iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** V **Month of publication:** May 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Customized Adaptation of Traditional Lesk Method for Sense Disambiguation of Punjabi Words from Medical Domain

Jaskiran Kaur[1], Amardeep Singh[2]
Department of Computer Engineering, Punjabi University Patiala, India

Abstract **- We report in this paper a way of doing Word Sense Disambiguation (WSD) that has its origin in multilingual MT. Various Machine Learning approaches have been demonstrated to produce relatively successful Word Sense Disambiguation systems. For English and other European languages, a lot of research has been done. Indian languages are still intact in this area. A WordNet for Indian languages, IndoWordNet was developed by IIT Bombay which is the only source of information when it comes to Indian languages. We have proposed a algorithm for disambiguating polysemous words in Punjabi words particularly belonging to medical domain and compared the results with the classical Lesk algorithm.**
Keywords- **Lesk algorithm, distance-based methods, dictionary based methods, Indo WordNet**

I. INTRODUCTION

The natural language processing involves a set of tasks and phases that evolves from the lexical text analysis to the pragmatic one in which the author's intentions are shown. One natural language problem is ambiguity, as we can see in the following sentence: "I made her duck". This is a classical example of ambiguity; someone who hears this phrase understands the speaker's intention, but it is harder make the computer understands it. First, the words duck and her are morphologically or syntactically ambiguous in their part-of-speech. Duck can be a verb or a noun, while her can be an object pronoun or a possessive adjective. Second, the word make is semantically ambiguous; it can mean create or cook. Finally, the verb make is syntactically ambiguous in a different way. Make can be transitive, that is, taking a single direct object, or it can be intransitive, that is, taking two objects, meaning that the first object (her) got made into the second object (duck). Finally, make can take a direct object and a verb, meaning that the object (her) got caused to perform the verbal action (duck).

This paper presents an analysis of the existing disambiguation algorithms and it analyses the quality of each of them taking into account the metrics that have been establish for the evaluation. At the same time, it shows a possible combination of features and classifiers to propose a word sense disambiguation algorithm that resolves some deficiencies detected before and improve the evaluation parameters.

Currently efforts are on in India to build large scale Machine Translation and Cross Lingual Search systems in consortia mode. These efforts are large, in the sense that 10-11 institutes and 6-7 languages spanning the length and breadth of the country are involved. The approach taken for translation is transfer based which needs to tackle the problem of word sense disambiguation. Since 90s machine learning based approaches to WSD using sense marked corpora have gained ground. However, the creation of sense marked corpora has always remained a costly proposition. The above situation brings out the challenges involved in Indian language. Lack of resources coupled with the multiplicity of Indian languages severely affects the performance of several NLP tasks. In the light of this, we focus on the problem of developing methodologies that could efficiently disambiguate words in Punjabi language. The idea is to do the annotation work for one language and find ways of using them for another language.

Our work on WSD is done for Punjabi language. The process of word sense disambiguation is a task of assigning the most precise sense to the ambiguous word in the context. Let us consider a scenario with the intention of recognizing the crucial role of word sense disambiguation in Punjabi language.

ਜਗਤੇ ਦੇ ਢਿਡ ਵਿਚ v`t ਪੈ ਰਿਹਾ ਹੈ[

In the above sentence, polysemous word is 'ਵੱਟ' and in the above context 'ਵੱਟ' means 'ਢਿਡ ਵਿਚ ਮਰੋੜ ਪੈਣਾ'.

There are other meanings to 'ਵੱਟ also, which are as follows-

1. ਮੂੰਹਸੁਜਾਉਣਾ

2. ਜ਼ੋਰ ਨਾਲ ਕਿਸੇ ਤੇ ਵਾਰ ਕਰਨਾ

3. ਢਿਡ ਵਿਚ ਮਰੋੜ ਪੈਣਾ

4. ਕੋਈ ਚੀਜ਼ ਦੇ ਬਦਲੇ ਪੈਸੇ ਜਾ ਹੋਰ ਕੋਈ ਚੀਜ ਲੈਣੀ

So, it is evident that disambiguation poses a problem in machine translation and information retrieval which calls for immediate resolution. We tried to develop an algorithm which could disambiguate such examples as precisely as possible. For that, we analyzed the classical WSD algorithm i.e Lesk algorithm.

The roadmap of the paper is as follows. Section II describes methodology used. In section III we demonstrate the experiments and results. Section IV discusses the future work and concludes the paper.

II. METHODOLOGY AND PROPOSED ALGORITHM

We begin by elaborating the working of classical Lesk algorithm. Lesk algorithm was developed by M. Lesk in 1986. According to this algorithm, all words in the sentence are compared with set of words in the dictionary. A variable is used to record the number of overlaps. The sense with maximum number of overlaps is retrieved.

Function lesk(word,sentence) returns best_sense

A. Max_overlap = 0
B. Context = words in sentence
C. For each sense in senses of word do
D. Signature = set of words in gloss
E. Overlap = compute_overlap(signature,context)
F. If overlap>max_overlap then Set Max_overlap = overlap
G. End
H. Return sense

The new algorithm is again based on same grounds of comparison with the words available in the glossary. Over and above some assumptions are made in order to retrieve more accurate results.
These are the couple of assumptions which are incorporated in the basic method-
The glossary is extended by tally up of one more column. This column contains the words which are very closely associated with the ambiguous word. Presence of these words improves the likelihood of pulling out the correct sense.
The overlapping context words nearest to ambiguous word aid in establishing the exact meaning of the ambiguous word.
Here is the flowchart of the proposed algorithm

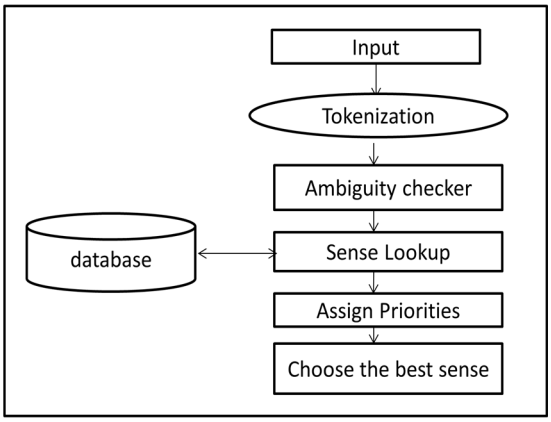


Figure 1. Flowchart of the proposed algorithm

The algorithm takes a punjabi sentence and carry out tokenization on it. The tokenization is the process in which the sentence is broken into separate lexical items. Following tokenization, ambiguous word is found in the sentence. If the sentence includes an ambiguous word then the context words are put side by side to words available in glossary. Whenever a match is found, equivalent word is dispensed priority according to its proximity to the ambiguous word. To finish off, the meaning having highest priority is retrieved.

Below is the step by step illustration of the algorithm-

1) Start
2) Input the sentence.
3) Ambiguous_word=Find_ambigous(context_dataset)
4) If ambiguous_word is not null then
5) Tokenization
   Context_dataset = individual words from sentence
6) Calculate max_overlap = overlap(context_dataset,word_gloss)
7) End if
8) Meaning = get_meaning(max_overlap)
9) Stop

The algorithm consists of 4 steps-

1) Determine ambiguous word - From the input sentence find the ambiguous word w i by looking up in the database. The following steps are executed only if the sentence carries an ambiguous word.
2) Building the context - The context context_dataset is represented by the words to the left and to the right of wi . We also adopt a particular configuration in which the context is represented by all the words that occur in the text.
3) Calculate overlap – Determine the rank of the possible meanings of the word taking into account the two assumptions i.e the words which are closer in meaing and the distance from the ambiguous word.
4) Selecting the correct meaning - After assigning ranks to all the possible meanings, the one with highest overlap is extracted.

III. SIMULATION AND RESULTS

Since Lesk method is simplest and basic method for disambiguation and considering no significant work has done in WSD for Punjabi language, we are comparing our algorithm to the Lesk algorithm. The lesk method and the proposed method are weighed up against 10 punjabi words which are regularly used in medical domain. We worked out baseline precision, baseline recall and F1 measure for these ten words. The metrics that are used to evaluate a disambiguation algorithm are the following:

Precision is the number of relevant documents a search retrieves divided by the total number of documents retrieved.

$$\text{Precision} = \frac{\#\text{ correctly disambiguated words}}{\#\text{ disambiguated words}} \quad \text{- - - - - - - - - - - Eq.1}$$

Ambigous word	Number of senses	Baseline Precision	
		Case 1	Case 2
ਸੱਟ	3	0.4132	0.4655
ਡੋਲ	3	0.5672	0.5060
ਵੱਟ	6	0.5598	0.5734
ਉਲਟੀ	2	0.3228	0.3330
ਜੋੜ	3	0.3228	0.4420

Volume 3 Issue V, May 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

ਦੌਰਾ	2	0.4600	0.4450
ਫੱਟ	3	0.6128	0.6440
ਮਾਤਾ	4	0.5004	0.5449
ਚੱਕਰ	6	0.5401	0.6311
ਖੁਰਾਕ	2	0.4072	0.4212
Average		0.4706	0.5066

TABLE 1. BASELINE PRECISION OF CASE1 AND CASE2

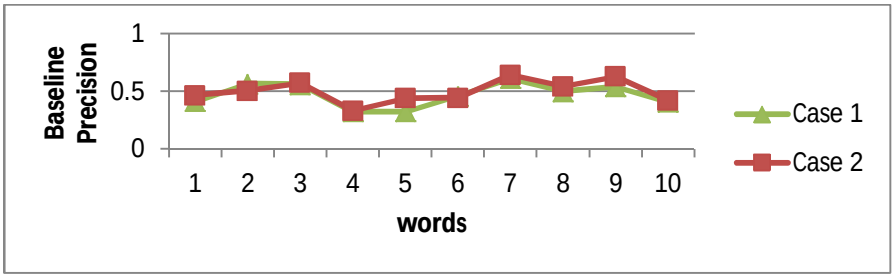


Figure 2. Baseline precision graph of case1 and case2

In the above graph and the table, case 1 represents the traditional lesk method and the case 2 represents the modified version of lesk with two assumptions. The baseline precision demonstrates the elevated bend of case 2 over case1. This is due to the reason that the proposed algorithm considers the distance and the degree of proximity of the overlapping words to the ambiguous word before assigning any sense to the word in contrast to the traditional lesk algorithm senses are assigned barely on the basis of overlapping.
Recall is the number of relevant documents retrieved divided by the total number of existing relevant documents that should have been retrieved.

$$\text{Recall} = \frac{\#\text{ correctly disambiguated words}}{\#\text{ tested set words}} \text{ - - - - - - - - - - Eq.2}$$

Ambigous word	Number of senses	Baseline Recall	
		Case 1	Case 2
ਸੱਟ	3	0.3990	0.4043
ਡੋਲ	3	0.3129	0.3652
ਵੱਟ	6	0.4021	0.4872
ਉਲਟੀ	2	0.5221	0.5392
ਜੋੜ	3	0.5476	0.5173
ਦੌਰਾ	2	0.4397	0.4180
ਫੱਟ	3	0.2649	0.3102
ਮਾਤਾ	4	0.3102	0.3978
ਚੱਕਰ	6	0.3045	0.2965
ਖੁਰਾਕ	2	0.5498	0.5163
Average		0.4052	0.4252

TABLE 2. BASELINE RECALL OF CASE1 AND CASE2

Volume 3 Issue V, May 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

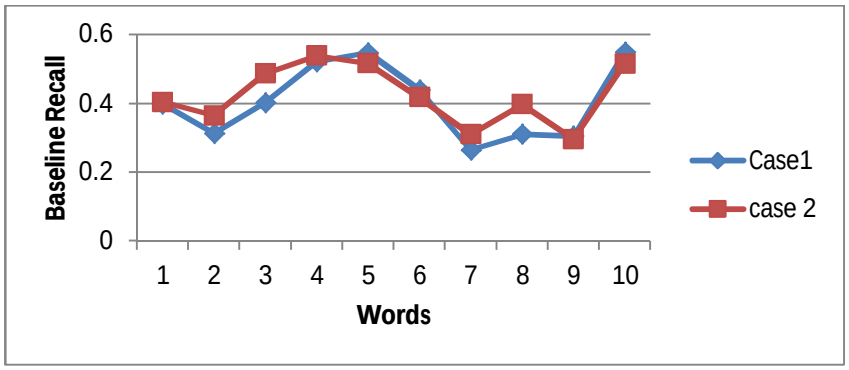


Figure 3. Baseline recall graph of case1 and case2

Recall is also a parameter for evaluation but it is not as significant as precision. In general, when precision decreases it is likely that value of recall increases. Though both are not inverse of each other, yet this trend is observed in most of the information retrieval systems.

We have appraised the algorithms against F measure as well. F measure is the combination of precision and recall and it is calculated by the following formula:

$$F1 = \frac{2 * \text{precision} * \text{recall}}{(\text{precision} + \text{recall})} \quad \text{- - - - - - - - - - - Eq.3}$$

The precision-recall graph illustrates the tradeoff made by the searching algorithm.

The F1 score lies between the value of the recall and the value of the precision, and tends to lie closer to the smaller of the two, so high values for the F1 score are only possible if both the precision and recall are large.

Ambigous word	Number of senses	F1 Measure	
		Case 1	Case 2
ਸੱਟ	3	0.4059	0.4327
ਡੋਲ	3	0.4033	0.4242
ਵੱਟ	6	0.4680	0.5267
ਉਲਟੀ	2	0.3989	0.4117
ਜੋੜ	3	0.4061	0.4766
ਦੌਰਾ	2	0.4496	0.4310
ਫੱਟ	3	0.3699	0.4187
ਮਾਤਾ	4	0.3829	0.4598
ਚੱਕਰ	6	0.3894	0.4034
ਖੁਰਾਕ	2	0.4678	0.4869

TABLE 3. F1 MEASURE OF CASE1 AND CASE2

Volume 3 Issue V, May 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

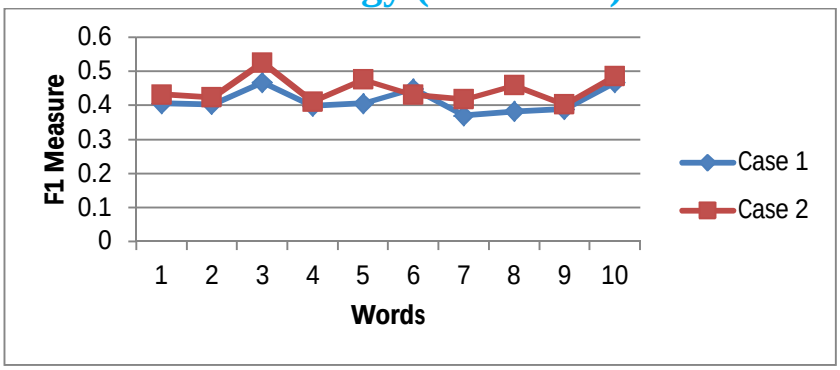


Figure 4. F1 Measure graph of case1 and case2

As shown in results the graph of case 2 algorithm is slightly mounting the case 1 graph. This is because the proposed algorithm contemplates two assumptions while disambiguating the Punjabi words. Case 2 algorithm takes into account the some words are more closely related than others and also distance of the overlapping context words matters. While opting for the correct meaning of a Punjabi word both the parameters are considered. However, in some cases there is a plunge which might be due to the reason that there are certain words which coincide for more than one polysemous word. This increases the likelihood of getting more than one meaning for a word. But, by and large the accuracy of case 2 which is 0.5066 is higher than that of case 1. This is significant improvement of 7.64% over the baseline performance of 0.4706. The overall % improvement in recall over the baseline performance is 5%. Thus taking into account that how closely an overlapping context word is related to the ambiguous word enhances the chance of obtaining the accurate meaning of the word.

IV. FUTURE WORK AND CONCLUSION

.In this paper, we presented an efficient modified version of Lesk algorithm, which provides yet another way of disambiguating the polysemous medical terms in Punjabi language. Taking into account the no significant work has been done in Punjabi language, our algorithm is just baby steps in this domain. For the realization of the this modified version we took into account two assumptions, one being that in every sentence certain words are strongly correlated to the ambiguous word as compared to others which are loosely correlated and the other the neighboring words which overlap are more significant than the distant ones. We evaluated the results of the proposed algorithm against Lesk algorithm and found that even when miniature assumptions are incorporated in the fundamental method available, it could generate better results. Parameters like precision, recall and F measure are used for assessing the functionality of the proposed algorithm. This research work is instigation of what could be done in the area of word sense disambiguation for Punjabi language. We focused our work on Punjabi medical terms. In future, others domains can be exercised. Also, it would be interesting to test our algorithm on other domains and other languages to conclusively establish the effectiveness of parameter projection for multilingual WSD. Moreover the algorithm can be open out to efficiently disambiguate sentence comprising of two or more ambiguous Punjabi words.

REFERENCES

[1] Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC '86, pages 24–26, New York, NY, USA. ACM

[2] Zhi Zhong and Hwee Tou Ng, "Word Sense Disambiguation Improves Information Retrieval", Department of Computer Science National University of Singapore 13 Computing Drive, Singapore 117417, In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 273–282, Jeju, Republic of Korea ,July 2012.

[3] Neetu Mishra, Shashi Yadav and Tanveer J. Siddiqui, "An Unsupervised Approach to Hindi Word Sense Disambiguation," IndianInstitute of Information Technology, Allahabad. UP, India, 2009.

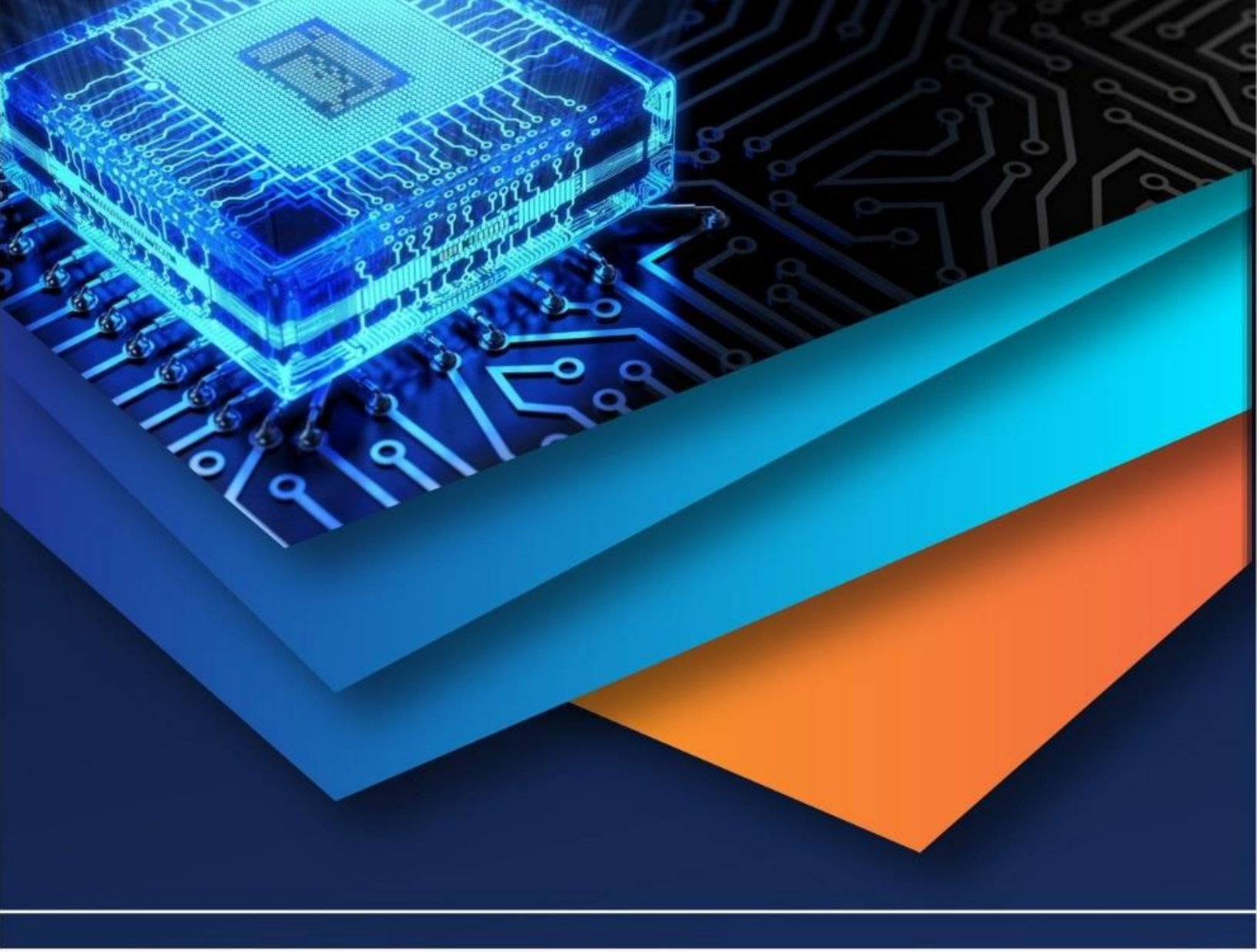
[4] Ashish Narang, R. K. Sharma, Parteek Kumar "Development of Punjabi WordNet", CSIT (December 2013) 1(4):349–354 DOI 10.1007/s40012-013-0034-0

[5] Sandeep Kumar Vishwakarma, Chanchal Kumar Vishwakarma "A Graph Based Approach to Word Sense Disambiguation for Hindi Language", IJSRET, Volume 1 Issue5 pp 313-318 August 2012

[6] Deepti Goyal,Deepika Goyal, Dr. Manjeet Singh," A Hybrid Approach to Word Sense Disambiguation", IJCST Vol. 1, Issue 2, December 2010 ISSN : 2229 - 4333 (Print)

[7] UmrinderPal Singh, Vishal Goyal, Anisha Rani," Disambiguating Hindi Words Using N-Gram Smoothing Models",An International Journal of Engineering

Sciences, Issue June 2014, Vol. 10 ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online)
[8] Parul Rastogi and Dr. S.K. Dwivedi, "Performance comparison of Word Sense Disambiguation (WSD) Algorithm on Hindi Language Supporting Search Engines", International Journal of Computer Science Issues, vol. 8, issue.2, March 2011.
[9] Gurleen Kaur Sidhu and Navjot Kaur"Role of Machine Translation and Word Sense Disambiguation in Natural Language Processing", IOSR Journal of Computer Engineering,e-ISSN: 2278-0661,p-ISSN: 2278-8727,volume 11, Issue 3,pp78-83, may jun 2013
[10] Rakesh Kumar, and Ravinder Khanna "Natural Language Engineering: The Study of Word Sense Disambiguation in Punjabi," In an International Journal of Engineering, Vol. 1, 2011
[11] S.K. Naskar and S. Bandyopadhyay "word sense disambiguation using extended Word Net," In proceedings of ICCTA'07, 2007



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY