Word Sense Disambiguation for Marathi language using Supervised Learning

Rasika Ransing
Datta Meghe College of Engineering, Mumbai
Vidyalankar Institute of Technology, Mumbai
rasikaransing275@gmail.com

Dr. Archana Gulati
School of Business Management, NMIMS University, Mumbai
archana.gulati@nmims.edu

Abstract

The task of disambiguating word senses, often referred to as Word Sense Disambiguation (WSD), is a substantial difficulty in the realm of natural language processing. Marathi is widely acknowledged as a language that has a relatively restricted range of resources. Consequently, there has been a paucity of academic research undertaken on the Marathi language. There has been little research conducted on supervised learning for Marathi Word Sense Disambiguation (WSD) mostly owing to the scarcity of sense-annotated corpora. This work aims to construct a sense-annotated corpus for the Marathi language and further use supervised learning classifiers, such as Naïve Bayes, Support Vector Machine, Random Forest, and Logistic Regression, to disambiguate polysemous words in Marathi. The performance of these classifiers is evaluated.

1 Introduction

Word Sense Disambiguation (WSD) is the process of determining the exact interpretation of a polysemous term in a given context (Pal et al., 2021). Many Natural Language Processing (NLP) applications employ WSD, either directly or indirectly. Sentiment Analysis, Machine Translation, Information Retrieval, Text summarization, etc. are some of the applications in NLP where WSD is employed. Many researchers have surveyed the various methods for Word Sense Disambiguation in various languages (Tatar, 2005; Zhou and Han, 2005; Navigli, 2009; Pal and Saha, 2015; Bevilacqua et al., 2021; Ransing and Gulati, 2022).

The existing methods for word sense disambiguation are generally categorized into two distinct groups (Bevilacqua et al., 2021; Navigli, 2009):

- The corpus-based technique that utilizes a dataset to extract features that convey linguistic information pertaining to the contextual aspects of each phrase. Corpus-based methods may be further classified into the following techniques:
  - Supervised techniques that need a corpora with sense annotations.
  - Unsupervised methodologies that do not need corpora with sense annotations.
  - Semi-supervised techniques may be characterized as a blend of supervised and unsupervised approaches. The researchers use a limited quantity of sense-tagged corpora with a substantial quantity of unlabeled corpora.
- Knowledge-based methodologies that employ lexical resources such as ontologies, machine-readable dictionaries, and thesauri.

Marathi is an Indo-Aryan language mostly used in the state of Maharashtra, India. Nevertheless, the progress in constructing Natural Language Processing (NLP) systems for Marathi has been rather restricted when juxtaposed with prominent languages like English. The insufficient focus on Marathi language processing presents difficulties and impedes the development of sophisticated Natural Language Processing (NLP) applications for those who speak Marathi (Lahoti et al., 2022).

Supervised methodologies for Word Sense Disambiguation (WSD) have shown superior accuracy in comparison to other techniques. Nevertheless, one drawback of these methods is their reliance on corpora that have been annotated with sense information. Several commonly used supervised learning methods include Naïve Bayes, Support Vector Machine, Decision Tree, etc. (Navigli, 2009).

In this paper, we present the development of a sense-annotated corpus and the use of several supervised machine learning methods, including Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest and Logistic Regression, for the

purpose of disambiguating words having multiple meanings in the Marathi language. The performance of these algorithms for Word Sense Disambiguation (WSD) is assessed. This study represents the first use of these algorithms for Marathi Word Sense Disambiguation (WSD) and the subsequent evaluation of their performance.

The subsequent sections of this work are organized as follows: Section 2 provides an overview of the existing literature and research that is relevant to the topic at hand. In Section 3, the implementation of supervised learning algorithms is shown. In the next section 4, an analysis of the acquired findings is presented. Section 5 provides a conclusion to the material presented and offers recommendations for future research endeavours.

2 Related Work

Kumari et al. examine several supervised machine learning algorithms used for the purpose of disambiguating the proper interpretation in Hindi text. They have constructed a sense-tagged corpus by selecting 20,000 phrases from news websites like as Danik Bhaskar and Navbharat Times. Three classifiers, including the Naïve Bayes classifier, Support Vector Machine (SVM) based classifier, and Multiple Layer Perceptron (MLP) based classifier, have been used on the defined feature sets. Upon conducting a comparison between the judgments made by the classifier and those made by human annotators, they concluded that the SVM Classifier had superior values for Accuracy, Recall, Precision, and F-Score (Kumar et al., 2016).

Faisal et al. propose the use of Support Vector Machine (SVM) algorithm in conjunction with the Term Frequency-Inverse Document Frequency (TF-IDF) technique as the feature extraction method, and utilize Wikipedia as the training data, to address the Word Sense Disambiguation (WSD) challenge in the Indonesian language. Initially, the training data is sourced from Wikipedia articles. The articles are then subjected to pre-processing techniques aimed at reducing word variation. Subsequently, the articles are transformed into classifiable features using TFIDF, followed by the utilization of a Support Vector Machines classifier to ascertain the meaning of an unclear word inside a phrase. The suggested technique has an accuracy level of 87.7% across all classifiers (Faisal et al., 2018).

Walia et al. have used a supervised methodology, namely the k-Nearest Neighbors (k-NN) algorithm, to resolve word ambiguities in the Gurmukhi language. They have used the Punjabi Corpora, sourced from the Evaluations and Language Resources Distribution Agency in Paris, France. The corpora has been sense-tagged with 100 words. The algorithm's efficiency was evaluated by conducting tests on a collection of 120 phrases, each containing 8 confusing terms. The k-NN based method to Word Sense Disambiguation (WSD) in Gurmukhi demonstrated an average accuracy ranging from 53% to 76% (Walia et al., 2018a).

Pal et al. propose to use supervised approach for the purpose of Word Sense Disambiguation (WSD) in Bengali, with appropriate adjustments. Their study is conducted using four widely used supervised approaches, namely the Decision Tree (DT), the Support Vector Machine (SVM), the Artificial Neural Network (ANN), and the Naïve Bayes (NB) algorithm, for the purpose of sense categorization in the baseline experiment. The aforementioned methods are independently used on a dataset consisting of the 13 most often utilized ambiguous terms in the Bengali language. The aforementioned approaches provide accuracy results of 63.84%, 76.9%, 76.23%, and 80.23% respectively (Pal et al., 2019).

Pal et al. use a supervised technique to address the challenge of word meaning disambiguation in the Bangla language. The Naive Bayes probabilistic model is often used as a baseline approach for sense classification. When applied to a database including the 19 most frequently used ambiguous terms in the Bangla language, it achieves a reasonable level of accuracy, namely 81%. Here, Pal et al. provide two modifications to the baseline approach. Firstly, they include a lemmatization process into the system. Secondly, they use a bootstrapping technique to enhance the operational process. Consequently, the approach exhibits a modest improvement in accuracy, reaching a level of 84% precision (Pal et al., 2018).

Walia et al. present the implementation of three supervised approaches, namely Naïve Bayes, k-NN, and Decision Trees classifiers, for the purpose of word sense disambiguation in the Punjabi language. To provide a comparative analysis of the three supervised approaches, a set of 20 ambiguous terms has been chosen. Additionally, the researchers have conducted experiments using three distinct window context sizes, namely 3, 5, and 7, in order to evaluate the efficacy of each of the

strategies. The findings of the study suggest that increasing the window size leads to improved accuracy. Additionally, the findings indicate that among the three supervised procedures, Naïve Bayes exhibits superior performance compared to the other two techniques (Walia et al., 2018b).

Singh and Kumar have conducted an analysis of a Punjabi language Word Sense Disambiguation (WSD) system that employs supervised approaches. A manually produced corpus consisting of 150 ambiguous Punjabi noun terms has been created. This study examines six supervised machine learning algorithms, namely Decision List, Decision Tree, Naive Bayes, K-Nearest Neighbour (K-NN), Random Forest, and Support Vector Machines (SVM). The word embedding features have been tested on six classifiers for the Punjabi Word Sense Disambiguation (WSD) challenge. The use of word embedding features has been shown to improve the performance of supervised classifiers in the task of Word Sense Disambiguation (WSD) for the Punjabi language. The present study demonstrates that the LSTM classifier, when using word embedding feature, has reached an accuracy rate of 84% (Singh and Kumar, 2019).

Naseer and Hussain employ a statistical method known as "Bayesian Classification" to address the issue of Word Sense Disambiguation for certain Urdu terms. This study examines four Urdu words, consisting of one noun and three verbs, as its primary emphasis. Their experiment involves using various window widths where n represents the number of words to the left and right of the ambiguous word, with n being equal to 3, 5, and 7, respectively. The method demonstrated superior performance when applied to a specific term that had a significantly high frequency within the corpus. However, the accuracy of the algorithm was considerably lower when dealing with words that occurred less often. The augmentation of window size yielded improved performance outcomes. The 7x7 window configuration yielded the highest accuracy (98.35%) and recall (92.17%) values (Naseer and Hussain, 2009).

Lee et al. provide a description of the participating systems in the SENSEVAL-3 English lexical sample task and multilingual lexical sample task, as presented by the authors. The Word Sense Disambiguation (WSD) systems used Support Vector Machine (SVM) learning techniques and included many knowledge sources. The features used in this approach include the Part-of-Speech (POS) of neighboring words, individual words within the surrounding context, local collocations, and syntactic relations. The omission of feature selection is justified based on the findings of their earlier study (Lee and Ng, 2002), which demonstrated that SVM achieves superior performance without the inclusion of feature selection. The findings of this inquiry demonstrate that the incorporation of four pre-existing knowledge sources resulted in an enhancement of the micro-averaged recall performance on the training data, increasing from 0.628 to 0.638 (Lee et al., 2004).

Abid et al. address the issue of word sense disambiguation (WSD) within the specific linguistic framework of the Urdu language. In this study, the authors use machine learning (ML) techniques, including the Bayes net classifier (BN), support vector machine (SVM), and decision tree (DT), to perform word sense disambiguation (WSD) on Urdu literature written in its original script. The findings indicate that Bayesian Networks (BN) exhibit a higher F-measure compared to Support Vector Machines (SVM) and Decision Trees (DT). The Bayes net classifier achieved a maximum F-measure of 0.711 when applied to a raw Urdu corpus consisting of 2.5 million words (Abid et al., 2018).

Junaida et al. present a hybrid methodology that combines a multi-class Support Vector Machine (SVM) with a corpus-based technique for the purpose of Malayalam word sense disambiguation. The training set consists of a restricted number of ambiguous words that have been annotated with 16 Word Sense Disambiguation (WSD) classes. The system undergoes evaluation using manually generated words, and its correctness is assessed by n-fold cross-validation. The findings from the 10-fold cross-validation demonstrate the suitability of the suggested multi-class Support Vector Machine (SVM) for the Malayalam word sense tagger. The evaluation was conducted using a one-against-one technique, using both word-only and word plus part-of-speech (POS) features. The findings indicate that the one versus one technique yields the highest performance outcomes, with an overall average accuracy, recall, and F-measure values of 63.058, 57.78, and 57.9, respectively, for the 10-folds (Junaida et al., 2015).

Singh et al. examine the use of the Naive Bayes (NB) classifier in Hindi Word Sense Disambiguation (WSD). The research focuses on the utilization of eleven distinct characteristics, including local context, collocations, unordered list of words,

nouns, and vibhaktis. The evaluation process was conducted on a manually constructed sense annotated corpus in the Hindi language. This corpus included 60 polysemous nouns in Hindi. A accuracy of 77.52% was found while using an unordered list of words in the feature vector. The use of nouns in the feature vector, after the application of morphology, resulted in an accuracy rate of 86.11%. A accuracy of 56.49% was achieved by including vibhaktis in the feature vector (Singh et al., 2016).

Sarmah et al. aim to offer a supervised machine learning strategy, namely Decision Tree, for the job of Word Sense Disambiguation for Assamese language. The training and test dataset consisted of a collection of polysemous terms in Assamese language, each having various genuine occurrences and manual sense annotation. The DT method yielded an average F-measure of 0.611 when a 10-fold cross-validation assessment was conducted on a set of 10 Assamese ambiguous phrases (Sarmah and Sarma, 2016).

Parameswarappa and Narayana provide a study that focuses on the use of compound words hint and syntactic properties within a limited context for the purpose of target word sense disambiguation for Kannada language. The Kannada Shallow parser has been used for the purpose of doing syntactic analysis. The process of resolving the ambiguity of the target word is accomplished by using a Naive Bayes classifier. The accuracy values of the models vary between about 51% and 71% when evaluated on typical corpora. Their system handles only one target word (Parameswarappa and Narayana, 2011).

Borah et al. have devised an automated method for Word Sense Disambiguation (WSD) in the Assamese language, using a Naive Bayes classifier. The Assamese language has a high degree of morphological complexity. The used features include Unigram Co-Occurrences (UCO), POS of Target Word (POST), POS of Next Word Feature, and Local Collocation. The system achieved optimal performance, with an F1-measure of 86%, when all four characteristics were aggregated (Borah et al., 2014).

Gopal et al. introduced a method for Malayalam word sense disambiguation that employs a Naive Bayes classifier inside a supervised framework. This framework primarily utilizes two corpora known as the sense corpus and the ambiguous corpus. An ambiguous corpus encompasses a comprehensive collection of words that have many alternative meanings, whereas a sense corpus has synsets, synonyms, and antonyms. The findings demonstrate a 90% level of accuracy when using the Naïve Bayes classifier algorithm on a corpus consisting of 100,000 words (Gopal and Haroon, 2016).

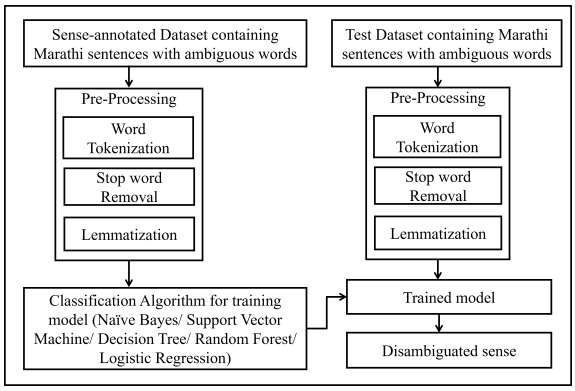

Figure 1: Implementation of Supervised Machine Learning algorithm for Marathi WSD

3 Implementation

Based on the survey of work done for word sense disambiguation of various languages in the previous section 2, we propose to use supervised learning approaches for word sense disambiguation of Marathi language. We have used baseline supervised learning classifiers like Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest and Logistic Regression. This is the first attempt in our knowledge to implement these supervised algorithms for Marathi Word Sense Disambiguation and comparison of their performance. The proposed system is shown in Figure 1.

The input to this system is a sense-annotated dataset containing Marathi sentences with ambiguous words. The sentences in the sense-annotated dataset are pre-processed by removing the special symbols, removing stopwords and performing lemmatization. Five separate models are trained on the sense-annotated dataset using algorithms like Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest and Logistic Regression. The sentences in test dataset are also pre-processed. Then the models are tested separately on a few test sentences to determine the most appropriate sense of the ambiguous word.

A collection of 650 sense-annotated Marathi sentences has been compiled, which have been derived from several Marathi websites and publications, including different genres such as news, philosophy,

sports, and fiction. Some sentences have also been generated manually by us. These sentences were carefully selected to include 12 terms that possess multiple meanings, resulting in a total of around 42 distinct senses. The set of ambiguous terms under consideration comprises 2 verbs, 2 adverbs, 2 adjectives, and 6 nouns. The current dataset has around 15-16 sentences per sense.

The training set consists of 500 sense-annotated Marathi sentences. The performance of these supervised learning algorithms is evaluated on the remaining 150 test sentences.

4 Results and Discussions

In this section, we present and discuss the results of the application of the supervised algorithms for Marathi Word Sense Disambiguation. The results obtained are illustrated in the following Table1 and Figure 2.

Table 1: Evaluation of various Supervised Learning algorithms for Marathi WSD.

	Precision	**Recall**	**Accuracy**	**F-1 score**
Naive Bayes	**0.39**	**0.53**	**0.53**	**0.43**
Support Vector Machine	0.36	0.53	0.53	0.4
Decision Tree	0.01	0.1	0.11	0.02
Random Forest	0.00	0.06	0.05	0.00
Logistic Regression	0.37	0.47	0.47	0.41

It is observed that the accuracy of Naive Bayes and Support Vector Machine (SVM) is the same but the precision and F-1 score of Naive Bayes is better than that of SVM. Logistic Regression gives average results as compared to Naive Bayes and SVM. Decision Tree and Random Forest perform poorly for this task of Marathi Word Sense Disambiguation on the test dataset.

5 Conclusion

Our study encompasses a comprehensive analysis of the literature surveyed, which together assert that the use of supervised machine algorithms

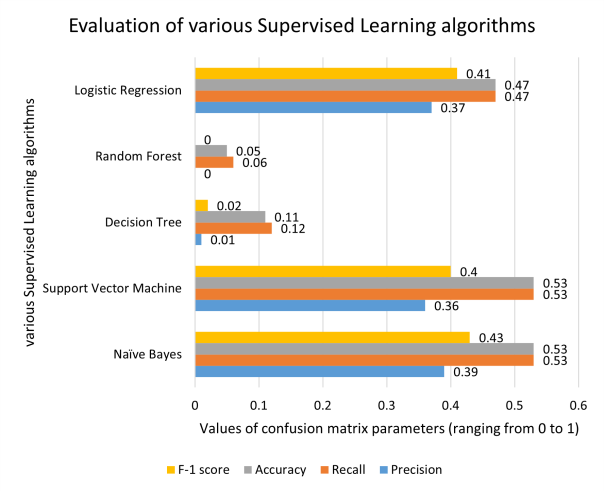

Figure 2: Comparison of various Supervised learning algorithms on our dataset

is viable for Marathi Word Sense Disambiguation (WSD). Additionally, researchers have shown that the use of sense-annotated datasets might potentially enhance the accuracy of disambiguating Marathi words. This study focuses on the implementation of several supervised machine learning methods, including Naive Bayes, Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and Logistic Regression (LR), for the purpose of Marathi Word Sense Disambiguation. Our study revealed that Naive Bayes (NB) exhibit superior performance on our test dataset. However, the Precision, Recall, and F1-score metrics for all the implemented algorithms do not meet the desired standards due to the limited quantity of the training and testing datasets. Supervised learning algorithms may exhibit superior performance relative to unsupervised and knowledge-based techniques when the size of the dataset is expanded.

References

Muhammad Abid, Asad Habib, Jawad Ashraf, and Abdul Shahid. 2018. Urdu word sense disambiguation using machine learning approach. *Cluster Computing*, 21:515–522.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Pranjal Protim Borah, Gitimoni Talukdar, and Arup Baruah. 2014. Assamese word sense disambiguation using supervised learning. In *2014 International*

Conference on Contemporary Computing and Informatics (IC3I), pages 946–950. IEEE.

Edi Faisal, Farza Nurifan, and Riyanarto Sarno. 2018. Word sense disambiguation in bahasa indonesia using svm. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 239–243. IEEE.

Sreelakshmi Gopal and Rosna P Haroon. 2016. Malayalam word sense disambiguation using naïve bayes classifier. In *2016 International Conference on Advances in Human Machine Interaction (HMI)*, pages 1–4. IEEE.

MK Junaida, Jisha P Jayan, and Sherly Elizabeth. 2015. Malayalam word sense disambiguation using yamcha. In *2015 International Conference on Computing and Network Communications (CoCoNet)*, pages 720–724. IEEE.

Ajai Kumar, Iti Mathur, Hemant Darbari, GN Purohit, and Nisheeth Joshi. 2016. Implications of supervised learning on word sense disambiguation for hindi. In *Proceedings of the second international conference on information and communication technology for competitive strategies*, pages 1–5.

Pawan Lahoti, Namita Mittal, and Girdhari Singh. 2022. A survey on nlp resources, tools, and techniques for marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–34.

Yoong Keok Lee and Hwee Tou Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 41–48.

Yoong Keok Lee, Hwee Tou Ng, and Tee Kiah Chia. 2004. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 137–140.

Asma Naseer and Sarmad Hussain. 2009. Supervised word sense disambiguation for urdu using bayesian classification. *Center for Research in Urdu Language Processing, Lahore, Pakistan*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Alok Ranjan Pal and Diganta Saha. 2015. Word sense disambiguation: A survey. *arXiv preprint arXiv:1508.01346*.

Alok Ranjan Pal, Diganta Saha, Niladri Sekhar Dash, Sudip Kumar Naskar, and Antara Pal. 2019. A novel approach to word sense disambiguation in bengali language using supervised methodology. *Sādhanā*, 44:1–12.

Alok Ranjan Pal, Diganta Saha, Niladri Sekhar Dash, and Antara Pal. 2018. Word sense disambiguation in bangla language using supervised methodology with necessary modifications. *Journal of The Institution of Engineers (India): Series B*, 99:519–526.

Alok Ranjan Pal, Diganta Saha, Sudip Kumar Naskar, and Niladri Sekhar Dash. 2021. In search of a suitable method for disambiguation of word senses in bengali. *International Journal of Speech Technology*, 24:439–454.

S Parameswarappa and VN Narayana. 2011. Target word sense disambiguation system for kannada language. In *3rd International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2011)*, pages 269–273. IET.

Rasika Ransing and Archana Gulati. 2022. A survey of different approaches for word sense disambiguation. In *ICT Analysis and Applications: Proceedings of ICT4SD 2022*, pages 435–445. Springer.

Jumi Sarmah and Shikhar Kr Sarma. 2016. Decision tree based supervised word sense disambiguation for assamese. *Int. J. Comput. Appl*, 141(1):42–48.

Satyendr Singh, Tanveer J Siddiqui, and Sunil K Sharma. 2016. Naïve bayes classifier for hindi word sense disambiguation. In *Proceedings of the 7th ACM India computing conference*, pages 1–Assamese Word Sense Disambiguation –48.

Varinder Pal Singh and Parteek Kumar. 2019. Sense disambiguation for punjabi language using supervised machine learning techniques. *Sādhanā*, 44:1–15.

Doina Tatar. 2005. Word sense disambiguation by machine learning approach: A short survey. *Fundamenta Informaticae*, 64(1-4):433–442.

Himdweep Walia, Ajay Rana, and Vineet Kansal. 2018a. A supervised approach on gurmukhi word sense disambiguation using k-nn method. In *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 743–746. IEEE.

Himdweep Walia, Ajay Rana, and Vineet Kansal. 2018b. Word sense disambiguation: Supervised program interpretation methodology for punjabi language. In *2018 7th international conference on reliability, infocom technologies and optimization (Trends and future directions)(ICRITO)*, pages 762–767. IEEE.

Xiaohua Zhou and Hyoil Han. 2005. Survey of word sense disambiguation approaches. In *FLAIRS conference*, pages 307–313. Citeseer.