin, Su Nam Kim, Francis Bond, Sanae Fujita, David Martinez and Takaaki Tanaka, “MRD-Based Word Sense Disambiguation: Further Extending Lesk”, 2008

AUTHOR’S BIOGRAPHY

Mr. Sandy Garg, Mtech., Dept of Computer Science and Engineering, Guru Kashi University Talwandi Sabo, Punjab.