NAIVE BAYES CLASSIFIER FOR WORD SENSE DISAMBIGUATION OF PUNJABI LANGUAGE

Varinder pal Singh, Parteek Kumar

Computer Science and Engineering Department, Thapar University,
Patiala 147004, Punjab, India

Email : varinderpal@thapar.edu, parteek.bhatia@thapar.edu

DOI: https://doi.org/10.22452/mjcs.vol31no3.2

ABSTRACT

Word Sense Disambiguation (WSD) is the process of identifying the correct sense of the word in the context. The most leading scheme used by WSD is machine learning approach, where a human expert provides examples of correctly disambiguated words, and a machine learning algorithm is used to induce a model from these examples. In this paper, Naive Bayes supervised classifier has been used to disambiguate words of Punjabi language. The feature extraction process plays a vital role in building the supervised machine learning models. For the proposed Punjabi WSD system, Bag of Words (BoW) and collocation models are used separately to extract relevant features. BoW model has used all words around target word while collocation model has used two words before and two words after the target word as features. Both the models have used a common training data set to build the model. It has been observed that the selection of smoothing parameter for Naive Bayes has a significant impact on its performance. This proposed work has been tested on 150 most ambiguous noun words selected form Punjabi WordNet having 6 or more senses. During the process of building the model, fine senses of ambiguous words have been merged to produce coarse sense on the basis of manual analysis of lexical relations of WordNet. The accuracy of the proposed system has been calculated independently for BoW and collocation model. The proposed WSD system achieves an accuracy of 89% for BoW model and 81% for collocation model. It has been concluded that BoW model performs better than the collocation model for WSD task for Punjabi language.

Keywords: Word sense disambiguation, Bag of words model, Collocation model, Naive Bayes classifier.

1.0 INTRODUCTION

Natural languages often have words which have more than one meaning (senses) in the context. These ambiguities of a language are easily interpreted by humans from the context. For example, in sentence "*deposit money in the bank*", the ambiguous word "*bank*" can be categorized into "*financial institute*" sense and in the sentence "*he sat on the bank of the river*", the word bank refers to *the slope of land at the side of a river*" sense. The identification of the right sense of a word in the given context by a computer is a complex task. For Natural Language Processing (NLP) applications, Word Sense Disambiguation (WSD) is considered as an important transitional step. WSD is defined as a task of automatically assigning a sense to an ambiguous word in the context from a given set of senses. For an automatic WSD system, computers need a knowledge base (machine-readable dictionaries, semantic networks, *etc.*) that will help the computer to disambiguate the different senses of the words. Collection of knowledge resources for a machine is an expensive, time consuming and repetitive task. Further, the granularity of sense inventories (knowledge resources) account for the hardness of the WSD task. This is a fundamental problem which pervades the field of WSD [1]. Hence, WSD problem is treated as an artificial intelligence complete complexity problem [2, 3].

In various applications of NLP like Machine Translation, Information Retrieval, Information Extraction, Text Mining, Word Processing, Lexicography, Content Analysis, Semantic Web, Speech Processing, Social Network Analysis [2,4,5] *etc.,* the accuracies are hindered due to the presence of ambiguous words. In this case, WSD plays a vital role for reducing such ambiguities. Hence, there is a need to develop a robust WSD system, which can be used to find the right meaning of the word in the context to remove its ambiguity.

There are three different approaches for WSD task namely, knowledge-based, supervised, and unsupervised approaches. The knowledge-based approach uses the lexical resources like dictionary, thesauri, ontologies, and WordNet to yield the right sense of the target word in the context [2]. Supervised approaches use labeled training data to learn and classify an ambiguous word. While in case of unsupervised approaches, the raw corpus is used to extract the feature vectors of the target word in context without any training dataset. The knowledge-based approach is scalable and does not require sense annotated data for WSD task. However, the precision of supervised approach is better than the knowledge-based approach. The supervised approach uses the previously seen sense annotated examples to perform WSD task and usually outperforms unsupervised approaches. However, supervised approaches need manually sense annotated training data, which is time-consuming and expensive. The availability of sense tagged corpus and sense repositories have encouraged the application of supervised technique for English language WSD tasks [6].

1.1 Background

Research work for WSD task using different machine learning approaches for the English language has been established in last 20 years. At the initial stage, the sense-tagged corpus of words '*line', 'interest', 'hard'*, and '*serve'* was developed and this corpus was used by the researchers to explore the different supervised approaches. Monney [7] compared five supervised techniques using sense-tagged dataset of '*line'* and '*interest'* for WSD task and reported the Naive Bayes (NB) as the best technique. An ensemble model of multiple NB classifiers by varying context window size was explored by Pedersen [8] for the disambiguation of words '*line'* and '*interest'*. The Naive Bayesian classifier was applied on the corpus of '*interest'*, '*line'*, '*hard'* and 'serve' words for disambiguation by Le and Shimazu [9] and it had yielded an accuracy of 92.3%. However, they also reported the accuracy of 66.4% for verbs and 72.7% for nouns when tested on Defense Science Organization (DSO) large corpus. Florian *et al.* [10] evaluated the combinations of well-established (NB, Cosine, Decision List) and novel (Transformation-based Learning, MMVC) classifiers on the data sets of four languages English, Spanish, Swedish and Basque and they had found that individually NB and Bayes ratios classifiers performed best in all the four datasets of languages.

Non-availability of standard sense-tagged corpuses has made the comparison and evaluation of different WSD system difficult. Senseval/Semeval [11] competitions used WordNets to produce a standardized sense tagged corpus for different languages. In these competitions, different supervised approaches were successfully applied by researchers for WSD task for Basque, Chinese, Czech, Danish, Dutch, English, Estonian, Italian, Japanese, Korean, Spanish, Swedish languages *etc*. and they generated the state of art accuracies [12, 13]. The NB classifier was the most common classifier for different WSD systems that had participated in Senseval/Semeval competitions [14] and it had also been effectively applied on Japanese, Chinese [15], and Indo-Aryan [16] WSD systems.

NB classifier was evaluated using a different set of parameters and was compared with different machine learning methods. Yarowsky and Florian [17] investigated the performance of Naive Bayes and five other supervised algorithms on thirteen different set of parameters. Features like collocation knowledge, word order in local context helped in improving the accuracy of NB classifier [9]. NB classifier used keywords, labels, and neighboring words as features for the classification of Portuguese noun [18]. Liu [19] concluded that the local features were more than sufficient to disambiguate the words of Chinese language using NB classifier. NB classifier for Hindi language WSD task by Singh *et al.* [20] used eleven features such as local context, collocations, nouns, vibhaktis and unordered list of words on 60 Hindi polysemous noun words. Borah *et al.* [21] described the use of NB classifier for disambiguation of Assamese language by using unigram co-occurrences, local collocation, part of speech of target word and part of speech of next word as features. Parameswarappa and Narayana [22] exploited the compound words clue and syntactic features in a local context, for Kannad language words disambiguation using NB classifier. It has been observed that WSD research on different languages has been concerned with the experimental comparison and performance of different classification methods. The researchers have found that NB machine learning classifier is the most successful and widely used supervised classifier. Its efficiency and ability to combine evidence from a large number of features is the main reason for its frequent usage in WSD tasks for different languages.

The little work has been reported for WSD of Punjabi language. Jason and Lehal [23] used the *n*-gram model for the WSD in Punjabi language and shown that lower order *n*-gram models were sufficient for sense disambiguation than the larger *n*-gram model. The use of knowledge base Lesk overlap approach for WSD of Punjabi language has been implemented by Rana and Kumar [24]. There is a dearth of sense annotated corpus for Punjabi language and no work has been reported in WSD task of Punjabi language using supervised NB classifier.

1.2 Need for proposed WSD system

NLP tasks for the Punjabi language are in the nascent stage. Currently, no WSD system exists to disambiguate the ambiguous words for the Punjabi language. Therefore, there is a need to develop an explicit WSD system, which can be incorporated into various NLP tasks of Punjabi language like machine translation, information retrieval, name entity tagging, question answering, and *etc.* to significantly enhance their accuracies. The availability of lexical resource Punjabi WordNet has motivated us to disambiguate Punjabi noun words by applying NB supervised learning technique for WSD task of Punjabi language.

For the proposed system, the Naive Bayes classifier was used to disambiguate 150 most ambiguous Punjabi noun words having more than 6 senses. These words and their fine senses have been extracted from the Punjabi WordNet. Coarse senses of selected Punjabi ambiguous words are formed manually by merging fine senses which have overlapping lexical relations defined in Punjabi WordNet. The sense annotated corpus is manually prepared by extracting data from various Punjabi online and offline resources. Feature extraction is an important phase in the supervised machine learning approaches. For this proposed work, the Bag of Words (BoW) and collocation models were used to extract the relevant features. The BoW model uses all words around the target word as features, while the collocation model trains on the annotated corpus with two words around the target word as its features. The accuracy of the proposed system was analyzed by using both approaches of the feature extraction (BoW and collocation) models individually.

The rest of the paper is organized as follows: Section 2 of the paper describes Naive Bayes classifier used for WSD task. Section 3 introduces the Punjabi language and describes its lexical resource. Disambiguation models for the Punjabi language are defined in Section 4. The proposed WSD system using both feature extraction models (BoW and collocation) is explained in Section 5. Experimental results are discussed in Section 6 and Section 7 concludes the paper.

2.0 NAIVE BAYES CLASSIFIER

NB, also known as the statistical classifier is based on the Bayes theorem [2]. In equation (1) the Bayes theorem determines the class of hypothesis H_i in a given evidence (or sample) X.

$$P(H_i|X) = \frac{P(H_i)P(X|H_i)}{P(X)} \quad \ldots(1)$$

The prior probability $P(H_i)$ is the independent occurrence of hypothesis H_i in the data set. The class conditional probability or likelihood probability $P(X|H_i)$, calculates the occurrence H_i for the specific class attribute in the evidence X. Lastly, $P(X)$ is an unconditional prior probability of the evidence X. The prior probability $P(X)$ is independent of H_i and it is not required to be known. Because $P(X|H_i)$ is independent of $P(X)$ and thus $P(X)$ is ignored in NB calculations.

In the context of WSD, let there be m senses of an ambiguous Target Word (TW) in the training dataset. The m senses $s_1 \ldots s_m$ has been extracted from a standard sense inventory, *i.e.,* WordNet. WSD using NB is formulated in equation (2) to determine the sense s_i of TW. It determines the probability of s_i sense of an ambiguous word TW with a given set of contextual features $C = c_1 \ldots c_n$ of the evidence.

$$P(s_i|C) = argmaxP(s_i)\prod_{j=1}^{n} P(c_j|s_i) \quad \ldots(2)$$

The *argmax* [2] method is used to maximize the product of prior probability $P(s_i)$ of sense s_i with likelihood probability estimation $P(c_j|s_i)$ in the evidence C. Prior probability $P(s_i)$ is the count of sense s_i (Ns_i) of an ambiguous word TW to the total count (N_{tw}) of different senses of an ambiguous word TW as shown in equation (3).

$$P(s_i) = \frac{Ns_i}{N_{tw}} \quad \ldots(3)$$

Equation (4) shows the likelihood probability estimation. The probability $P(c_j|s_i)$ is the frequency Nc_j,s_i of a context feature c_j occurring with the sense s_i in the evidence C.

$$P(c_j|s_i) = \frac{N_{cj,si}}{N_C} \qquad \ldots(4)$$

Naive Bayes theorem assumes independence of features among each other. Each contextual feature c_j independently contributes to probability estimation. The position of the features with respect to target word does not matter. The likelihood probability is mostly zero because of the sparse presence of features in training dataset. This anomaly is overridden using Laplace smoothing, *i.e.*, small quantity is added in numerator and denominator of likelihood probability [8].

3.0 PUNJABI LANGUAGE AND ITS LEXICAL RESOURCE

Punjabi is an Indo-Aryan language and is used in Punjab region of South Asia with approximately 118 million people spoke this language [25]. This makes it the 10th most spoken language in the world [26]. The Punjabi language can be written in four scripts Shahmukhi, Gurmukhi, Devnagri and LaNDA [26]. Gurmukhi script is widely used in offline as well as online documents. The work presented in this paper is based on this script of Punjabi language.

Availability of computational resources like the Punjabi WordNet in Gurmukhi script has introduced research in WSD task for Punjabi language [27]. WordNet is a collection of words along with their senses in a database. This resource explores the rich features of a language along with the examples, concept, synonyms and different relations among the words. Synonymous words are grouped together to form synonym sets called synsets. In WordNet, each synset represents a single distinct sense or concept. Synsets are interlinked based on semantic and lexical relations. There are total 32334 synsets in Punjabi WordNet that are classified into different POS classes [28]. The details of Punjabi WordNet are given in Table 1.

Table 1: Details of Punjabi WordNet

Noun	Verb	Adjective	Adverb	Total
23225	2836	5830	443	32334

For this proposed system, Punjabi WordNet outlines 150 most ambiguous noun words along with their fine senses. The glosses and example sentences of these selected words constitutes the sense-tagged corpus.

4.0 DISAMBIGUATION MODELS FOR PUNJABI LANGUAGE

In this paper, two independent models of feature extraction, *i.e.*, BoW and collocation model, based on NB classifier are used to classify an ambiguous Punjabi word in the context. These models classify the ambiguous words with the help of distinctly observed parameters from the given context of the target word. The set of unique parameters that are used to classify the different senses of an ambiguous word in the given context are known as features. Directly captured features from text or instance are known as lexical features. The lexical features used in the proposed WSD system are the words around the target word. NB classifier has been trained on a different set of lexical features to generate individual models. In BoW model, the NB classifier uses all words around the target ambiguous word as feature set. While only two words before and two words after the ambiguous target word have been used as the feature set in collocation model. The difference between BoW and collocation models on different parameters has been described in Table 2.

Table 2: Differences between BoW and Collocation models of feature extraction

Parameter	BoW Model	Collocation Model
Feature Set	It includes all words around target ambiguous word as feature set.	Two words right and two words left of the target word from the feature set in collocation model.
Feature Scope	The feature set of this model consists of all words of the context. These words give general intent of the target word. Feature scope for this model is topical.	Only a few vicinity words constitute feature set of this model. The scope of feature for this model is local.
Window Size	Window size determines the words to be extracted from context. This model uses wider window size and the whole instance is used in the window for feature extraction.	The narrow window size is used by this model to extract the local features. Window size for local features is ±*n* neighboring words of the target word.
Sparsity	The absence of test instance features from extracted features from training instances leads to sparsity. Sparsity exists in this model.	As compared to BoW model, the problem of the sparsity is severe in collocation model.
Computation and Training time	BoW model is computation intensive, needs more memory and takes more training time to build the model.	This model is less computation intensive, requires less memory and takes less training time to build the model.
Results	This model uses all words as the feature set and it gives better results for WSD.	Local features used by collocation model have less significance for WSD.

5.0 WORKING OF THE PROPOSED WSD SYSTEM

The complete process of disambiguation of ambiguous Punjabi words has been illustrated in Fig. 1. In the proposed system, the target word has been eliminated from the training and test datasets and no stemming has been performed on the data sets. In feature extraction phase, all words from the training dataset have been used to train model in BoW approach, while only local features are extracted in collocation model. The complete description of the different components of the proposed WSD system is as follows.

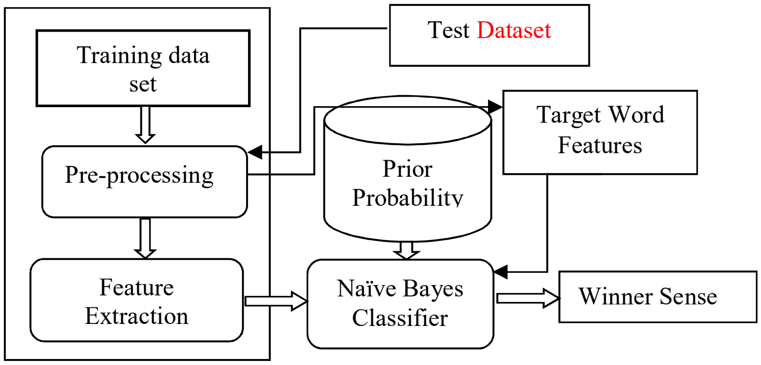

Fig. 1: Process of disambiguating an ambiguous word

5.1 Dataset preparation

Dataset used for this experiment has been prepared manually by extracting the concepts and examples of ambiguous words from Punjabi WordNet and from other resources like online newspapers, blog articles, and Punjabi Wikipedia text. From this corpus, 66% of the instances have been used for training of the system and the rest of approximately 33% has been used for testing purpose. The training data set is a collection of text files that consists of sense annotated context of 150 most ambiguous noun words selected from Punjabi WordNet. For each sense of an ambiguous word, independent text files have been created. Each text file contains the complete sentence which includes target word. There is 6104 sense tagged instances in the training data set. The test dataset of 150 ambiguous words have been randomly selected from the training dataset. The text file of the test data set of an ambiguous word consists of instances of each sense of the word without any label.

To understand the working of proposed WSD system, let us consider, an ambiguous target word ਜੱਗ (jagg) having three senses stored in three different sense files. These sense files contain 68, 35 and 38 instances of target word ਜੱਗ (jagg) respectively as shown in Table 3. There are total 141 instances of the target word ਜੱਗ (jagg) in the training dataset.

Table 3: Description of sense files of word ਜੱਗ (jagg)

Senses of ਜੱਗ(jagg)	Gloss	Target word count	Prior probability
ਸੰਸਾਰ, saṃsār, world	everything that exists anywhere	68	68/141
ਭਾਂਡਾ, bhāṇḍā, vessel	an object used as a container	35	35/141
ਰੀਤ, rīt, ritual	the prescribed procedure for conducting religious ceremonies	38	38/141
Total number of instances for a given word		141	

To disambiguate an ambiguous word ਜੱਗ (jagg) in test instance given in Table 4, data pre-processing has been performed as described in next sub-section.

Table 4: Test sentence of ਜੱਗ (jagg)

Sentence	ਰੱਬ ਨੇ ਇਸ ਜੱਗ ਵਿਚ ਭਾਂਤ ਭਾਂਤ ਦੇ ਜੀਵ ਬਣਾਏ ਹਨ।
Transliteration	rabb nē is jagg vic bhānt bhānt dē jīv baṇāē han.
Translation	God has created a variety of living beings in this world.

5.2 Data pre-processing

Pre-processing of training data set removes the punctuation marks along with the sentence ending symbols. This process also eliminates target word from every instance of training dataset. After pre-processing of training data set of target word ਜੱਗ (jagg), the count of words in its sense files SF1, SF2 and SF3 are 1039, 733 and 763 respectively. Similarly, test data is also pre-processed to remove punctuation marks and other sentence ending symbols from it.

5.3 Feature extraction

The feature extraction process transforms the pre-processed text into a set of features that are further examined in the classification process. Lexical features that are used in this proposed system are bag-of-words and collocations. The process of feature extraction for both models has been explained by using the data sets of ambiguous word ਜੱਗ (jagg).

5.3.1 BoW model

BoW model is built upon the set of all words surrounding the target word in the training data set. This feature set ignores the exact ordering of words in the training dataset but it considers the number of occurrence of ~~the~~ each word. Each distinct word in the training data set therefore constitutes an individual feature of the target ambiguous word. BoW model of an example instance containing ambiguous word ਜੱਗ (jagg) consists of all words present in its three pre-processed sense files. Features of the test instance given in Table 4 are all the words of the example instance except the target ambiguous word ਜੱਗ (jagg). The test features are overlapped with the BoW model features and the frequency of occurrence of the words is shown in Table 5.

Table 5: BoW features of test word ਜੱਗ (jagg)

Feature		Count of features in the dataset		
Name	Value	SF1	SF2	SF3
c_0	ਰੱਬ , rabb	5	0	2
c_1	ਨੇ , nē	9	15	6
c_2	ਇਸ, is	8	1	4
c_3	ਵਿਚ, vic	10	9	16
c_4	ਭਾਂਤ , bhānt	0	0	0
c_5	ਭਾਂਤ , bhānt	0	0	0
c_6	ਦੇ, dē	19	15	14
c_7	ਜੀਵ , jīv	0	0	0
c_8	ਬਣਾਏ, baṇāē	0	0	0
c_9	ਹਨ, han	6	8	0

5.3.2 Collocation model

Collocation model limits the context to a window of two words, *i.e.,* two words immediate prior and two words immediate after the target word and these words are considered as features for further processing. Collocation model for example sentence given in Table 4 is trained on all local collocation features that are determined from each sense tagged instance of the training dataset. The local companions for the target ambiguous word ਜੱਗ (jagg) are extracted and their frequency in the training data set is computed as shown in Table 6.

Table 6: Collocation features of test word ਜੱਗ (jagg)

Feature		Count of feature in data set		
Name	Value	SF1	SF2	SF3
w_{+1}	ਵਿਚ, vic	3	0	0
w_{+2}	ਭਾਂਤ, bhānt	1	0	1
w_{-1}	ਇਸ , is	3	2	2
w_{-2}	ਨੇ, nē	0	0	0

5.4 NB classifier

NB classifier requires evidence and knowledge for estimating the probabilities of different senses of an ambiguous word. The feature extraction (BoW and collocation) models of the proposed WSD system provide the necessary knowledge needed by NB classifier. The evidence or prior probability is the ratio of frequency of one sense of target word to the total count of all senses of target word from training dataset. This classifier has been applied independently on the two different feature extraction models.

5.4.1 NB classifier using BoW model

Identification of the winner sense is the final objective of this experiment. The count of feature vectors of test data observed in the training data set of the target word ਜੱਗ (jagg) is shown in Table 5. The zero count of features is called data sparseness problem. This is corrected by a method called smoothing. The smoothing parameter used for BoW model is 1, *i.e.*, 1 is added to numerator and denominator when 0 occurs in likelihood probability calculations as shown in equation (5). The variable 65536, *i.e.*, the size of vocabulary is added in denominator of equation (5).

$$P(S_1)P(c_0|s)\ P(c_1|s)\ldots\ P(c_8|s)\ P(c_9|s)= \qquad \ldots(5)$$
$$68/141\ (5+1)(9+1)\ldots(0+1)(6+1)/\ (1039+65536)^{10}$$

The training data set of sense S1 has six features of the test feature vector while the other four have no match as shown in Table 5. Substituting the count of the features in the equation (5) gives the probability of the sense S1 of an ambiguous target word ਜੱਗ (jagg). Similarly, the probabilities of senses S2 and S3 of the target word ਜੱਗ (jagg) has been calculated by substituting the values from Table 5 into the equation (5). Table 7 shows the calculated probabilities percentage of the senses S1, S2, S3 for the target word ਜੱਗ (jagg) in the test sentence. The sense S1 is selected as winner sense which has the highest probability for the target word in test sentence.

Table 7: Estimated probabilities using BoW model

S1	S2	S3
88.69	0.47	10.83

5.4.2 NB classifier using collocation model

Local collocation features have been used by NB classifier to identify the right sense of the target word ਜੱਗ (jagg). The four features are extracted from the test sentence, two words on right and two words on left of the target word. Table 6 shows that the words w_{-1} and w_{-2} are present on left of the target word while the words w_{+1} and w_{+2} are present on right side of the target word. These features are compared with the training dataset. The data sparseness is also persistent in this model. To alleviate data sparseness problem the smoothing parameter is used by collocation model. In this model, the numerator is added with 0.15 when test feature is absent in training corpus, *i.e.*, for zero likelihood probability of feature. The smoothing parameter 5000 is added to the denominator of equation (6) as the approximate size of the corpus. The equation (6) shows the probability calculation of sense S1 for the target word ਜੱਗ (jagg) using Naive Bayes classifier.

$$P(S_1)P(w_{-2}|s)\ P(w_{-1}|s)P(w_{+1}|s)\ P(w_{+2}|s)= \qquad \ldots(6)$$
$$68/141(3+0.15)(1+0.15)(3+0.15)(0+0.15)/(1039+\ 5000)^4$$

Similarly, the probabilities of senses S2 and S3 are estimated by substituting the values from Table 6 into the equation (6). The winner sense of the target word ਜੱਗ (jagg) is the sense S1 as shown in Table 8.

Table 8: Estimated probabilities using collocation model

S1	S2	S3
86.44	4.28	9.27

6.0 RESULTS AND DISCUSSIONS

In this section, the evaluated results of the proposed WSD system for the Punjabi language have been presented. This system has been tested on BoW and collocation models of feature extraction separately using NB classifier. The same training and test data have been used for training and testing of both the models. This manually created corpus is based on coarse senses of the ambiguous words. The overlapping fine senses of an ambiguous word extracted from Punjabi WordNet are merged to produce its respective coarse senses. For merging of fine senses into coarse senses, different relationships (hypernymy, hyponymy *etc*.) of Punjabi WordNet have been manually analyzed for all 150 most ambiguous nouns. The fine senses of an ambiguous word which have common relationships are merged to generate the coarse senses. For WSD system the defined coarse senses are more useful than the fine senses of the ambiguous words [29].

Fig. 2 delineates the fine senses of 150 most ambiguous nouns along with their corresponding coarse senses. The horizontal axis in Fig. 2 shows ambiguous words used in this experiment while the vertical axis is their respective count of coarse and fine senses. The graph of fine senses of words starts with the highest number of senses (18) of a word to the lowest number of senses (6) of the word.

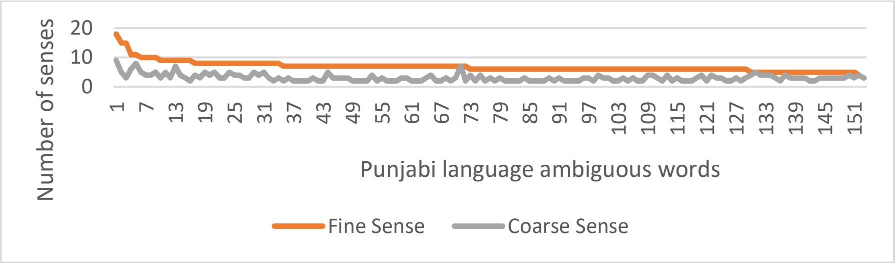

Fig. 2: The fine and Coarse sense of ambiguous Punjabi words

It has been observed that this merging of fine senses to coarse senses has a major impact on those words having more fine senses as compared to words having less fine senses. For example, an ambiguous word "ਭਾਗ" (bhāg) has 15 fine senses and after merging, only 5 coarse senses have been formed. While for the ambiguous word "ਸਿੱਲ" (sill) with 6 fine senses was merged to 4 coarse senses.

In the proposed system, disambiguation has been performed on the coarse senses of the ambiguous words. The results of BoW and collocation models using NB classifier for some of the selected Punjabi language ambiguous noun words are presented in Table 9 and these are arranged in descending order of their coarse senses.

Table 9: Accuracy of proposed models on some of the ambiguous Punjabi words

Words (Transliteration)	No. of Senses		Training instances	Testing instances	BoW Model accuracy	Collocation Model accuracy
	Fine	Coarse				
ਜੋੜ(jōṛ)	11	8	108	16	81.25	68.75
ਤਲਾ(talā)	9	7	84	14	78.57143	71.42857
ਮੇਲ(mēl)	11	6	90	18	83.33333	72.22222
ਨਿਸ਼ਾਨ(nishān)	10	5	75	15	86.66667	80
ਚਾਲ(cāl)	10	5	65	15	80	73.33333
ਗੱਠ(gaṭṭh)	9	5	55	10	80	70
ਕਲਮ(kalam)	8	5	60	15	93.33333	80
ਝਾੜ(jhāṛ)	8	5	65	15	66.66667	60
ਕਲੀ (kalī)	8	5	60	15	86.66667	80
ਸੱਤ (satt)	10	4	52	12	83.33333	75
ਸਮਾਂ(samāṃ)	10	4	48	16	81.25	75
ਅੰਤ(ant)	9	4	44	12	83.33333	75
ਖਿਆਲ(khiāl)	8	4	40	12	91.66667	83.33333
ਨਿਯਮ(niyam)	8	4	48	16	87.5	75
ਰਸ(ras)	9	3	42	14	92.85714	85.71429
ਚਾਹ(cāh)	9	3	45	12	91.66667	75
ਸੱਤਰ(sattar)	9	3	39	13	92.30769	84.61538
ਬੇਨਤੀ(bēntī)	8	3	36	12	83.33333	75
ਮੁਕਤੀ (muktī)	8	3	36	15	86.66667	80
ਇੱਛਾ(icchā)	8	3	39	15	80	80
ਘੜਾ (ghaṛā)	9	2	36	10	90	80

Table 9 indicates, that accuracies of both the models are better when the number of coarse senses of an ambiguous word is less. When the coarse senses of an ambiguous word increases, then there is a drop in the accuracy of both the models. It has also been observed that ambiguous words with a higher number of senses have more overlapping features among the senses; this causes a drop in prediction rate of both the models. While for the words with a lesser number of coarse senses, there are more unique features among the different senses which are the cause of high prediction rate for both the models.

The details of the corpus and result of this experiment have been shown in Table 10. It indicates that average coarse sense of the corpus of 150 ambiguous words is 3.07. It also shows approximately 40 instances per ambiguous word (approximately 66% of the corpus) have been used for training the models while approximately 13 instances per ambiguous word (33% of the corpus) have been used for testing the models. The average accuracy of BoW model for all Punjabi language ambiguous words used in this experiment is 89.62%, while it is 81.97% for the collocation model.

Table 10: Average details of corpus and results

Number of words	Coarse sense	Training instance per word	Test instance per word	BoW model accuracy in Percentage	Collocation model accuracy in Percentage
150	3.07	40.15	12.76	89.62	81.97

The comparison of accuracies of both the models has been given in Fig. 3. It has been observed that the accuracies of the system are always higher or equal when BoW model is used for feature extraction as compared to the usage of collocation model for feature extraction.

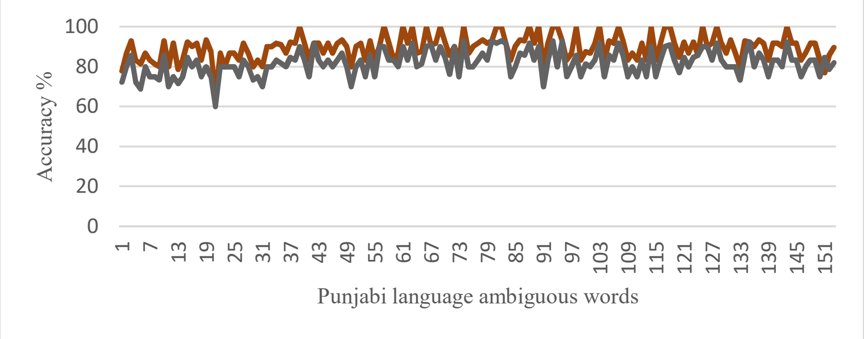

Fig. 3: Accuracies of models for 150 target words

7.0 CONCLUSIONS AND FUTURE SCOPE

In this paper, BoW and collocation models have been used for NB classifier to disambiguate the Punjabi language ambiguous words. It is observed that the accuracy of both models increases with the decrease in the number of senses of the target ambiguous word. It has been further observed that the BoW model outperforms the collocation model for the Punjabi language WSD.

The results of both the models can be further improved by adding more annotated data to the existing corpus. To increase the coverage of the proposed models, sense-tagged corpus for all multi-sense words present in the Punjabi WordNet can be created. The proposed work uses one-word disambiguation per context which can be extended to all word disambiguation per context for the Punjabi language. The standalone WSD model for the Punjabi language described in this paper can be integrated with the other NLP tasks to further improve their outcomes.

REFERENCES

[1] N. Ide and J. Vronis, “Word sense disambiguation: The state of the art”, *Computational Linguistics*, Vol. 24, No. 1, 1998, pp. 1-40.

[2] R. Navigli, “Word sense disambiguation: A survey”, *ACM Computing Surveys (CSUR)*, Vol. 41, No. 2, Article 10, 2009, pp. 1-69.

[3] R. G. Raj and S. Abdul-Kareem, "A pattern based approach for the derivation of base forms of verbs from participles and tenses for flexible NLP", *Malaysian Journal of Computer Science*, Vol. 24, No. 2, 2011, pp. 63-72.

[4] Y.S. Chan, H.T. Ng and D. Chiang, "Word sense disambiguation improves statistical machine translation", *In Annual Meeting-ACL*, Vol. 45, No. 1, 2007, pp. 33.

[5] M. D. R-Moreno, A. Cuesta and D. F. Barrero, "A Framework for massive twitter data extraction and analysis", *Malaysian Journal of Computer Science*, Vol. 27, No. 1, 2014, pp. 50-67.

[6] R.V. Bhala and S. Abirami, "Trends in word sense disambiguation", *Artificial Intelligence Review*, Vol. 42, No. 2, 2014, pp. 159-171.

[7] R. J. Mooney, "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning", *arXiv preprint cmp-lg/9612001,* 1996, pp. 1-10.

[8] T. Pedersen, "A simple approach to building ensembles of naïve bayesian classifiers for word sense disambiguation", in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference,* Association for Computational Linguistics, 2000, pp. 63–69.

[9] C. A. Le, A. Shimazu, "High wsd accuracy using naïve bayesian classifier with rich features", in *Proceedings of Pacific Asia Conference on Language, Information and Computation*, Tokyo, Vol. 18, 2004, pp. 105–113.

[10] R. Florian, S. Cucerzan, C. Schafer and D. Yarowsky, "Combining classifiers for word sense disambiguation", *Natural Language Engineering*, Vol. 8, No. 4, 2002, pp. 327–341.

[11] R. Mihalcea, T.A. Chklovski and A. Kilgarriff, "The senseval-3 english lexical sample task", in *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Stroudsburg, Pennsylvania, 2004.

[12] R.G. Raj and S. Abdul-Kareem. "A Pattern Based Approach for The Derivation Of Base Forms Of Verbs From Participles And Tenses For Flexible NLP", *Malaysian Journal of Computer Science*, Vol. 24, No. 2, 2011, pp 63-72.

[13] R.G. Raj and S. Abdul-Kareem, "Information Dissemination And Storage For Tele-Text Based Conversational Systems' Learning", *Malaysian Journal of Computer Science*, Vol. 22 No. 2, 2009. pp. 138-159.

[14] R. Mihalcea, Evaluation Exercises for the Semantic Analysis of Text*,* 2015, Available at: *//http://www.senseval.org/*.

[15] G. Jiang, Z. Yangsen, "Study on multiple classifiers for Chinese word sense disambiguation", in *International Conference on Artificial Intelligence and Computational Intelligence (AICI),* China, Vol. 1, 2010, pp. 433–437.

[16] T.A. Chklovski, R. Mihalcea, T. Pedersen and A. Purandare, "The SENSEVAL-3 multilingual English-Hindi lexical sample task*"*, *Association for Computational Linguistics*, 2004, pp. 1-4.

[17] D. Yarowsky and R. Florian, "Evaluating sense disambiguation across diverse parameter spaces", *Natural Language Engineering*, Vol. 8, No. 4, 2002, pp. 293–310.

[18] M. Zampieri, "A supervised machine learning method for word sense disambiguation of Portuguese nouns", *Bulletin de Linguistique Aplique et Gnrale-BULAG*, Vol. 34, 2010, pp. 187–203.

[19] P. Liu, "The effect of the number of features to supervised Chinese word sense disambiguation", *Journal of Computers*, Vol. 8, No. 2, 2013, pp. 313–318.

[20] S. Singh, T. J. Siddiqui and S. K. Sharma, “Naive Bayes classifier for Hindi word sense disambiguation”, in *Proceedings of the 7th ACM India Computing Conference*, ACM, Nagpur, India, 2014, pp. 1-8.

[21] P. P. Borah, G. Talukdar and A. Baruah, “Assamese word sense disambiguation using supervised learning”, in *Proceedings of International Conference on Contemporary Computing and Informatics (IC3I),* 2014, pp. 946–950.

[22] S. Parameswarappa and V. Narayana, “Kannada word sense disambiguation for machine translation”, *International Journal of Computer Applications*, Vol. 34, No. 10, 2011, pp. 1-8.

[23] G. Jason and G. Lehal, “Size of n for word sense disambiguation using n-gram model for Punjabi language”, *International Journal of Translation*, Vol. 20, No. 1-2, 2008, pp. 47–56.

[24] P. Rana and P. Kumar, "Word Sense Disambiguation for Punjabi Language Using Overlap Based Approach", *Advances in Intelligent Informatics,* Springer International Publishing, 2015, pp. 607-619.

[25] M. P. Lewis (ed.), *Ethnologue: Languages of the World, Sixteenth edition*, Dallas, Tex.: SIL International, *https://www.ethnologue.com/browse/names/*, 2009.

[26] B. Kachru, Y. Kachru, S. Sridhar, *Language in South Asia,* Cambridge University Press*, https://books.google.co.in/books?id=O2n4sFGDEMYC*, 2008.

[27] P. Kumar and R. K. Sharma, “Generation of UNL attributes and resolving relations for Punjabi enconverter”, *Malaysian Journal of Computer Science*, Vol. 24, No. 1, 2011, pp. 34-46.

[28] A. Narang, R. Sharma and P. Kumar, “Development of Punjabi wordnet”, *CSI Transactions on ICT*, Vol. 1, No. 4, 2013, pp. 349–354.

[29] M. Palmer, H. T. Dang and C. Fellbaum. "Making fine-grained and coarse-grained sense distinctions both manually and automatically", *Natural Language Engineering*, Vol. 13, No. 2, 2007, pp. 137-163.