A Naïve Bayes Approach for working on Gurmukhi Word Sense Disambiguation

Himdweep Walia[1], Ajay Rana[2], Vineet Kansal[3]

[1]*GNIOT, Gr. Noida, Uttar Pradesh, India*
[2]*Amity University Uttar Pradesh, Noida, India*
[3]*CSED, IET, Lucknow, Uttar Pradesh, India*

himdweep@yahoo.com, ajay_rana@amity.edu, dir.engg@its.edu.in

Abstract **– Natural Language Processing is a technique which allows communication between the human and the machine. In this technique the major problem has been Word Sense Disambiguation (WSD). WSD is the process of uniquely identifying the correct usage of the given word, of the multiple meanings that the word may have. A lot of work is going on in this field, especially in English and European Languages. In recent years, significant work has been done in Indian Regional Languages also. Punjabi is an Indian Regional Language and Gurmukhi is its script. The WSD applies three approaches – knowledge based, corpus based and hybrid approach. The corpus based approach can be further divided into - supervised and unsupervised approach. Off the many algorithms implemented under supervised approach, Naive Bayes Approach has shown higher accuracy in WSD. For this paper we have used the Punjabi Corpora (obtained from Evaluations and Language Resources Distribution Agency, Paris, France) which has been sense-tagged with 100 words.**

Keywords **– Word Sense Disambiguation, Supervised Approach, Naïve Bayes Classifier, WSD in Punjabi, Sense-tagged corpora**

I. INTRODUCTION

The study of Word Sense Disambiguation (WSD) is very significant in Natural Language Processing as it forms the base for the different usages in this domain of Artificial Intelligence, like, machine translation, speech recognition, natural language understanding and generation. The word "ambiguous" means having more than one meaning. Every Natural Language has multiple ambiguous words. It is easy for the human to understand that in which context the word is being used but that may not be the case for the machines. To explain the concept of WSD, let us look at the following sentences:

1. The little boy knocked the door.

2. The little boy knocked himself in the door.

As we can see that the word "knocked" is used in the two sentences but the context in which the word is being used is entirely different. The meaning of "knocked" in the first sentence is to strike a surface noisily, whereas in the second sentence it implies, collision.

In a similar fashion, all the languages have this sort of ambiguity which is difficult for the machine to detect easily. In our language of study, Punjabi, we too have many ambiguous words. Consider the word, ਉੱਤਰ .

Context 1: ਰਾਮ ਪੌੜੀਆਂ ਉੱਤਰ ਰਿਹਾ ਹੈ |

Context 2: ਮੈਂ ਪ੍ਰਸ਼ਨ ਦਾ ਉੱਤਰ ਸੋਚ ਕੇ ਲਿਖਿਆ |

Context 3: ਸੂਰਜ ਉੱਤਰ ਵਲ ਡੂਬ ਗਿਆ |

Context 4: ਮਹਿੰਦਰ ਡਿਗ ਪਿਆ ਤੇ ਉਸ ਦਾ ਘੁਟਣਾ ਉੱਤਰ ਗਿਆ |

The word " ਉੱਤਰ ", is common in all the four sentences and convey different meanings in all of the above. It is easier for a human to differentiate that in which context the word is being used. For example, in the first context the given usage means "climbing down", in the second context it means "answer", in the third context it means "direction" and in the fourth context it means "dislocation of bone" [25].

The different techniques that are used in WSD have been categorized into three main segments i.e. knowledge-based technique, corpus-based technique and hybrid technique. The knowledge-based technique uses the concept of dictionary i.e. Machine Readable Dictionary or thesauri. The corpus-based technique is further categorized into two i.e. supervised learning techniques and unsupervised learning techniques. The supervised techniques are better if the language has rich resource and also these techniques give more promising results on implementation. In unsupervised techniques, the concept of sense-tagged text is not required. The hybrid technique mixes all the above techniques. In this paper, we are using one of the most successful algorithms, in terms of results, under the supervised approach – Naive Bayes for Gurmukhi Word Sense Disambiguation. For the training and testing, we are using the Punjabi Corpora (obtained from Evaluations and Language Resources Distribution Agency, Paris, France) which has been sense-tagged with 100 words. This paper has six sections. Section I gives a brief introduction about WSD in general and also illustrates examples of WSD in Gurmukhi. The rest of the sections are structured as: Section II discusses the related work

978-1-5090-3012-5/17/$31.00 ©2017 IEEE

in other languages. Section III explains the Naïve Bayes Algorithm and Bayesian Classifier in WSD. Section IV discusses the execution and implementation of the system. Section V shows the results and Section VI concludes the paper.

II. PREVIOUS WORK

Commendable work has been done in Word Sense Disambiguation for English and other European Languages [8]. In recent years, a lot of work has been carried out in Indian Regional Languages as well [1]. Our language of focus is Gurmukhi, more popularly known as Punjabi [7]. The major amount of work in this language has been limited to machine translation and building of machine readable dictionaries due to lack of sense-tagged corpora [4, 12]. For this paper, we have studied the work done in Hindi Language as Gurmukhi drives closely to this language.

In their paper on Hindi WSD, S. Singh and T. Siddiqui [20] have discussed three algorithms based on corpus statistics. The algorithms have shown good results and the authors have remarked that if a sense annotated corpus is available then they can pre-compute the conditional probability of the co-occurring words. In their paper, C. A. Le and A. Shimazu [9] have worked on English Language and have used the Forward Sequential Selection approach which has shown 92.3% accuracy.

In the paper by S. Singh and T. Siddiqui [18], the authors have worked with the Bayesian Classifier disambiguating the words in Hindi Language. In their paper, they have identified 11 different features like nouns, collocations, pronouns, prepositions, etc. In this paper, the authors have sense-coded 60 polysemous Hindi nouns in the Hindi corpora. After applying the Bayesian Classifier on the same, they observed a precision of 77.52% on unordered list of words in feature vector and on the same corpora the results were improved upto 86.11% by utilizing nouns in feature vector.

III. NAÏVE BAYES THEOREM AND BAYESIAN CLASSIFIER

The Bayesian Classifier is a statistical classifier. This classifier helps in calculating the probability of a given sample, say M, belonging to the given set or class, say N. This is known as the class membership probability. Bayesian classification has been derived from the Bayes Theorem. The main property of this classifier is that, for a given attribute, say A, and the effect that the value of A has on the given class, say N, is independent of the values of other attributes belonging to N. This property of the Bayesian Classifier is known as the class conditional independence.

Let A be the given data sample which is taken from an unknown class, say N.

Let H be a hypothesis that A belongs to class M for classification problems.

Calculate P(H/A): the probability that the hypothesis holds given the observed data sample A.

P(H): prior probability of hypothesis H (i.e. the initial probability before we observe any data, reflects the background knowledge)

P(A): probability that sample data is observed

P(A|H) : probability of observing the sample A, given that the hypothesis holds

Given training data A, posteriori probability of a hypothesis H, P(H|A) follows the Bayes theorem:

$$P(H \mid A) = (P(A \mid H).P(H)) \,/\, P(A)$$

The Naive Bayes Theorem works on the assumption that the attributes which are used for the descriptions, for the given set or class, are all conditionally independent; although it does states two important observations:

- The structure and linear ordering of words within the context are ignored, leading to a so-called “bag of words model”.
- The presence of one word in the bag is independent of another, which is not possible in the case of Natural Languages.

The concept of Bayesian Classifier is used in Word Sense Disambiguation. The concept is implemented by looking into the cluster of words around the given word, which is ambiguous in nature. This clustering often helps in determining the correct context in which the word is being used. It is then, that the supervised training of our Bayesian Classifier is used on the sense-tagged corpora to find the context closest to the actual usage of the given word.

IV. EXECUTION AND IMPLEMENTATION

The algorithm used in this paper has been proposed by N. T. T. Aung, K. M. Soe, N. L. Thein [24]. Their proposed system consists of four parts, namely, preprocessing, multi-sense look-up, calculating probability based on Bayes Theorem, and, disambiguation i.e. calculating maximum scores using Bayes decision rule.

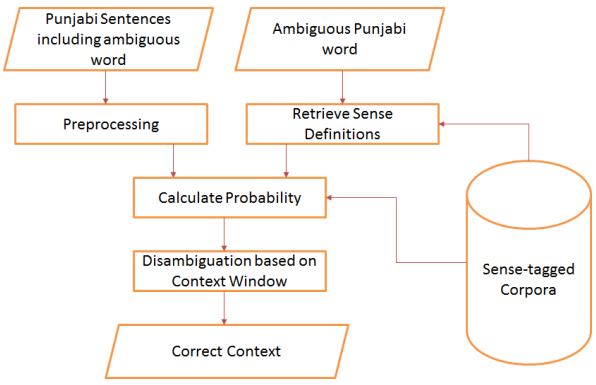

FIGURE-1 APPROACH FLOWCHART

We have implemented the approach for Gurmukhi WSD. We take any sentence from the corpora. In preprocessing stage, it generates tokens of the given sentence. This process is carried out manually as of now. The idea is to identify the stop words i.e. prepositions, pronouns, conjunctions, etc. and consequently removing them. After gathering this information in the preprocessing step, the system then uses the rest of the words from the given sentence as the input, commonly referred as "bag of words". The system then makes a grouping of these words depending on the size off the context window that we are experimenting with. This window then helps in determining the possible context in which the given word is being used from the sense-tagged corpus. In our case study we have sense-tagged 100 Punjabi nouns. The system then calculates the prior probability and the likelihood based on Bayes Theorem. Finally, disambiguation process is performed using Bayes decision rule. The theorem helps in calculating the probability of nearness to the context in which the word has been used. The higher the calculated value, the closer it is to the approximation to the context in which it is being used.

The process of finding the correct context of the given word depends largely on the group of words that we are associating with the said word. This is known as the window and the number of words that we decide will help in determining the right context is known as the window size. In this paper, we took two window sizes – 5 and 7, i.e. we made a grouping of 5 words in the first instance and then the grouping of 7 words in the second instance in order to try to determine the correct context.

V. RESULTS AND DISCUSSION

In this paper, we have used the corpora obtained from Evaluations and Language Resources Distribution Agency, Paris, France, which has been sense-tagged with 100 Punjabi nouns. We took three words, "waar", "kacha", and "uttar" in order to find that how accurately we are able to find the correct reference of the words. The first word has 4 defined meanings, while the second word has 3 and the third words also has 4 different meanings.

The given words were tested multiple times with respect to the different meanings that they have. The reason for doing this is to check if the system is able to rightly pick the correct meaning of the given word in the given context window size.

The idea of having two different size of windows was to determine how would that effect in reaching the correct conclusion.

The results calculated are shown below in Table – 1.

TABLE-1 CALCULATED RESULTS

Ambiguous Word	Number of Senses of the word	Number of instances tested	Disambiguated correctly (Window size 5)	Disambiguated correctly (Window size 7)
waar	4	167	152 (91.01%)	165 (98.80%)
kacha	3	131	129 (98.47%)	131 (100%)
uttar	4	206	198 (96.12%)	201 (97.57%)

VI. CONCLUSION

Based on the observations that we had for the proposed system we have come to the following conclusion:

- With lesser number of senses to be disambiguated, the chances of it being disambiguated correctly are more.
- By increasing the size of the context window, the accuracy with which the correct context is found is also increased.

ACKNOWLEDGEMENTS

The author acknowledges the Management of Greater Noida Institute of Technology, Gr. Noida, which gave permission to pursue doctorate and have also provided necessary help in that regard.

The author also acknowledges Amity University, Uttar Pradesh, which gave the opportunity to pursue doctorate and has provided library facility and extensive online resources.

REFERENCES

[1] M. Bansal, "Word Sense Disambiguation: Literature Survey for Indian Languages", International Journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277 128X), Volume 5, Issue 12, December 2015.

[2] D. Jurafsky, J. H. Martin, "Naive Bayes Classifier Approach to Word Sense Disambiguation", Chapter 20 Computational Lexical Semantics, available at (http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/presentations/Olango-Naive-Bayes-2009.pdf), last seen 2015.

[3] J. Kaur, "Word Sense Disambiguation (WSD)", International Journal For Technological Research In Engineering, Volume 1, Issue 5, January-2014 ISSN (Online) : 2347 - 4718

[4] R. Kaur, R. K. Sharma, S. Preet, P. Kumar, "Punjabi WordNet Relations and Categorization of Synsets", available at (http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/12_iwn_Punjabi%20WordNet%20Relations%20and%20Categorization%20of%20Synsets.pdf), last visited 2016.

[5] M. M. Khapra, P. Bhattacharyya, S. Chauhan, S. Nair, A. Sharma, "Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting", available at Research Gate (https://www.researchgate.net/publication/228981276), last visited 2013.

[6] M. M. Khapra, S. Shah, P. Kedia, P. Bhattacharyya, "Projecting Parameters for Multilingual Word Sense Disambiguation", Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapore, 2009.

[7] R. Kumar, R. Khanna, "Natural Language Engineering: The Study of Word Sense Disambiguation in Punjabi", IJES (ISSN: 2229-6913), July, 2011.

[8] R. Kumar, R. Khanna, V. Goyal, "A Review of Literature on Word Sense Disambiguation", International Journal of Engineering Sciences, ISSN: 2229-6913, Vol. 6, July, 2012.

[9] C. A. Le, A. Shimazu, "High WSD accuracy using Naive Bayesian classifier with rich features", Proceedings of PACLIC 18, Tokyo, Japan, pp.105-113, 2004.

[10] Y. K. Lee, H. T. Ng, T. K. Chia, "Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources", 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July 2004.

[11] A. Montoyo, A. Suarez, G. Rigau, M. Palomar, "Combining Knowledge and Corpus-based Word-Sense Disambiguation Methods", Journal of Artificial Intelligence Research, March, 2005.

[12] A. Narang, R. K. Sharma, P. Kumar, "Development of Punjabi WordNet", Springer, CSIT, December 2013.

[13] R. Navigli, "Word Sense Disambiguation: A Survey", ACM Computing Surveys, Vol. 41, no.2, Article 10, February, 2009.

[14] P. Rana, P. Kumar, "Word Sense Disambiguation for Punjabi Language using Overlap Based Approach", Chapter in Advances in Intelligent Informatics, Volume 320 of the series Advances in Intelligent Systems and Computing, pp 607-619, Springer.

[15] J. Singh, I. Singh, "Word Sense Disambiguation: Enhanced Lesk Approach in Punjabi Language", International Journal of Computer Applications (0975-8887), Volume 129 – No.6, November 2015.

[16] S. Singh, T. J. Siddiqui, "Role of Karaka relations in Hindi Word Sense Disambiguation", Journal of Information Technology Research (JITR), IGI Global, Volume 8, Issue 3, 2015, pages 21 - 42.

[17] S. Singh, T. J. Siddiqui, "Role of Semantic Relations in Hindi Word Sense Disambiguation", Proceedings of the International Conference on Information and Communication Technologies (ICICT 2014), Kochi, India, 3-5 December, 2014, Elsevier Procedia Computer Science, Volume 46, 2015, pages 240-248.

[18] S. Singh, T. J. Siddiqui, S. K. Sharma, "Naïve Bayes classifier for Hindi Word Sense Disambiguation", Proceedings of 7th ACM India Compute Conference (Compute'14), Nagpur, India, 9 – 11 October, 2014, Article No. 1, ACM Digital Library.

[19] S. Singh, V. K. Singh, T. J. Siddiqui, "Hindi Word Sense Disambiguation using Semantic Relatedness measure", Proceedings of 7th Multi-Disciplinary Workshop on Artificial Intelligence (MIWAI 2013), 9-11 Dec. 2013, Krabi, Thailand, pages 247-256, LNCS, Springer.

[20] S. Singh, T. J. Siddiqui, "Utilizing Corpus Statistics for Hindi Word Sense Disambiguation", International Arab Journal of Information Technology (IAJIT), Volume 12, No. 6A, December 2015, pages 755 - 763 (SCI Expanded, Impact Factor 0.582).

[21] S. Singh, T. J. Siddiqui, "Evaluating Effect of Context Window Size, Stemming and Stop Word Removal on Hindi Word Sense Disambiguation" , Proceedings of the International Conference on Information Retrieval & Knowledge Management, CAMP2012, 13-15 March, 2012, Malaysia, pages 1-5, IEEE Explorer.

[22] F. Vasilescu, P. Langlais, G. Lapalme, "Evaluating Variants of the Lesk Approach for Disambiguating Words", available at www.iro.umontreal.ca/~felipe/Papers/paper-lrec-2004.pdf, last visited 2012.

[23] T. Pedersen, "Unsupervised Corpus-based Methods for WSD", Chapter - Word Sense Disambiguation, Volume 33 of the series Text, Speech and Language, pp 133-166, Springer.

[24] N. T. T. Aung, K. M. Soe, N. L. Thein, "A Word Sense Disambiguation System Using Naïve Bayesian Algorithm for Myanmar Language", International Journal of Scientific & Engineering Research, Volume 2, Issue 9, September-2011 1 ISSN 2229-5518.

[25] H. Walia, A. Rana, V. Kansal, "Different Techniques Implemented in Gurumukhi Word Sense Disambiguation", International Journal of Advanced Technology in Engineering and Science, Volume 05, Issue 06, June 2017 ISSN 2348-7550.

[26] J. Kaur, V. Gupta, "Effective Approaches for extraction of Keywords", International Journal of Computer Science, Issue 6, Vol. 7, November 2010, ISSN 1694-0814.