INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Building Kashmiri Sense Annotated Corpus and its Usage in Supervised Word Sense Disambiguation

Tawseef Ahmad Mir[1], **Aadil Ahmad Lawaye**[2]*, **Parveen Rana**[2], **Ghayas Ahmed**[1]

1 Research Scholar, Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri, India
2 Assistant Professor, Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri, India

OPEN ACCESS

Received: 15-12-2022

Accepted: 04-03-2023

Published: 06-04-2023

Citation: Mir TA, Lawaye AA, Rana P, Ahmed G (2023) Building Kashmiri Sense Annotated Corpus and its Usage in Supervised Word Sense Disambiguation. Indian Journal of Science and Technology 16(13): 1021-1029. https://doi.org/10.17485/IJST/v16i13.2396

* **Corresponding author**.

aadillawaye@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: In this research work maiden attempt is made towards developing a sense annotated corpus for Kashmiri Lexical Sample Word Sense Disambiguation (WSD). Sense annotated dataset is required to use Supervised WSD techniques which are the most effective techniques to carry out WSD. As developing a sense-tagged dataset is an arduous task such datasets are not available for all natural languages. Kashmiri being computationally a low-resource language does not have a sense-tagged corpus available for research purposes. **Methods:** To develop the sense annotated dataset we selected 60 commonly used ambiguous Kashmiri words and annotated the dataset using the manual annotation method. The usefulness of the dataset is also examined by implementing machine learning algorithms (k-NN, Decision Tree (DT) and Support Vector Machine (SVM)) on it. Part of Speech (PoS) and Bag of Words (BoW) features are used to train the classifiers. **Findings:** The performance of the machine learning algorithms for Kashmiri WSD is evaluated using accuracy metric. Out of the different classifiers used SVM showed the best performance with an average accuracy of 75.74%. **Novelty:** This research is the first attempt to develop a sense-tagged dataset for Kashmiri language. The developed dataset would be of great importance to the research community and can be used in various Natural Language Processing tasks like WSD, part-of-speech tagging.

Keywords: Sense Annotation; Machine Learning; Word Sense Disambiguation; WordNet; Part-of-Speech Tagging

1 Introduction

Data-driven approaches to solve Natural Language Processing (NLP) have been more commonly used due to their remarkable performance for a quite long time. With the popularity of machine learning and deep learning techniques to solve NLP problems in recent years, data-driven approaches have attained further importance. This necessitates the development of large-sized datasets for NLP research to flourish. Supervised machine learning techniques which are data-driven have been more successful in NLP than other approaches but require large-scale datasets to train the model. Unfortunately, such datasets are rare and focus has been limited to a few languages to develop such datasets. As a result of this data sparseness has been a challenge in various NLP systems. This fundamental problem in NLP is termed as knowledge acquisition bottleneck.

Acquiring knowledge about different interpretations of words in the given text is required in different NLP tasks. The automated attribution of felicitous meaning to an uncertain term based on its usage is known as Word Sense Disambiguation (WSD). WSD is a problem in computational linguistics that has applications in different NLP tasks like Machine Translation, Information Extraction and Retrieval[1]. Due to the lack of resources like sense-tagged datasets WSD systems have not achieved performance as per expectations which in turn affects the development of efficient NLP systems for different tasks. Although resources are available for languages that are of international importance number of other languages do not have these resources available. Development of such resources is a dare need especially for under-resource languages like Kashmiri.

Based on these limitations this research effort aims at developing a maiden sense-annotated dataset for Kashmiri. Kashmiri language lacks the availability of resources to progress the research in the NLP domain in comparison to other languages like English, Hindi, Urdu, Punjabi, Assamese etc. No sense-tagged dataset is available in Kashmir for research purposes. To carry out the research work we surveyed the research work related to the development of such datasets in other languages.

The research for the creation of sense-annotated corpora that can be utilized for WSD has been ongoing for quite some time. Researchers have analyzed two variants of the WSD problem namely a) Lexical Sample WSD and b) All Words WSD. In Lexical Sample WSD task models are implemented with the purpose to disambiguate a preselected set of ambiguous words. On the other hand, the All Words WSD task aims at disambiguating all dubious words existing in the given text.

The annotated corpora have been created to handle both the Lexical Sample WSD and the All-words WSD tasks. During the early phase of research, the developed sense-tagged datasets were limited in language coverage and size. With the passage of time, large-sized datasets were developed but for limited languages. To develop the sense-tagged datasets that can be used as a standard to evaluate the WSD systems SENSEVAL organized a series of competitions resulting in the development of important datasets for majorly used languages. A survey related to important sense annotated datasets developed for major natural languages to evaluate WSD systems is presented in[2]. Other important All-Words WSD datasets developed recently that are worthy to mention include All-Words Turkish WSD corpus[3], UAW-WSD-18 dataset[4],SBU-WSD-Corpus[5]. All-Words Turkish WSD corpus contains around 83500 terms, 51117 of which are ambiguous. The corpus is syntactically annotated in parallel to semantic annotation, making it helpful for various NLP applications. UAW-WSD-18, a standard All-Words WSD dataset has 5042 Urdu words with 856 occurrences of ambiguous terms. SBU-WSD-Corpus is the first standard test corpus for Persian All-Words WSD. There are 3371 sense-tagged items among the 5892 content Persian words, with 2073 nouns, 566 verbs, 610 adjectives and 122 adverbs.

Lexical sample WSD datasets that we studied to carry out this research work include the Turkish WSD Lexical Sample dataset[6], Arabic Sense-tagged corpus[7], ULS-WSD-18[8]. Turkish Lexical Sample WSD dataset encompasses instances for 30 ambiguous terms (15 nouns and 15 verbs). There are at least 100 occurrences for each ambiguous term obtained from various web sources. Arabic Sense-tagged corpus is built using Wikipedia for annotation purpose. Wikipedia article corresponding to the appropriate sense in the Arabic WordNet is chosen using a mapping process. There are 30961 occurrences in the corpus for 50 dubious Arabic terms with 148 meanings. ULS-WSD-18 is a baseline Lexical Sample Urdu WSD corpus produced for 50 frequently used ambiguous terms. For each selected ambiguous word, the corpus contains 75+15n occurrences (n represents total senses for a target lexical item). The existing research work on the Kashmiri language does not mention any sense-annotated dataset hence there is an extreme need for such a dataset to smoothen the NLP research work in the Kashmiri language. In this research work our objective is to fill this research gap and develop the first-ever Lexical Sample WSD dataset for the Kashmiri language.

Kashmiri is spoken by around 6.8 million people, most of whom live in Jammu and Kashmir. It is an Indo-Aryan language belonging to the Dardic subgroup. It's a verb-second (V2) word order language with a lot of inflection. In the realm of NLP, research in the Kashmiri language is in primitive phase and hence needs proper attention. Only a few research efforts have been made with their scope constrained to machine translation, part-of-speech (PoS) tagging, machine transliteration. Recent research efforts in the NLP domain with respect to the Kashmiri language include[9,10]. In[9] brief account of the research works pertaining to the NLP domain with respect to the Kashmiri language is presented. In[10] research effort Kashmiri-to-English machine translation system based on Long-Short-Term-Memory (LSTM) is presented. To our knowledge, there is no accessible sense-tagged dataset in Kashmiri for research purposes. The major goal of this research is to create a novel Kashmiri Lexical Sample WSD dataset, KLS-WSD-Corpus, as a part of the research presented in[11].

KLS-WSD-Corpus contains a total of 9673 sense annotated instances for 60 commonly used ambiguous Kashmiri words. A set of machine learning algorithms are also evaluated against KLS-WSD-Corpus and results are provided as a baseline for future research on Lexical Sample WSD task in Kashmiri. The tokens in the dataset are associated with part-of-speech (PoS) tags hence can be helpful for PoS tagging and tokenization also.

The rest of the paper is laid out as follows: Section 2 discusses the proposed methodology adapted to carry out this research work. Results and observations that came out of the experiments carried out on the developed dataset are presented in Section 3. Section 4 concludes the research work.

2 Methodology

The overall research work is divided into two stages. The sense-tagged dataset is created in the first stage, and in the second stage, its effectiveness is assessed by using it to train various supervised WSD classifiers. Various tasks involved in the methodology adapted are discussed hereafter.

2.1 Raw Corpus

The major challenge that we come across in this research work is the unavailability of the dataset. Unlike other languages like English, Hindi, Urdu etc. for which enormous data is available in digital form, Kashmiri language is data deficient. After exploring various resources (Trilingual (English-Hindi-Kashmiri) E-Dictionary[12], Kashmiri WordNet[13], dataset used in[14] and other resources), we managed a raw corpus comprising of about 500K tokens. The overall corpus contains text from different domains like Sports, culture, science etc. Using PoS tagger created in research effort[14] thewhole corpus is PoS tagged with an accuracy of 94%.

2.2 Target Word Selection

In this study Kashmiri Lexical Sample WSD task is targeted for which target words needs to be selected first. For these target words, a sense-tagged corpus is constructed, which is then utilized to train the WSD classifier. To select the target words for this research work we extracted all ambiguous words from Kashmiri WordNet and picked out the top 60 among them (Table 1) having the maximum number of sentences present in the corpus collected.

2.3 Sense Annotation

Once we have selected the target words, we extracted instances for these target words from the raw corpus and also obtained senses that these target words can have from the Kashmiri WordNet. In the next step, we have to give the most suitable sense label to the target words in the instances based on the context. We used the manual sense annotation technique for the developed dataset. Kashmiri WordNet contains fine-grained senses, making the annotation task extremely difficult. This complexity is avoided by using coarse-grained senses in place of fine-grained senses. The total senses used in the annotation process for each target word range from 2 to 8. The senses used for annotation purposes are selected after a thorough analysis of all the senses present in Kashmiri WordNet and the advice of the linguistic expert. Senses which are not commonly used or do not have sufficient instances in the dataset are left.

The annotation is done manually by three annotators, all of whom are native Kashmiri speakers and are familiar with the WSD problem. There are two stages in the total annotation process. Two annotators tagged the data in the first phase. Following annotation, these two annotators reviewed their annotations, especially the ones that were unclear, and inter-annotator agreement is computed to ensure that annotation is not subjective. The inter-annotator agreement value of 92% achieved indicated a high level of agreement. In the second step, a linguist expert assessed the annotated corpus and re-annotated any terms that has previously been incorrectly annotated.

2.4 Corpus Encoding and Statistics

The KLS-WSD-Corpus contains 9673 sentences for the selected 60 target words. The corpus is encoded in standard XML format and comprises of 60 files one for each ambiguous word. Each file contains the instances of the target word with training and test splits. Figure 1 shows an example instance for Kashmiri word تیز (taiz) extracted from the developed corpus, KLS-WSD-Corpus.

<contextfile> indicates the beginning of the XML file and has fileno and filename as attributes. Attribute filename indicates the target word for which instances are contained in the file and fileno uniquely identifies the file. <sentence> tag identifies

Table 1. Target words along with Transliteration, number of instances & number of senses in Kashmiri WordNet

Serial No.	Target Word	Total Senses in Kashmiri WordNet	Total Instances	Serial No.	Target Word	Total Senses in Kashmiri WordNet	Total Instances
1	دور Door	9	162	31	کأم Kaam	5	575
2	آمُت Aamut	3	104	32	کأشُر Kashur	5	83
3	موج Mouj	3	164	33	غأر Gair	4	106
4	زیادٕ Zyad	2	216	34	سورُے Soori	4	132
5	آگُر Aagur	4	80	35	زمانہٕ Zaamane	4	141
6	دِل Dil	3	139	36	رۄد Roode	4	96
7	سیۆد Seud	13	150	37	راتھ Rath	4	85
8	مۄل Mool	7	103	38	خیال Khayal	7	153
9	اہم Aham	3	88	39	خاص Khas	4	111
10	شأمِل Shamil	2	88	40	حال Hall	5	120
11	جاے Jaai	9	174	41	جان Jaan	16	174
12	آب Aab	6	192	42	خبر Khabar	4	150
13	اِستعمال Istimal	5	154	43	پۄرٕ Poore	6	232
14	عمل Amal	2	73	44	پَنُن Panun	10	238
15	حِصہٕ Hiss	16	147	45	بۆڈ Boode	7	161
16	تھۆد Thud	5	120	46	بُتھ Buthe	6	132
17	پرٛون Prone	6	113	47	نظر Nazar	4	100
18	کَم Cum	4	186	48	اَکھ Akh	2	250
19	کۄر Koor	5	158	49	حوالہٕ Hawaale	2	76
20	مال Maal	5	87	50	باقٕے Baki	4	91
21	صاف Saaf	23	202	51	آرام Aram	5	124
22	کال Call	3	120	52	ۄنی Veni	3	72
23	کۆچ Kutch	8	132	53	تیز Teez	24	177
24	وَتھ Veth	6	156	54	لاگُن Laagun	22	136
25	سخت Sakath	4	70	55	دۄر Door	7	129
26	واریاہ Waryah	10	1011	56	ترٛاوُن Travun	20	102
27	نیران Neeran	2	127	57	بناوُن Banavun	2	80
28	نۆو Noove	11	111	58	واٹھ Wathe	6	116
29	ناوَ Naave	5	157	59	بِہتھ Behit	8	295
30	لُکھ Lukhe	6	232	60	مُشکِل Mushkil	20	220
Total instances in Dataset							9673

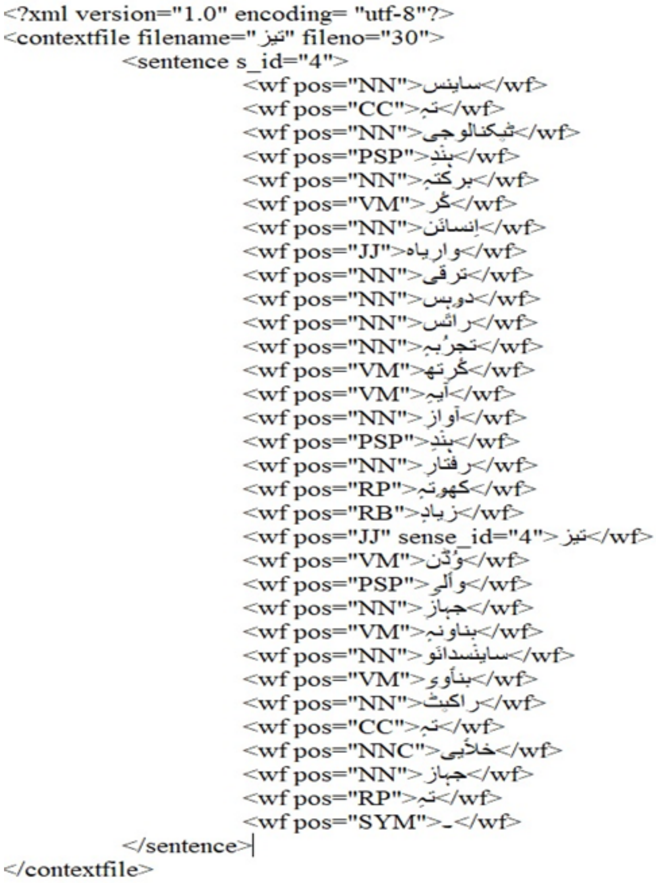
the sentences in the file and the s_id attribute has a numeric value assigned to it which specifies the sentence number of that particular sentence in the file. <wf> tag identifies a particular token in the sentence and has pos as the attribute which gives the lexical category of the token. sense_id attribute depicts the identification number of the sense that the target word takes in the particular sentence.

2.5 Experimental Setup

We investigated the usefulness of the sense annotated dataset, KLS-WSD-Corpus, in the construction and assessment of the Lexical Sample WSD system for the Kashmiri language in the last stage of this research work. The proposed WSD system for the Kashmiri language is depicted in Figure 2 and the various steps involved are discussed in the subsections below.

2.5.1 Data Collection

In the data collection phase raw Kashmiri data is collected from various resources and then a sense annotated dataset is created about which discussion is presented in Subsections 2.1 to 2.4 above.



```
<?xml version="1.0" encoding= "utf-8"?>
<contextfile filename="تیز" fileno="30">
          <sentence s_id="4">
                    <wf pos="NN">ساینس</wf>
                    <wf pos="CC">تہٕ</wf>
                    <wf pos="NN">ٹیکنالوجی</wf>
                    <wf pos="PSP">بنٛدِ</wf>
                    <wf pos="NN">برکتہٕ</wf>
                    <wf pos="VM">کٔر</wf>
                    <wf pos="NN">اِنسانٔن</wf>
                    <wf pos="JJ">واریاہ</wf>
                    <wf pos="NN">ترقی</wf>
                    <wf pos="NN">دوہس</wf>
                    <wf pos="NN">راتس</wf>
                    <wf pos="NN">تجرُبہٕ</wf>
                    <wf pos="VM">کٔرتھ</wf>
                    <wf pos="VM">آیہٕ</wf>
                    <wf pos="NN">آواز</wf>
                    <wf pos="PSP">بنٛدِ</wf>
                    <wf pos="NN">رفتار</wf>
                    <wf pos="RP">کھوتہٕ</wf>
                    <wf pos="RB">زیادٕ</wf>
                    <wf pos="JJ" sense_id="4">تیز</wf>
                    <wf pos="VM">وُڈن</wf>
                    <wf pos="PSP">والِی</wf>
                    <wf pos="NN">جہاز</wf>
                    <wf pos="VM">بناونہٕ</wf>
                    <wf pos="NN">سایٔنسدانَو</wf>
                    <wf pos="VM">بناوٕی</wf>
                    <wf pos="NN">راکیٹ</wf>
                    <wf pos="CC">تہٕ</wf>
                    <wf pos="NNC">خلأیی</wf>
                    <wf pos="NN">جہاز</wf>
                    <wf pos="RP">تہٕ</wf>
                    <wf pos="SYM">۔</wf>
          </sentence>
</contextfile>
```

Fig 1. Example Sentence from Sense Annotated Dataset

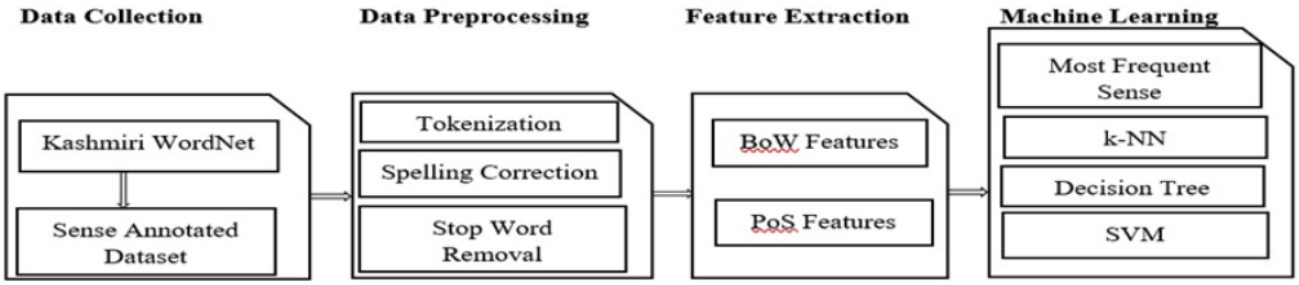


Fig 2. Kashmiri Lexical Sample WSD System

2.5.2 Data Preprocessing

The sense-tagged dataset collected contains data that is not useful for classification purposes. In the data preprocessing phase, the dataset is tokenized and all unwanted data like punctuations, stop words, noisy data, etc. are removed from dataset so that only data which is useful for sense prediction is left behind.

2.5.3 Feature Extraction

Previous research works have shown that the quality of features extracted from the sense-tagged training data set determines the effectiveness of WSD classifiers[15]. Features that have been most commonly used by different researchers in WSD task include Bag of Words, collocations, co-occurrences, PoS, and Word Embeddings. The proposed WSD classifiers are designed based on the following features:

Most Frequent Sense (MFS): A sense among all possible senses of a polysemous word that occurs more frequently than its counterparts is called MFS of that polysemous word. Many researchers have used MFS as a baseline to evaluate the WSD system[16]. In this study, a simple method based on MFS is used as a baseline to evaluate the performance of other machine learning algorithms on the developed dataset.

Part-of-Speech (PoS)**:** PoS of the content words in the instance of the target word is a very effective feature used in WSD systems[17]. Window size also impacts the performance of the WSD system and it has been found that very large Windows sizes do not show much improvement in the performance of the WSD system[18]. In this work Window Size of ±3 has been investigated to carry out disambiguation. If w_i is the target word then PoS based feature vector can be depicted as [PoS_{i-3}, PoS_{i-2}, PoS_{i-1}, PoS_{i+1}, PoS_{i+2}, PoS_{i+3}, sense_label] where PoS_{i-3}, PoS_{i-2}, PoS_{i-1} represents the PoS tags of the three words present before the target word; PoS_{i+1}, PoS_{i+2}, PoS_{i+3} represents PoS tags of three words present after the target word in the given sentence and sense_label specifies the sense which the target word takes in the given sentence. The example below shows the PoS feature for an instance of the target word تیز (taiz) extracted from the dataset.

شُمال مشرِقہٕ طرفہٕ آیہِ تیز ہَوا لہر تہٕ تَمہِ پَتہٕ گۆو روٗد پیوٚن شروع

shumal mashrik tarfe aayi hawa laher te tem pate gov rood peun shuru (Transliteration)

Eliminating the stop words تہٕ, تَمہِ, پَتہٕ from the above sentence the PoS vector with window size of ±3 is extracted as:

Three words before the target word تیز (taiz) in the above sentence are طرفہٕ, آیہِ and مشرِقہٕ, and their PoS tags are VM (verb), NN (noun) and NN respectively, i.e., PoS_{i-1}=VM, PoS_{i-2} =NN and PoS_{i-3} =NN. Similarly, three words present after the target word تیز (taiz) in the sentence are ہَوا, لہر and گۆو having PoS tags as NN, NN, VM respectively, i.e., PoS_{i+1}=NN, PoS_{i+2}=NN and PoS_{i+3}=VM.

The target word تیز (taiz) in the above sentence takes sense یَتھ منٛز واریاہ زور آسَن hence the PoS based feature vector becomes:

[NN, NN, VM, NN, NN, VM, یَتھ منٛز واریاہ زور آسَن].

Bag of Words (BoW): This feature is also important while dealing with the WSD problem. Bag of Words represents the bunch of words that exist along with the target word in the given sentence. In this feature, the position of the words is not taken into consideration instead the presence or absence of the words matters. This feature has also been explored by many researchers in WSD systems and has shown a significant role in sense disambiguation[19]. In order to obtain the BoW feature we first specify vocabulary and then count the frequency with which words from this vocabulary exist in the context of the target word. Sometimes instead of frequency binary digits 0 and 1 are used to show the absence or presence of words from the vocabulary. Suppose the following vocabulary is set for "ناو" (naav) instances.

[شہرس, دٔریاو, گوڑٕ, یاد, لِبکھنہٕ, آبَس, چلاوان, وٕنو, چیز, آبس, سَوأری, سمنٛدر, پھٔٹہٕ,ہأنز, جایہِ, مےٚ, ییٚمہٕ, اَصلی]

The bag-of-words vector for the instance

ناو چھا گوڑٕ کنہٕ ناوِ وول چیز ؟

naav cha gude kin naave vole cheez?

may be represented as:
[0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0]

2.6 Classification

In this research work, we implemented four machine learning approaches on our dataset namely; Most Frequent Sense (MFS), Decision Tree (DT), Support Vector Machine (SVM), and k nearest neighbor (k-NN) algorithm. The four classifiers MFS, DT, k-NN and SVM are trained on selected 60 ambiguous terms separately with distinct lexical and syntactic properties. The hand constructed string files with lexical and syntactic characteristics have been transformed into binary vector files that indicate the existence of a word with a 1 and the absence of a word with a 0. The accuracy of the proposed WSD system is measured as the ratio of cases for which the model properly assigns a sense label to the total number of occurrences examined. Suppose for a target word w having s senses, K be the correctly labeled instances from a total of n instances tested then accuracy Ac may be calculated as per equation 1 below:

$$Ac_n^w = \frac{k_{s\{s\in\{1,2,\ldots,n\}\}}}{n} \tag{1}$$

Furthermore, 5-fold cross-validation approach is used to determine the individual classifier's resilient accuracy. The training data set is divided into five equal subgroups using cross-validation. One subgroup is periodically utilized as a test data set, while the leftover data sets are used for training. The resultant accuracy is obtained by calculating the average of the overall accuracies produced by the classifiers.

3 Results and Discussion

Table 2 depicts the results produced by machine learning algorithms (MFS, k-NN, SVM, DT) implemented using KLS-WSD-Corpus. MFS is used asa baseline in the WSD system designed and gives the average accuracy of 59% which is lower than the other classifiers tested. Classifiers for individual words are trained using PoS features and BoW features separately to analyze the impact of these features on system performance. Experiments are also carried out using both BoW features and PoS features at the same time. From the results obtained it is observed in most cases classifiers for individual words show better performance when trained using BoW features than by using PoS features. The main reason behind the lower performance of PoS based approaches is the smaller size of sense tagged dataset and lesser syntactic combinations. The findings in the table also demonstrate that combining both BoS and PoS features improves the performance of the proposed system significantly than employing either PoS or BoW features separately.

Table 2. Accuracy of Various WSD Approaches on Developed Dataset

Approach	Algorithm	Accuracy
MFS		59
BoW	k-NN	65.72
	Decision Tree (DT)	61.23
	Support Vector Machine (SVM)	63.9
PoS	k-NN	63.53
	Decision Tree (DT)	60.77
	Support Vector Machine (SVM)	66.18
Bow + PoS	k-NN	69.84
	Decision Tree (DT)	66.37
	Support Vector Machine (SVM)	**75.74**

k-NN produced the highest accuracy (65.72%) when used with the BoW feature set. DT produced 61.23% accuracy and SVM showed an accuracy of 63.9% with BoW based model. When PoS based approach is used SVM showed better results (66.18%) than other algorithms (k-NN produced 63.53%, DT produced 60.77% accuracy). DT showed little variation in performance irrespective of using PoS or BoW based approach. All the machine learning algorithms tested produced better results when BoW and PoS based features are used in combination. Among the all algorithms used SVM produced the best results (75.74%) when both PoS and BoW features are used. k-NN produced 69.84% and DT showed 66.37% accuracy when used with both BoW and PoS features at the same time.

According to the examination of the findings of the various machine learning algorithms, it is observed ambiguous words with a greater number of senses perform worse than words with fewer senses. It is also observed that the classifiers performed worse for ambiguous words with a smaller number of instances in the training set. The performance of the proposed system is also analyzed from the PoS category point of view. Table 3 shows the performance of various algorithms used based on PoS categories. When compared to ambiguous terms from other PoS categories, nouns being more ambiguous performed worse whereas adverbs being least ambiguous performed best. It is also evident from Table 3 that based on the PoS categorization SVM with BoW and PoS features recorded the highest accuracy for nouns (62.62%), for adjectives highest accuracy reported is 71.03%, words with PoS as verbs showed the highest accuracy as 73.41% and adverbs produced highest accuracy of 85.56%. The lower performance shown by the ambiguous nouns may be attributed to their higher level of ambiguity level in the sense-tagged dataset as compared to other PoS categories.

Table 3. Performance of various machine learning approaches based on PoS Categories

Approach	Algorithm	Noun	Adjective	Verb	Adverb
MFS		38.4	46.2	64.45	78.6
	k-NN	54.12	59.33	66.71	75.92
BoW	Decision Tree (DT)	48.01	53.73	65.13	70.28
	Support Vector Machine (SVM)	51.23	59.44	63.76	69.46
	k-NN	50.11	58.28	65.3	70.15
PoS	Decision Tree (DT)	44.34	54.23	67.91	71.13
	Support Vector Machine (SVM)	56.61	61.28	67.73	79.14
	k-NN	57.39	65.22	70.24	81.45
BoW + PoS	Decision Tree (DT)	56.81	61.31	67.80	79.20
	Support Vector Machine (SVM)	62.62	71.03	73.41	**85.56**

4 Conclusion

In this research work, we introduced the first-ever Lexical Sample WSD dataset for Kashmiri language which is a highly under-resourced language in terms of linguistic resources. The produced Lexical Sample WSD dataset, KLS-WSD-Corpus, comprises of 9673 occurrences for 60 frequently used ambiguous Kashmiri words. On the dataset developed, various machine learning techniques (MFS, k-NN, DT, SVM) were also examined to demonstrate its usefulness in the construction and assessment of the Lexical-Sample WSD system for Kashmiri. Performance of the WSD system is examined by using BoW and PoS based features separately as well as in combination. MFS based approach produced an accuracy of 59% whereas k-NN, DT and SVM reported accuracy of 65.72%, 61.23% and 63.9% respectively when used with BoW features. When PoS based features were used k-NN, DT and SVM produced an accuracy of 63.53%, 60.77% and 66.18% respectively. When both BoW and PoS features were used then k-NN produced 69.84%, DT produced 66.37% and SVM produced 75.74% accuracy. The results of various experiments carried out during the study reveal that the performance of the system improves when both BoW and PoS feature sets are used at the same time. Out of the different classifiers (MFS, k-NN, DT, SVM) used, SVM produced the best results when used with both BoW and PoS features. The novel dataset produced is a great contribution to the NLP domain for the Kashmiri language as it is the first of this type and can be used in various NLP tasks.

The research work presented here would be extended in the future from different ways:

- The dataset would be extended by adding instances for other ambiguous Kashmiri words to enhance its language coverage.
- Experiments would be carried out using other feature extraction techniques like word embeddings and sense embeddings.
- Additional WSD machine-learning techniques would be examined in the future on the created dataset.

References

1) Zhang G, Lu W, Peng X, Wang S, Kan B, Yu R. Word Sense Disambiguation with Knowledge-Enhanced and Local Self-Attention-based Extractive Sense Comprehension. *Proceedings of the 29th International Conference on Computational Linguistics*. 2022. Available from: https://aclanthology.org/2022.coling-1.357.pdf.
2) Pasini T, Camacho-Collados J. A short survey on sense-annotated corpora. 2018. Available from: https://doi.org/10.48550/arXiv.1802.04744.
3) Akcakaya S, Yildiz OT. An all-words sense annotated Turkish corpus. *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*. 2018;p. 1–6. Available from: https://dx.doi.org/10.1109/icnlsp.2018.8374368.
4) Saeed A, Nawab RMA, Stevenson M, Rayson P. A Sense Annotated Corpus for All-Words Urdu Word Sense Disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 2019;18(4):1–14. Available from: https://dx.doi.org/10.1145/3314940.
5) Rouhizadeh H, Shamsfard M, Tajalli V, Rouhziadeh M, Persian-Wsd-Corpus. A Sense Annotated Corpus for Persian All-words Word Sense Disambiguation. . Available from: https://doi.org/10.48550/arXiv.2107.01540.
6) Ilgen B, Adali E, Tantug AC. Building up Lexical Sample Dataset for Turkish Word Sense Disambiguation. *2012 International Symposium on Innovations in Intelligent Systems and Applications*. 2012;p. 1–5. Available from: https://dx.doi.org/10.1109/inista.2012.6247026.
7) Saif A, Omar N, Zainodin UZ, Aziz MJA. Building Sense Tagged Corpus Using Wikipedia for Supervised Word Sense Disambiguation. *Procedia Computer Science*. 2018;123:403–412. Available from: https://dx.doi.org/10.1016/j.procs.2018.01.062.
8) Saeed A, Nawab R, Stevenson M, Rayson P. A word sense disambiguation corpus for Urdu. 2018. Available from: https://dx.doi.org/10.1007/s10579-018-9438-7.

9) Lone NA, Giri KJ, Bashir R. Natural Language Processing Resources for the Kashmiri Language. *Indian Journal Of Science And Technology*. 2022;15(43):2275–2281. Available from: https://doi.org/10.17485/ijst/v15i43.1964.

10) Lone NA, Giri KJ, Bashir R. Machine Intelligence for Language Translation from Kashmiri to English. *Journal of Information & Knowledge Management*. 2022. Available from: https://doi.org/10.1142/S0219649222500745.

11) Mir TA, Lawaye AA. Word Sense Disambiguation For Kashmiri Language Using Supervised Machine Learning. In: Proceedings of the 17th International Conference on Natural Language Processing. 2020;p. 243–245. Available from: https://aclanthology.org/2020.icon-main.32.pdf.

12) Kak AA, Mehdi N, Lawaye AA, Lone FA. English-Hindi-Kashmiri E- Dictionary: A Case Study. *Linguistic Data Consortium For Indian Languages (LCD-IL)*;p. 21–27. Available from: https://www.academia.edu/35871451/English_Kashmiri_Hindi_e_dictionary_A_Case_Study.

13) Kak AA, Ahmad F, Mehdi N, Farooq M, Hakim M. Challenges, Problems, and Issues Faced in Language-Specific Synset Creation and Linkage in the Kashmiri WordNet. *The WordNet in Indian Languages*. 2017;209:209–220. Available from: https://dx.doi.org/10.1007/978-981-10-1909-8_12.

14) Lawaye AA, Purkayastha BS. Kashmir Part of Speech Tagger Using CRF. *Paripex - Indian Journal Of Research*. 2012;3(3):37–38. Available from: https://dx.doi.org/10.15373/22501991/mar2014/11.

15) Liu P. Another View of the Features in Supervised Chinese Word Sense Disambiguation. *2011 Seventh International Conference on Computational Intelligence and Security*. 2011;p. 1290–1293. Available from: https://dx.doi.org/10.1109/cis.2011.286.

16) Navigli R. Word sense disambiguation. *ACM Computing Surveys*. 2009;41(2):1–69. Available from: https://doi.org/10.1145/1459352.1459355.

17) Abid M, Habib A, Ashraf J, Shahid A. Urdu word sense disambiguation using machine learning approach. *Cluster Computing*. 2018;21(1):515–522. Available from: https://doi.org/10.1007/s10586-017-0918-0.

18) Walia H, Rana A, Kansal VA. A Naïve Bayes Approach for working on Gurmukhi Word Sense Disambiguation. *2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. 2017;p. 432–435. Available from: https://doi.org/10.1109/icrito.2017.8342465.

19) Singh VP, Kumar P. Naive bayes classifier for word sense disambiguation of Punjabi language. *Malaysian Journal of Computer Science*. 2018;31(3):188–199. Available from: https://dx.doi.org/10.22452/mjcs.vol31no3.2.