Techniques for Lexical Semantics in Hindi Language

Mohd Zeeshan Ansari, Lubna Khan

Abstract**:** ***A word having multiple senses in a text introduces the lexical semantic task to find out which particular sense is appropriate for the given context. One such task is word sense disambiguation which refers to the identification of the most appropriate meaning of the polysemous word in a given context using computational algorithms. The language processing research in Hindi, the official language of India, and other Indian languages is constrained by non-availability of the standard corpora. For Hindi word sense disambiguation also, the large corpus is not available. In this work, we prepared the text containing new senses of certain words leading to the enrichment of the available sense-tagged Hindi corpus of sixty polysemous words. Furthermore, we analyzed two novel lexical associations for Hindi word sense disambiguation based on the contextual features of the polysemous word. The evaluation of these methods is carried out over learning algorithms and favourable results are achieved.***

Keywords**:** ***Lexical Semantics, Word Sense Disambiguation, Text Classification, Naïve Bayes.***

I. INTRODUCTION

All natural languages have certain words with multiple meanings called homonyms. The automatic selection of the appropriate meaning of such words is a challenging task in natural language processing. Human beings can easily arrive at the appropriate meaning of such word using the context in which it is used. However, the relationship between the meaning and the context is not well understood by machines because the computational representation of context is considerably difficult. When a particular word has different meanings, also called senses, pertaining to the contexts in which it is used, it is called polysemous. This characteristic involves a great deal of complexity in the processing of natural languages. In lexical semantics of natural languages, the context is closely related to the specific task, domain and underlying language. Since the words that occur in a given text may be interpreted in more than one way, the context is significant to determine its appropriate sense.

Therefore, for the automatic identification of the particular sense of a word, the context analysis is required. Consequently, the Word Sense Disambiguation (WSD) task is to determine the most appropriate sense of a word in a given context. Most of the WSD techniques consider context as the text surrounding the polysemous word, usually in a fixed size window keeping the word in the middle. Several approaches to solve WSD task for English and European languages are present in literature. Most of them are classified under three major approaches: Knowledge-based, supervised and unsupervised approaches [1, 2, 7, 8, 13, 15, 24]. Enough WSD research works based on semi-supervised and hybrid approaches is carried out, with English being the primary language.

The studied on complex language processing tasks such as machine translation, information extraction, question answering, sentiment analysis, etc. involving Indian languages, especially Hindi, the official language of India, is constrained by unavailability of large standard corpora. As WSD is involved in many such tasks, it is therefore, a challenging task for Indian languages since these are morphologically rich in nature and development of various resources like machine-readable dictionaries, WordNet, language corpora etc. are under progress [32, 35]. In this article, we present the work on Hindi written in Devanagari script, which is the official language of India. We explore the Hindi WSD task by the interpretation and analysis of the context in a variety of ways. The Sense Tagged Hindi Corpus, used in this work was developed under the Technology Development for the Indian Languages (TDIL) project, Government of India [30]. This corpus is available for research on Hindi Word Sense Disambiguation consist of polysemous words and their instances, see Table-I. We enriched this corpus by the inclusion of more senses in the case of two existing words. This enriched sense tagged Hindi corpus is used to investigate lexical and semantic attributed significantly to Hindi WSD task. Our contribution to Hindi WSD is three folds (1) propose additional sense of two words (i) 'बाल' meaning 'भुट्टे का बाल' (corn silk), and (ii) 'कदम' meaning 'कदम का पेड़' (bur flower) and their instances to the existing corpus (2) we explore two novel attribute associations for Hindi WSD and test them on a range of window size, (3) present comparative analysis of their performance with respect to the methods found in literature. For a comparative evaluation of our methods, we also constructed the attributes defined by Singh et al. [29].In this work, we investigate various attribute associations for Hindi WSD task. The feature vector is constructed using the associations of local context, collocation, bag of words after stop word removal and *vibhakti.*

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Mohd Zeeshan Ansari*, Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India. Email: mzansari@jmi.ac.in

Lubna Khan, Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India. Email: lubna.java@gmail.com

Table-I: Corpus statistics.

	Original Corpus	Enriched Corpus
Number of words	381875	383008
Number of instances	7506	7570
Number of polysemous words	60	60

These vectors of features when tested with the different window size of text produce significant results. The overall work is organized in the sections as follows. Section 2 presents related work on general WSD task, highlighting the challenges faced by researchers in WSD. Section 3 introduces Hindi WSD tasks. Section 4 gives the basic definitions of attributes of WSD task. Section 5 gives the detailed description of our work, i.e., the suggested attribute associations, the existing methods with illustrative examples. Section 6 gives detailed description of the dataset preparation, experimental setup. Section 7 gives analysis and discussion of results. Section 8 presents the conclusion and future work.

II. RELATED WORD

In lexical semantics, the word sense disambiguation task typically involves two main subtasks, one determining the different possible senses (or meanings) of each word, and second, the tagging of each word with its appropriate sense with high accuracy and efficiency. The challenges identified for the WSD task are (1) discreteness of the senses, (2) differences between dictionaries, (3) amount of samples and semantic knowledge available. Discreteness of the senses can be divided into coarse-grained and fine-grained levels. A human can easily understand coarse-grained which deals with homographs, but it is quite difficult for a human to understand fine-grained level. The WSD accuracy for English is reported around 90% for coarse-grained and 65% for fine-grained [6]. The number of samples and semantic knowledge available can be enhanced by building them manually which involves large costs. Several approaches are there for assigning the correct sense to a word in a given context, some of them achieving high accuracy figures. Initially, these methods were usually tested only on a small set of words with few and clear sense distinctions, e.g. Yarowsky [4] reports 96% precision for twelve words with only two clear sense distinctions each. Despite the wide range of approaches investigated and the large effort devoted to tackle this problem, no large-scale broad coverage and highly accurate WSD system is built. It still remains an open problem if we look at the main conclusions of the ACL SIGLEX Workshop: “Tagging Text with Lexical Semantics: Why, What and How?, WSD is perhaps the great open problem at the lexical level of NLP” [7,8,10] or to the results of the Senseval, Senseval-2 and Senseval-3 [14,16,23] in which none of the systems presented in these conferences achieved 80% accuracy on both English Lexical Sample and All Words tasks.

Table-II: Excerpts of Hindi text containing the polysemous word ‘हार’ in two senses.

Sense 1	Paragraph1. निर्माता–निर्देशक करण जौहर ने ट्विटर पर लिखा है केकेआर की **हार** से बहुत दुख हुआ। खेल को खेल भावना से देखना चाहिए और **हार** स्वीकार करनी चाहिए।
Sense 2	Paragraph 2. न्यूयॉर्क। हीरे का **हार** पहनी एक बार्बी गुडिया न्यूयॉर्क में रेकॉर्ड कीमत में नीलाम हुई है। अपनी तरह की ये अकेली बार्बी डॉल काला लिबास पहने हुई है और उसके गले में एक कैरेट का चौकोर गुलाबी हीरे का **हार** है। ये गुडिया में बनाया गया था और तबसे लेकर आज तक इसका रूप कई बार बदला है। सबसे बडी नीलामी का रेकॉर्ड बनाने वाली बार्बी गुडिया को ऑस्ट्रेलिया के एक गहनों के डिजायनर स्टीफानो कैन्टुरी ने बनाया है।

Bayesian classification models applied to several investigations in WSD achieved considerable success [22,31]. Bruce and Wiebe [11] presented a more complex model known as the decomposable model which considers different characteristics dependent to each other. The main drawback of this approach is that its enormous amount of parameters to be estimated, as they are proportional to the number of different combinations of the interdependent characteristics. Therefore, this technique requires a large number of training examples so as to appropriately estimate all the parameters. In order to solve this problem, Pederson and Rebecca [9] proposed an automatic method for identifying the optimal model, a high performance and low effort in parameter estimation, by means of the iterative modification of the complexity degree of the model. Despite its simplicity, the Naive Bayes algorithm has been the first choice to obtain state-of-the-art accuracy on supervised WSD [12,18,21]. Mooney [6] considered seven supervised learning algorithms namely Naïve Bays, perceptron, decision-tree learner, k-nearest neighbor classifier, logic-based conjunctive and dis-junctive normal form learners and a decision-list learner and tested on the ambiguous word ‘line’ having six senses and made a comparison among them, which showed that Naïve Bayes classifier and perceptron are the best methods for WSD task. The words surrounding the ambiguous word were used as features for the classifiers. Lea-cock et al. [3] used Naïve Bayes classifier and combined topic context and local context to achieve higher accuracy for the combination, which disambiguated a noun, a verb and an adjective. Pederson [12] took an ensemble of nine simple Naïve Bayes classifiers to improve WSD accuracy using nine different window sizes of left and right of context: 0,1,2,3,4,5,10,20 and 50. An accuracy of 88% and 89% was achieved on the two datasets. Le and Shimazu [19] studied Naïve Bayes classifier for performing WSD task utilizing rich features. They added features represented by ordered words in a local context and collocations using features derived from a for-ward sequential selection algorithm. Results obtained were an accuracy of 92.3% for four common test words and an accuracy of 72.7% for nouns and 66.4% for verbs when tested on large DSO corpus. Zhong and Ng developed [27] a system based on a supervised learning approach, a flexible framework that achieved good results on several Senseval and Semeval tasks. Lee et. al. [20] studied the supervised learning approach for WSD task using SVM machines and multiple knowledge sources. Results on English lexical sample task indicated that their method achieved good accuracy. Chaplot, et al. [33] developed an unsupervised WSD model using Maximum A Posteriori Inference Query built on a Markov Random Field (MRF) using WordNet and Link Parser (Stanford Parser). It is a graphical model which was tested on English all word dataset showed better and fast results compared to the existing best un-supervised models [25]. The Maximum Entropy approach [5,17] provides a flexible way to combine statistical evidence from many sources.

The estimation of probabilities assumes no prior knowledge of data and it has proven to be very robust. It is applied to many NLP problems and it also appears as a competitive alternative in WSD [28,29].

Table-III: Senses of ‘हार’

Sense	Synonyms	Meaning	Use in sentence
Sense 1	पराजय, आपजय (defeat) विघात, असफलता (failure) मात (checkmate) अभिभव (disgrace)	पराजित होने की अवस्था या भाव (defeated state or expressions)	इस चुनाव में उसकी हार निश्चित है (in this election his defeat is certain)
Sense 2	माला (garland) नेकलेस (necklace) अवतंस,अवतन्स (garland) आभूषण (jewelry)	गले में पहनने का एक प्रकार का सोने, चाँदी आदि का गहना (a type of gold or silver jewelry used for wearing around the neck)	उसने हीरे का हार पहन रखा है (she is wearing a diamond necklace)

Singh, et. al. (2015) initiated work on Hindi WSD task using Naive Bayes classifier. They considered eleven feature vectors based on local context, collocations, unordered list of noun words and *vibhakti*. Results obtained by their model show a precision of 77.52% for an unordered list of words, and a maximum precision of 86.11% for nouns words in feature vector after applying morphology. A precision of 56.49%, by incorporating vibhakti in the feature vector was also reported. The task for Hindi WSD supervised approach was further extended using sense tagged training corpus, dictionary definitions and semantic relations to assign weights to words appearing in the context of polysemous words [29]. The context of target word was defined as a list of words appearing in ± n window with the target word in the middle. For evaluation, sense tagged corpus consisting of sixty Hindi words (nouns) using sense definitions and semantic relations obtained from Hindi WordNet, were utilized. Results showed that overall average precision and recall values were 78.98% and 73.41% respectively.

III. HINDI WORD SENSE DISAMBIGUATION

The application of Word Sense Disambiguation on polysemous words present in Hindi language text written in Devanagari script is called Hindi Word Sense Disambiguation. A Hindi polysemous word does also have a different meaning in different contexts. The text excerpts in Table-II. is the text in the Hindi language which illustrates one such word ‘हार’. It can be clearly observed that there are two senses of the word ‘हार’, in the first paragraph it means ‘पराजय’ (defeat) and in the second paragraph, it means ‘माला’ (necklace). The more details of senses of word ‘हार’ is presented in Table-III.

IV. LEXICAL ATTRIBUTES FOR HINDI WSD TASK

The basic lexical attributes used in sense disambiguation are local context, collocation, bag of words, bag of words after stop words removal and *vibhakti* which are defined below. Illustrative examples# of each method are given throughout this section.

Definition 1. Local context (l_j) is defined as the collection of words surrounding the ambiguous word in a given piece of text with a window size of j. The local context feature set denoted by l_2, contain words of local context in a window size ±2, i.e. two words from the left and two words from the right of the ambiguous word. The example may be given as l_2 = [‘हीरे’, ‘का’, ‘हार’, ‘पहनी’, ‘एक’]#.

Definition 2. Collocation (c_j) is defined as the group of those sequence of words that include the target word. In a window size of j, the collocation feature set, c_j, is the group of those sequences of size 2 to j+1 words which contain the ambiguous word. The collocation feature set denoted by c_2, contain collocation in a window size ±2, i.e. word sequence of size 2 and 3 which include the ambiguous word. The example may be given as c_2 = [‘हीरे का हार ’, ‘का हार’, ‘हार पहनी’, ‘हार पहनी एक’, ‘का हार पहनी’] #.

Definition 3. Bag of words (b_j) is the simple bag of j words in the left and right of the ambiguous words in a window size ±j, but without removing the stop words from the text. The example of window size ±2 may be given as b_2 = [‘हीरे’, ‘का’, ‘हार’, ‘पहनी’, ‘एक’] #.

Definition 4. Bag of words after stop word removal (b^*_j) is the bag of j words in the left and right of the ambiguous word in a window size ±j after removing the stop words from the text. The example may be given as b^*_2 = [‘’, ‘हीरे ’, ‘हार’, ‘पहनी’] #.

Definition 5. *Vibhakti* (v_j) are set of words considered important constituent of Hindi grammar, referring to the relationship between verbs and other constituents typically nouns or pronouns in a sentence. The *vibhakti* involved in this work are, *v* = [ने(ne) , को(ko) , से(se) , के(ke) , लिए(liye) , का(ka) , की(ki) , में(mein) , पर(par) , हे(hey) , अरे(arey)] [27]. The feature set is defined by making bag of words with only *vibhakti* in the left and the right of the target word with a window size ±j. The example feature set may be given as, v_2 = [‘’, ‘का’, ‘’,‘’] #

V. LEXICAL ATTRIBUTE ASSOCIALTIONS FOR HINDI WSD TASK

In order to construct a novel set of lexical attribute associations, we defined lexical association feature set on a range of window size.

This set is obtained by a unique combination of basic attributes, viz. local context, collocation, bag of words after stop words removal and *vibhakti*. We proposed two novel attribute associations and tested them on window size range from ±2 to ±5. These are defined as follows

Table-IV: Polysemous words for Hindi WSD task.

Number of senses	Hindi Polysemous words
2	अशोक, कांड, कोटा, क्रिया, गल्ला, गुना, गुरु, ग्राम, घटना, चंदा, चारा, जीना, जेठ, डब्बा, डाक, , सोना, हल, ढाल, तान, ताव, तिल, तीर, तुलसी, दक्ष, दर, दाद, दाम, धन, धुन, माँग, लाल, विधि, शेर, सीमा, हार
3	अंग , अंश, अचल, उत्तर, कमान, कुंभ, क्वार्टर, खान, चरण, तेल, थान, फल, बाल, मत, मात्रा, वचन, वर्ग, संक्रमण, संबंध
4	कदम, कलम, धारा, मूल
5	चाल, टीका

Table-V: Precision, Recall and F-Score

Method	Window	P	R	F
Collocation + bow	5	**0.80**	**0.85**	**0.82**
Collocation + local context + vibhakti		0.74	0.80	0.77
Singh et al. [29]		0.77	0.82	0.79
Bow after stop word removed		0.76	0.81	0.78
Collocation + bow	4	0.79	0.84	0.82
Collocation + local context + vibhakti		0.74	0.80	0.77
Singh et al. [29]		0.76	0.82	0.79
Bow after stop word removed		0.75	0.81	0.78
Collocation + bow	3	0.78	0.83	0.80
Collocation + local context + vibhakti		0.74	0.80	0.77
Singh et al. [29]		0.76	0.81	0.78
Bow after stop word removed		0.74	0.80	0.77
Collocation + bow	2	0.77	0.82	0.79
Collocation + local context + vibhakti		0.76	0.81	0.78
Singh et al. [29]		0.76	0.81	0.78
Bow after stop word removed		0.70	0.76	0.73

Collocation and bag of words after stop words removal ([c+b*]j). It is the combination of local context and bag of words after stop words removal of window size j. The feature set $[c+b^*]_j$ is obtained by merging the basic features of collocation c_j and bag of words after stop word removal b^*_j of window size ±j. Therefore, the feature set $[c+b^*]_2$ is defined as the combination of c_2 and b^*_2 of window size j=±2. Using the same example text as used in previous examples, the feature set of $[c+b^*]_2$ is obtained as [‘हीरे का हार’, ‘का हार’, ‘हार पहनी’, ‘हार पहनी एक’, ‘का हार पहनी’, ‘हीरे’, ‘हार’, ‘पहनी’] $^\#$. Similarly, the feature sets of window size ±3, ±4, ±5 are also obtained.

Local context, collocation and vibhakti ([l+c+v]j). It is the combination of three basic features, local context, collocation and *vibhakti*. The feature $[l+c+v]_j$ is obtained by merging the basic features of local context l_j, collocation c_j and *vibhakti* v_j of window size ±j. Therefore, the feature $[l+c+v]_2$ is defined as the combination of feature l_2, c_2 and v_2 of window size j=±2. Using the same example text as used in previous examples, the feature set of $[l+c+v]_2$ is obtained as [‘हीरे’, ‘का’, ‘हार’, ‘पहनी’, ‘एक’, ‘हीरे का हार ’, ‘का हार’, ‘हार पहनी’, ‘हार पहनी एक’, ‘का हार पहनी’, ‘’, ‘का’, ‘’, ‘’]$^\#$. Similarly, the feature sets of different window size of ±3, ±4, ±5 are also constructed.

The attribute association suggested by Singh et. al. (2015) [29] for Hindi Word Sense Disambiguation is *local context with collocation* ($[l+c]_j$). It is a combination of local context and collocation of window size j. This set is obtained by merging the basic feature sets l_j and c_j as defined at the beginning of this section. The feature set $[l+c]_2$ consists of words in local context and collocation in window size j=±2,

which can also be obtained by merging the basic features l_2 and c_2. The example may be given as

$[l+c]_2$ = [‘हीरे’, ‘का’, ‘हार’, ‘पहनी’, ‘एक’, ‘हीरे का हार’, ‘का हार’, ‘हार पहनी’, ‘हार पहनी एक’, ‘का हार पहनी’] $^\#$. This feature combination $[l+c]_j$ was analyzed for different window size having a range from j = ±5 to ±25 [27].

VI. EXPERIMENTAL SETUP

The dataset is generated using the sense tagged Hindi corpus, also used by Singh et al. [29] which consists of sixty polysemous Hindi nouns with their senses. Table-IV. presents the complete set of all such words along with the number of senses for each word. To this corpus, we added one more sense of two words, ‘बाल’ and ‘कदम’. It was identified that the word ‘बाल has one more sense ‘भुट्टे का बाल’ (corn silk) which is not present in the corpus.

Similarly, it was also identified word 'कदम' has one more sense 'कदम का पेड़' (bur flower) which is also not present in the corpus. Subsequently, the instances of both words were collected and merged with the corpus leading to the final dataset having additional 64 instances of 1133 words more, than original corpus.

Although the corpus is organized in a defined format, we further preprocess it for our experimental settings. Finally, we divide the dataset into training and testing data according to ratio 3:1. We performed a series of experiments using the proposed as well as previously existing attribute associations and also using the basic attributes alone. All the methods are examined on the window size of range from ±2 to ±5.

VII. RESULT AND DISCUSSION

By the application of the Naïve Bayes classification algorithm, we predicted the appropriate sense of each ambiguous word. We predicted it for each of the test samples, and reported the sense accuracy as the *number of correct sense predictions* divided by *total number of test sample*. We computed the metrics such as precision, recall and F1-measure for the two proposed attribute associations as well as methods of Singh et al. [29] which are presented in Table-V. It is observed from the analysis that the proposed attribute associations for collocation and bag of words after stop word removal with window size ±5, i.e. $[c+b^*]_5$, performs best and achieves the highest precision, recall and accuracy values of 0.80, 0.85 and 0.85 respectively. Moreover, this combination method with window size in the range from ±3 to ±5 outperforms the other methods. We deduce that this combination is best suited for our problem and on increasing the window size greater than ±5 much higher performance can be achieved. The second proposed combination of local context, collocation and *vibhakti* shows poor performance than other methods because of the introduction of *vibhakti*. The reason being that there is no sense related words in it. This can be endorsed by the observation that it alone shows the worst performance. Therefore, we deduce that the *vibhakti* do not make any contribution in the performance improvement of the model when used alone as well as when used in combination, moreover it degrades the performance when used in combination. The performance of basic attributes without combination show reduced performance as compared to the proposed and existing attribute associations. The bag of words after stop words removal of window size ±5, i.e. b^*_5 is the best among the basic attributes. It is, therefore, considered most appropriate for the comparison in the analysis throughout, although, the results of local context and collocation have also been obtained from the experiments.

VIII. CONCLUSION

The present work explores the lexical semantics for the Hindi language in Devanagari script. A contribution is made by the addition one new sense each in case of two polysemous words which contains 1133 Hindi Devanagari words under 64 instances. The investigation is carried out to examine the effect of several lexical attribute associations over Hindi sense disambiguation. Realizing that the neighbouring words in the context of the ambiguous word play a vital role in feature vector formation, two novel attribute associations are proposed. These methods are successfully tested along with the existing and basic methods on the corpus of 60 polysemous words. The precision, recall and F1- measure are computed for each feature vector. One of our proposed combination method performs best as compared to rest of the methods which is observed from the figures of outcomes obtained from the experiments. The second proposed method validates the fact that *vibhakti* do not make any contribution in disambiguation of senses. A scope of examining the proposed methods on a higher range of window size is left for future.

REFERENCES

1. Duda, R. O. and Hart, P.E. (1973): Pattern Classification and Scene Analysis, John Wiley & Sons.
2. Gale, W. A., Church, K., and Yarowsky, D. (1992b): A method for disambiguating word senses in a corpus, Comput. Human, 26, 415–439.
3. Leacock, C., Geoff T. and Ellen V. (1993): Corpus based statistical sense resolution, Proceedings of the ARPA Workshop on Human Language Technology, Plainsboro, U.S.A., 260–265.
4. Yarowsky, D. (1995): Decision lists for lexical ambiguity resolution: Application to accent resolution in Spanish and French, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (Las Cruces, NM), 88-95.
5. Berger, A., Pietra, S. D., and Pietra, V. D. (1996): A Maximum Entropy Approach to Natural Language, Processing Computational Linguistics, 22(1).
6. Mooney, R. J. (1996): Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning, Proceedings Conference on Empirical Methods in Natural Language Processing (EMNLP), 82–91.
7. Ng, T. H. (1997): Getting serious about word sense disambiguation, Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (Washington D.C.), 1–7.
8. Resnik, P. and Yarowsky, D, (1997): A perspective on word sense disambiguation methods and their evaluation. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (Washington, D.C.). 79–86.
9. Pederson, T. and Rebecca, B. (1997): Distinguishing word senses in untagged text, Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, Providence, U.S.A., 197–207.
10. Leacock, C., Miller, G. A., and Chodorow, M. (1998): Using Corpus Statistics and WordNet Relations for Sense Identification, Journal of Computational Linguistics, Special issue on word sense disambiguation, 24, I, 147-165.
11. Bruce, R. F. and Wiebe, J. M. (1999): Decomposable Modeling in Natural Language Processing, Computational Linguistics, 25(2).
12. Pederson, T. (2000): A Simple Approach to Building Ensembles of Naïve Bayesian Classifier for word sense disambiguation, Proceeding of the North American Chapter of the Association for Computational Linguistics, NAACL, 63-69.
13. Manning, C.D. and Schuze, H. (2000): Introduction to Statistical Natural Language Proceesing, The MIT Press, Massachusetts.
14. Kilgarriff, A. and Rosenzweig, J. (2000): English SENSEVAL: Report and Results, Proceedings of the International Conference on Language Resources and Evaluation, LREC.
15. Escudero, G., Mμarquez, L. and Rigau, G. (2000a): A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation, Proceedings of the Computational Natural Language Learning Workshop, CoNLL.
16. Kilgarriff, A. (2001): English Lexical Sample Task Description, Proceedings of the International Workshop on Evaluating Word Sense Disambiguation Systems, Senseval-2.
17. Suarez, A. and Palomar, M. (2002): A Maximum Entropy based Word Sense Disambiguation System, Proceedings International Conference on Computational Linguistics, COLING.

18. Banerjee, S., and Pedersen, T., (2003): Extended gloss overlaps as a measure of semantic relatedness; In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI, Acapulco, Mexico). 805-810.
19. Le C. A. and Shimazu, A. (2004): High WSD accuracy using Naïve Bayesian classifier with rich features, PACLIC 18, Waseda University, Tokyo, 105, 113.
20. Lee, Y. K., Ng, H. T., and Chia, T. K. (2004): Supervised Word Sense Disambiguation with Support Vector Machine and Multiple Knowledge Sources, Proceeding of SENSEVAL-3 : Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics,137-140.
21. Yuret, D. (2004): Some Experiments with a Naive Bayes WSD System, Proceedings of the International Workshop on Evaluating Word Sense Disambiguation Systems, Senseval3.
22. Suarez, A. (2004): Resolucion de la ambigauedad semantica de las palabras mediante modelos de probabilidad de maxima entrop³a, PhD thesis, Departamento de Lenguajes y Sistemas Informaticos, Universidad de Alicante.
23. Mihalcea, R. Chklovski, T. and Kilgarriff, A. (2004): The Senseval-3 English lexical sample task, Proceedings of SENSEVAL-3, the third international workshop on the evaluation of systems for the semantic analysis of text.
24. Rada Mihalcea, Courtney Corley, Carlo Strapparava (2006): Corpus-based and Knowledge-based measures of text semantic similarity; AAAI, Volume 6, 775-780.
25. Navigli, R. and Lapata, M. (2007): Graph Connectivity Measures for Unsupervised Word Sense Disambiguation, Proceedings of the 20th International Joint Conference on Artificial Intelligence Hyderabad, India, 1683–1688.
26. Naseer, A. and Sharmad, H. (2010): Supervised Word Sense Disambiguation for Urdu using Bayesian Classification, Proceeding of Conference on Language & Technology (CLT10).
27. Zhong, Z. and Ng, H. T., (2010): It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text, Proceeding of the ACL, System Demonstration, pp8-83, Uppsala, Sweden.
28. Singh, S., Siddiqui, T. J., (2013): A supervised algorithm for Hindi Word Sense Disambiguation, International Journal of Systems, Algorithms & Algorithms, Volume 3.
29. Singh, S., Siddiqui, T. J., Sharma, S. K., (2014): Naïve Bayes (NB) classifier for Hindi Word Sense Disambiguation (WSD), Proceedings of the 7th ACM India Computing Conference (Compute 14) , http://dx.doi.org/10.
30. Sense Annotated Hindi Corpus: Indian Language Technology Proliferation & DevelopmentCenter,tdil-de.in/index.php
31. Francisco, V. J., (2014): Word sense disambiguation through associative dictionaries, PhD thesis, Instituto Politecnico Nacional, Mexico, D. F.
32. Ansari, M.Z., Ahmad, T., Ali, M. A., (2018): Cross Script Hindi English NER Corpus from Wikipedia in Proceedings of International Conference on Intelligent Data and Internet of Things ICICI 2018.
33. Chaplot, D. S., Bhattacharyya, P and Paranjape, A. (2015): Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser; Proceedings of the 29 AAAI Conference on Artificial Intelligence.
34. Hadni, M., Alaoui, S. E., and Lachkar, A., (2016): Word Sense Disambiguation for Arabic Text Categorization, International Arab Journal of Information Technology, Vol. 13, No. 1A.
35. Bhattacharyya P. (2017) IndoWordNet. In: Dash N., Bhattacharyya P., Pawar J. (eds) The WordNet in Indian Languages. Springer, Singapore.

AUTHORS PROFILE

Mohd Zeeshan Ansari is currently Assistant Professor at Department of Computer Engineering, Jamia Millia Islamia (A Central University), New Delhi. He received M.Tech in Computer Science and Engineering from Delhi Technological University, New Delhi and B.Tech in Computer Science and Engineering from Uttar Pradesh Technical University, Lucknow, Uttar Pradesh. He has more than twelve years of teaching experience. His area of research is Code Mixing, Information Extraction and Retrieval, Text Mining, Natural Language Processing and Soft Computing Techniques. His field of Specialization is Information Extraction and Retrieval. He has active participation in several national and international workshops, seminars and conferences. He has also published articles in international conferences and refreed journals.

Lubna Khan has completed B.E. and M.Tech Computer Engineering from Department of Computer Engineering, Jamia Millia Islamia (A Central University), New Delhi. She has undertaken several research oriented projects at Jamia Millia Islamia. Her research includes Natural Language Processing, Text Mining, and Word Sense Disambiguation. She is a python expert and actively involved in NLP research.