RESEARCH ARTICLE

SAMRIDDHI Volume 15, Issue 2, 2023 Print ISSN: 2229-7111 Online ISSN: 2454-5767

Towards Developing Word Sense Disambiguation System for Kashmiri Language

Tawseef A. Mir, Aadil A. Lawaye

Department of Computer Science, Baba Ghulam Shah Badshah University Rajouri, Jammu and Kashmir, India.

Abstract

Background: A word, phrase, sentence or other communication is "ambiguous" if interpreted in multiple ways. The process of assigning the correct meaning to a word with respect to its context is known as Word Sense Disambiguation (WSD). WSD is intended to be a very imperious problem in Natural Language Processing (NLP) that requires proper attention as it impacts the performance of various NLP applications.

Objectives: In this paper first attempt is made to propose a supervised machine learning Kashmiri WSD system.

Material & Methods: The dataset comprising of 500K tokens for this research study has been collected from different resources. A sense annotated corpus for fifty commonly used ambiguous Kashmiri words has been created using the manual annotation method. Kashmiri WordNet is used to extract senses for the target words. Decision-tree based classifier is trained using the features extracted from annotated corpus for carrying out WSD task. We have used context widow of ±3 to extract features that are used to train the classifier.

Results: The proposed system is tested on all fifty target words and evaluation is carried using accuracy, precision, recall and F-1 measures. The proposed system reported 81.831% accuracy, 0.834 precision,0.816 recall and 0.824 F1-measure.

Conclusions: This was the initial step towards developing the WSD system for Kashmir and it has shown good results. In the future we expect to use other algorithms to carry out this task with greater language coverage.

Keywords: Information Storage and Retrieval, Machine Learning, Natural Language Processing, Decision Trees.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2023); DOI: 10.18090/samriddhi.v15i02.08

Introduction

Natural language processing (NLP) is one of the sub-fields of Artificial Intelligence (AI). The main concern of the NLP is to solve the problems like information extraction, information retrieval, natural language understanding and generation, Machine Translation, Question Answering with the help of computers. One of the long-standing problems in the NLP is Word Sense Disambiguation (WSD). WSD is the method of associating the best-fit sense from the set of possible senses to a polysemous word according to the context in which it has been used [1]. For example, consider the following sentences:

Double click the mouse to open an application.

The cat ate the mouse.

The word "mouse" has two different meanings in the above sentences. In the first sentence, it means electronic peripheral (pointing device) attached to the computer, whereas in the second sentence it refers to the rodent. In the same manner, consider the following example in Kashmiri:

ہتہِب ہٹ ٕپ یس رُک رادٕ ہتآ ٕز ہُچ ہُس

Su che ze ath daar kursi peth behit (Transliteration)

In the above sentence the word ہتہِب (behit) is ambiguous. Kashmiri WordNet [2] gives eight senses for this which are shown in the Table 1.

Corresponding Author: Tawseef A. Mir, Department of Computer Science, Baba Ghulam Shah Badshah University Rajouri, Jammu and Kashmir, India, e-mail: tawseefmir1191@gmail.com

How to cite this article: Mir, T.A., Lawaye, A.A. (2023). Towards Developing Word Sense Disambiguation System for Kashmiri Language. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 15(2), 234-240.

Source of support: Nil

Conflict of interest: None

In the above sentence the word ہتہِب (behit) takes sense ٔسئ ہِسآ اژھکَ ہِن (yus ne khada aasi) which translates to 'sitting/seated' in English.

As proposed by Ide and Veronis[3], WSD is considered as an AI Complete problem, that is, a problem whose solution presumes a solution to understand natural language or common-sense reasoning. The difficulty in resolving the WSD problem arises from two aspects. First, the sense definitions extracted from dictionaries are ambiguous. Second, world knowledge or common sense involved in WSD is difficult to verbalize in dictionaries.

Table 1: Senses for Kashmiri word بِہته (behit) in Kashmiri WordNet with Transliteration

Kashmiri Word	Sense
بِہته (behit)	ہِسآ ےرِاتفرَ ہِن هتی (Yeth ne raftari aasi) ہِکیٚہ هتکِ پُ ہِن سِیُ (Yus ne paketh heki) ہِسآ اڑھکَ ہِن سِیُ (Yus ne khada aasi) نہُبِ یلٲخ طقف یٚنرِکَ ہِن مٲک ہہحناک (Kenh kaam ne karen fakat khali behun) نارکَ ہِسآ مٲک ای ہِسآ روُآ زنٛم ہِماک ہِن سِیُ (Yus ne kaami manz aavur aasi ya kaam aasi karan) ہِسآ ہِلیٖسو ہہنٛاک کنُوامک راگزور ہِن سِیَ (Yas ne rozgar kamavnuk kenh waseel aasi) ہِسآ روُآ زنٛمَ ہِماک ہِن سِیُ (Yus ne kaam manz aavur aasi) ہِنسآ رادلوڈ کوٚژ هتیَ (Yeth cxook duldaar aasi ne)

Sense knowledge is commonly represented by sense knowledge vector (sense ID, features). Dictionaries have definitions and partial lexical knowledge for each sense but provide very little world knowledge (common sense). The world knowledge can be better learnt from manually sense tagged corpus using machine learning approaches. The words surrounding the target word forms its context and plays important role in assigning the correct sense to the ambiguous word. The context is represented by context vector (word, features). An ambiguous word can be disambiguated by comparing the sense knowledge vector and its context vector.

WSD is not a separate task, it is an intermediate step for a number of other NLP applications like Machine Translation, Information Retrieval (IR), text summarization, Information Extraction (IE), Question Answering etc. [4]. All these NLP applications use WSD module as preprocessing step to perform better. For languages like English, Spanish, French and other European languages WSD problem have been explored by number of researchers and efficient WSD systems exist for these languages but for low resource languages like Gujrati, Punjabi, Manipuri, Kashmiri and other South Asian languages this problem have not been explored thoroughly.

Kashmiri language is used by about seven million people majority of whom are residing in Jammu and Kashmir. It belongs to Dardic group of Indo-Aryan languages. Despite being spoken by millions of people research work in various NLP applications for Kashmiri language lacks far behind due to unavailability of corpus and other processing tools. As of now no Kashmiri WSD system on the basis of machine learning technique is available. No corpus is available in Kashmiri language that can be used for NLP tasks like WSD which restricts the research in higher level NLP tasks. These reasons lead us to carry out this research work. In this study we propose a supervised machine learning based WSD approach using Decision Tree for Kashmiri language. To develop the WSD system we created a sense annotated corpus of 8311 sentences for 50 commonly used ambiguous Kashmiri words. We used Kashmiri WordNet as a source to label the ambiguous words with most appropriate sense. The rest of the paper is divided into following sections: Section 2 discusses the various WSD research works surveyed to carry out this study; Section 3, explains the design methodology adopted; in Section 4 we discussed the model evaluation, Section 5 concludes the work done.

Literature Review

Research work to explore various approaches to solve WSD problem exists from a long time and different techniques have been developed with due course of time to do this task. Different WSD approaches used include supervised approaches, unsupervised approaches, knowledge-based approaches with supervised approaches showing better performance comparatively [4][5]. Although supervised machine learning approaches have been explored to a large extent to decipher correct interpretation of ambiguous words in different languages no research work exists to the best of our knowledge to explore WSD problem in Kashmiri. To take out this research work we have studied research works conducted to solve WSD problem in other languages and brief overview of the literature surveyed is given here.

Decision tree, a well-known supervised machine learning algorithm, have been used by many researchers to carry out WSD in various languages and have performed better in a number of comparative studies [6][7]. In [6] the WSD model is implemented using bag-of-words based features while in the study described in [7] different features like part of speech of neighboring words, morphology and collocations are used to build the model.

In [8] supervised machine learning algorithm based on decision tree algorithm (C4.5) is used to disambiguate text extracted from SEMCOR corpus. Experiments are carried out by using the sense-tagger on general corpus as well as restricting the corpus to news domain. The results showed that using the tagger on domain specific text shows better performance than on general corpus.

In [9] corpus-based approach using decision tree algorithm is implemented for WSD. In this study bigrams are used as features to carry out WSD. Results obtained using different classifiers; the majority classifier, decision stump and Naïve Bayes classifier are compared and contrasted and it is observed that decision tree classifier shows better performance relatively. The important point that comes out as conclusion from this study is, it is more significant to use appropriate feature set instead of variations in algorithms to enhance the system performance.

In [10] Case-Based Reasoning approach is used to disambiguate similar cases in Punjabi text. The system is

implemented using k-NN, decision tree and Naïve Bayes algorithms. Experiments are carried out using bi-grams, tri-grams and four (n-grams) as features. Out of the algorithms used decision-tree based approach produced highest accuracy (84.88%) when used with pre-bigrams.

In [11] authors have examined decision tree algorithm for Assamese WSD. A sense annotated corpus of 50K sentences for 160 ambiguous words is developed using Assamese Corpus [12] and Assamese WordNet [13]. The context window size of ±2 is used in the disambiguation process. On evaluation the system gave an average F-measure of 0.611 for 10 ambiguous words.

[14] proposed Manipuri WSD system based on decision tree algorithm. The experiments in this study are carried on a small dataset of 672 Manipuri sentences only. CART (Classification and Regression Tree) model when tested achieves an accuracy of 71.75%.

[15] used the decision-tree based algorithm for deciphering ambiguous Gurmukhi words. The proposed system is experimented using a sense annotated dataset for 100 ambiguous words. The authors have reported the F-measure of 73.1%.

In [16] a bunch of machine learning based WSD algorithms are compared for their performance. The algorithms are tested to disambiguate all ambiguous words present in the given input. Out of the different algorithms used the results reported in the study depict that decision-tree based approach (J48) defeats other counterparts for the performance. The main reason for the better performance of decision tree in comparison to its counterparts is its ability to rank the features as per the role they play in the disambiguation process.

In [17] supervised machine learning approaches (decision tree (DT), support vector machine (SVM), Naive Bayes) are used for disambiguation purpose in Urdu text. This study has also pointed out that increasing window size improves WSD system performance. For training and test purpose the manually created sense annotated dataset is divided into 80:20 ratio. This study showed that Naïve-Bayes classifier performs better than other algorithms tested.

Materials and Method

In this study we propose Supervised Machine Learning WSD technique based on Decision Tree for Kashmiri language. The flow diagram of the proposed WSD system is depicted in Figure 1 and the steps involved are discussed in the sub-sections below.

Data Collection

Kashmiri language mainly spoken by the people of the Kashmiri and is morphologically very rich but no dataset is available for research purpose which poses a great challenge in this study. Dataset used in this study is collected from Kashmiri WordNet, dataset used in [18], Trilingual Sense Dictionary [19]. In addition, sentences are manually entered

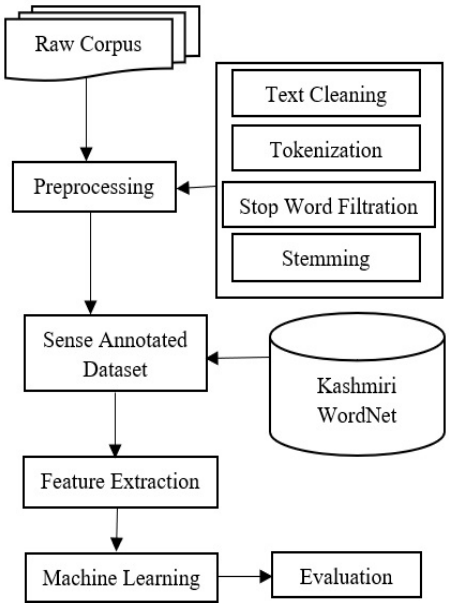

Figure 1: Proposed kashmiri WSD methodology

using keyboard. So, after aggregating the data from various resources the final raw dataset contains about 500K tokens.

Preprocessing

The data collected in the data collection phase is usually in the raw form and cannot be used directly for machine learning purpose. Before feeding the data to the machine learning model it needs to be preprocessed. Various steps used in the preprocessing phase are:

Text Cleaning

Data cleaning is the first step in the pre-processing phase of the data. In this step the data that is in unstructured form i.e., data which cannot be understood by the machines is removed. Also, unwanted characters present in the dataset like {",:;_<<"،,} are eliminated. The raw corpus also contained inconsistent data like missing values, typing errors, spelling mistakes etc. These inconsistencies are also removed in this step. Kashmiri language also contains various diacritic marks and using these change the meaning of the words in some cases and form different words. Due to the lack of standardization same word has different spellings in different sources from which data is extracted. For example, the word written in Kashmiri WordNet as رظن (*nazar*) is written as رظَنَ in Trilingual Sense Dictionary. Similarly, word رۄد (*Door*) has been written as رود in Trilingual Sense Dictionary however, both these words are totally different in meaning. All such issues are resolved in this step.

Tokenization

After cleaning the dataset, the sentences need to be divided into individual units called tokens and this process is called tokenization. Tokens may consist of words, punction marks, numeric values, special characters etc.

Stop Word filtration

Stop words are those words which occur much frequently in the data but does not provide any useful information for carrying out disambiguation process. Identifying these words correctly is very important and it affects the performance of the model. Stop words were carefully identified and removed. تہٕ (*te*), زنہٕ (*henz*), یٚ (ye), کنٔ (*ken*) are some of the examples of stop words that were removed.

Stemming

Stemming is the rule-based process of removing the suffixes from a word. It reduces an inflected word to its root form. For example, the word '*playing*' is inflected and its root word is '*play*'. Kashmiri language is morphologically very rich but no morphological analyzer is freely available. So, in this research work we used only the base forms of the target ambiguous words to simplify the process.

Preparing Sense Annotated Dataset

Supervised machine learning approaches depend on sense annotated dataset and the results solely rely on information gained from training dataset. Due to the absence of any sense annotated dataset for Kashmiri language the sense annotated dataset for this research work is created manually which was a very challenging and time-consuming task. To build a sense annotated dataset we picked out 50 continually used ambiguous words from the collected corpus. The overall procedure to detection and selection of ambiguous words is depicted in Figure 2.

After carrying out the preprocessing steps already discussed we eliminated the duplicates from the dataset. After removing duplicate words, we detected the ambiguous words present in the dataset using the Kashmiri WordNet. After getting the ambiguous words present in the dataset the final selection of target words was made on the basis of frequency with which the ambiguous words were present in dataset. Ambiguous words having maximum frequency in the dataset are considered.

In order to train the classifier efficiently it is required to have sufficient number of instances for each sense the word exhibits. Kashmiri WordNet however contains very fine-grained senses which are often difficult to differentiate from one another. Most of research related to sense disambiguation have targeted coarse-grained WSD instead of fine-grained WSD.

Fine grained senses are converted into coarse grained senses and the senses for which sufficient number of instances are not available are eliminated in this study. In order to convert fine-grained senses into coarse-grained senses relationships like (synonymity/hypernymy/hyponymy etc.) for the target words in the Kashmiri WordNet were thoroughly analyzed and the senses having similarity in these relationships or give same translation in English were treated same. For example, considering the ambiguous word وون (nuv) having Adjective part of speech, Kashmiri WordNet returns total eleven senses shown in Table 2.

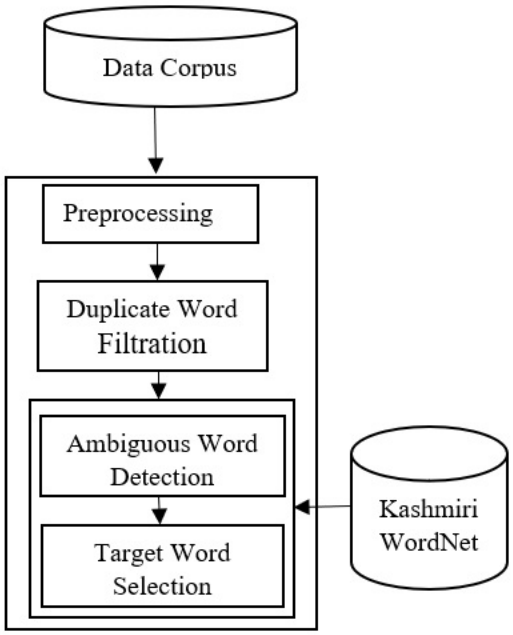

Figure 2: Target word selection

Sense with Sense ID 25 gives sense of 'new' in English, senses with Sense ID 2482 translates to 'unimpaired' in English, senses with Sense IDs 3064, 6375, 6377 translates to 'unused' in English, sense with Sense ID 25555 translates to 'latest' and sense with Sense ID 34223 translates to 'innovative' in English. So, at last the eleven senses for the word وۆن (nuv) are reduced to five for annotation purpose. For the target words number of senses used in the annotation process ranges from 2 to 8. The sense annotated corpus comprises of 8311 sense tagged instances for 50 target words. The sense tagged dataset so developed contains 54271 Nouns, 18533 verbs, 9695 adjectives, 5899 adverbs and 66692 tokens of other PoS category. The resultant corpus contains 50 files saved in XML format one for each target word. Figure 3 shows an example sentence from sense annotated dataset.

Machine Learning

Now we have sense tagged dataset in hand the next step is to train the machine learning model for WSD. The machine learning model is trained using learning sets to carry out WSD and then its performance is tested by supplying test sets as input. Before using learning sets to train the classifier these needs to be converted into feature vector representation. Selection of features for training the WSD classifier have significant impact on system performance hence features should be carefully selected. Commonly used features are collocation, co-occurrence, POS tagging, bag of words. In this research work we used bag of words with a window size of ±3 i.e., three content words before target word and three content words after target word to train the decision-tree based classifier. Thus, feature vector consists of three content words existing before and after target word and the sense label which best suits the target word in the underlying instance.

Table 2: Senses for Kashmiri Word ووٚن (nuv) in Kashmiri WordNet with Transliteration

Word	*Sense ID*	*Sense*
	25	ہِچیٚہ ہِسآ ےلِاح مأک ہہنٛاک ہِمیٚی ژٕم (Yem kenh kaam haali aasi hichmucx)
	2482	ہسِآ صقُون ہتِ ہہناک ہٕن ھتیَ (Yeth ne kenh ten ukus aasi)
	3064	ستِمٕرِ کئ رایتَ ای ستِمیٕنبَ ھتیَ تمُوٚگ ہِسآ لاک ےمٕکَ (Yeth banaimets ya tayar gemtes kami call aasi gumut)
	4531	ہٕننآ ہِنوٕ زنٛم سلامعتسا ہٕن سیُ تمُآ ہِسآ (Yus ne ven istimaals munz aasi aanene aamut)
	6375	ہٕنرکَ ہِسآ دٕأپ ای ہٕنوانبَ یٕنوٕ سیُ تمُآ (Yus ven banavene ya pade aasi karne aamut)
ووٚن (nuv)	6377	ہسآ تمُآ ہنوانب ینٛوِ ینٛوِ سیُ (Yus ven ven banavene aasi aamut)
	15300	ژمآ ہٕنرَک ہسِآ داورٛش ھتٛیَ (Yeth shuruvat aasi karne aamucx)
	22481	ہِنوُ ہٕن سیُ ای ہِسآےنَگال ہِنوُ سیُ تمُروٚک ہِسآ لامعتسِا (Yus ven laagni aasi yay us ne ven istimaal aasi kurmut)
	25372	ہِسآ ہٕبرُجت ہٕن کیُمیٚی (Yemeuk na ven tajrube aasi)
	25555	نورٛپ ہٕن سیُ ای ہٕسآ کیٚکُنوٕ سیُ ہِسآ (Yus venkeenuk yay us ne ven proun aasi)
	34223	ےہسآ زنٛم سدوٗجوٗ ڑوگ ہٕن سیُ (Yus ne gude wajoodes manz aasihy)

In this research work we used decision-tree algorithm (J48) which is an improved version of C4.5[20] machine learning algorithm. Decision tree algorithm is one of the popular algorithms used in classification problems. Based on the features vectors discussed above J48 builds decision tree which can be used for future sense prediction. The nodes of the decision tree generated by J48 evaluate the features present in feature vector. While traversing a path from parent node to leaves a series of such tests are carried out which leads to the decision about final sense of the target word. At every step the features which are most relevant are picked up to construct the decision trees in top-down fashion. The decision about which feature is more significant is taken based on the information-theoretic measure, which pinpoints the role that the feature plays in classification. Learning sets are then partitioned into subsets corresponding to the different values of chosen feature and the same procedure

```
<?xml version="1.0"  encoding ="utf-8"?>
<contextfile filename="دور" fileno="1">
<sentence s_id="43">
                <wf pos="NN">سٔرکیوٗلیٹری</wf>
                <wf pos="NNP">سِسٹَم</wf>
                <wf pos="PSP">سِتۍ</wf>
                <wf pos="VAUX">چُھ</wf>
                <wf pos="NN">خوٗن</wf>
                <wf pos="NN">جِسمِکِن</wf>
                <wf pos="JJ">سأرِنے</wf>
                <wf pos="NN">تانَن</wf>
                <wf pos="PSP">منٛز</wf>
                <wf pos="NN" sense_id="3">دور</wf>
                <wf pos="VM">کران</wf>
                <wf pos="SYM">۔</wf>
        </sentence>
</contextfile>
```

Figure 3: Example sentence extracted from sense annotated dataset

is replicated for each subset until all instances in each subset belong to a single class.

Evaluation

The last step is to evaluate the effectiveness of the proposed system. In this research work we chose the evaluation measures those are typically used in machine learning i.e., accuracy, precision, recall and F1-measure. Accuracy calculates the percentage of the correct classifications out of the total classifications produced by the system. Precision calculates the number of words accredited with correct senses, out of total word-senses considered positive by the system, i.e., the ratio of true-positive to false and true positive examples. Recall calculates the words accredited with correct sense, out of the number of words to be accredited with sense. In other words, recall calculates ratio of true positive examples to the sum of positives in the test set. F1-measure combines precision and recall measures and is obtained by

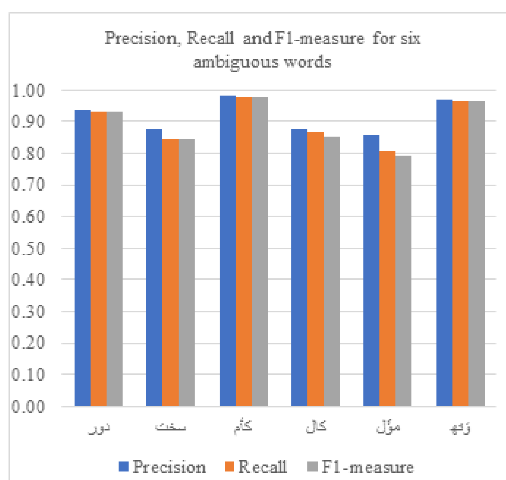

Figure 4: Precision, Recall & F-1 measure for six target words

Table 3: Results produced for six target words by proposed WSD system

Ambiguous word	*No. of instances used*	*Accuracy (%)*	*Precision*	*Recall*	*F1-measure*
رود	208	93.269	0.936	0.933	0.932
تخس	178	84.269	0.878	0.843	0.844
مأک	450	98.000	0.981	0.980	0.980
لاک	127	86.614	0.877	0.866	0.852
لۄم	87	80.459	0.857	0.805	0.793
هتَو	88	96.590	0.968	0.966	0.965

Table 4: Ambiguous word مأک (kaam) along with senses and example sentences

S. No.	Sense	Instances in the dataset	Example Sentence
1	task	219	واڑد هترِ کئ رِوٚپ مأک یننِپَ رِگَ ہسُ (He went home after completing his task)
2	embroidery	119	ہاڑکِ مأک چوِلپَ ہِمیٚی هِچ ترۄصبۄخ (How much beautiful embroidery this cloth has)
3	construction	112	نالچَ مأک ےٚهچ رِگَ (Construction is going on at home)

calculating their harmonic mean. Mathematical equations to obtain accuracy, precision, recall and F-measure are given below:

$$\text{Accuracy (A)}=(TN + TP)/(TN + TP + FN + FP) \quad (1)$$

$$Precision\ (P)=TP/(TP+FP) \quad (2)$$

$$Recall\ (R)=TP/(TP+FN) \quad (3)$$

$$F1\text{-}Measure=(2\times P\times R)/(P+R) \quad (4)$$

where TP, TN, FN and FP represent true-positives, true-negatives, false negatives and false positives.

The system is evaluated on all 50 target words existing in different contexts. Accuracy, precision, recall and F1-measure is obtained for each word individually. At last, the results obtained for all these measures is averaged. To get robust estimate of the performance of proposed system we used 10-fold cross validation. This evaluation dataset is partitioned into ten equal parts, and each part is used as test case once while the remaining nine subparts are used as training case. The sense labels predicted by the proposed system for a particular instance of an ambiguous word are compared with the sense labelled to the word in manually annotated data. The System produces an accuracy = 81.831, precision = 0.834, recall = 0.816 and F-1 = 0.824 on average when tested on all target words.

To get the better understanding of the performance shown by the proposed system we discuss the results produced by the system on some ambiguous words. Table 3 shows the results produced by the WSD system on six target words.

From Table 3 it is evident that ambiguous words with fewer instances show poor performance compared to the ambiguous words with larger instances used. Figure 4 plots the precision, recall and F-1 measures depicted in Table 4. As shown in the Table 3 the proposed system produced best results for ambiguous word مأک (*kaam*). There are 450 instances for word مأک (kaam) with three different meanings in the dataset. Table 4 shows the different meanings, number of instances, and example sentence for each sense in the dataset.

Out of 219 instances belonging to sense '*task*' 211 instances were corrected classified, 118 instances out of 119 instances belonging to sense 'embroidery' were correctly classified and 112 out of 112 instances belonging to sense 'construction' were correctly classified for the word مأک (*kaam*) by the proposed system.

Conclusion

In this study we put forward supervised machine learning based WSD for Kashmiri. The sense annotated dataset is prepared manually for a set of 50 commonly used ambiguous Kashmiri words. Content words surrounding the target word were used as features to build the model. On evaluation system produced accuracy of 81.831%, precision of 0.834, recall of 0.816 and F1-measure of 0.824. In future we are excepting to enhance the dataset coverage by adding instances for more ambiguous words and use more features and carry out experiments using other machine learning and deep learning algorithms.

References

[1] Mittal K, Jain A. WORD SENSE DISAMBIGUATION METHOD USING SEMANTIC SIMILARITY MEASURES AND OWA OPERATOR. ICTACT Journal on Soft Computing. 2015 Jan 1;5(2).

[2] 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. In The WordNet in Indian Languages 2017 (pp. 209-220). Springer, Singapore.

[3] Ide N, Véronis J. Word sense disambiguation: The state of the art. Computational linguistics. 1998 Mar;24(1):1-40.

[4] Agirre E, Edmonds P, editors. Word sense disambiguation: Algorithms and applications. Springer Science & Business Media; 2007 Nov 16.

[5] Pal AR, Saha D. Word sense disambiguation: A survey. arXiv preprint arXiv:1508.01346. 2015 Aug 6.

[6] Mooney RJ. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. arXiv preprint cmp-lg/9612001. 1996 Dec 9.

[7] Pedersen T, Bruce R. A new supervised learning algorithm for word sense disambiguation. InAAAI/IAAI 1997 Jul 1 (pp. 604-609).

[8] Paliouras G, Karkaletsis V, Spyropoulos CD. Machine learning for domain-adaptive word sense disambiguation. InProceedings of the Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, International Conference on Language Resources and Evaluation, Granada, Spain 1998 May 26.

[9] Pedersen T. A decision tree of bigrams is an accurate predictor

of word sense. arXiv preprint cs/0103026. 2001 Mar 29.
[10] Walia H, Rana A, Kansal V. Case Based Construal using Minimal Features to Decipher Ambiguityin Punjabi Language. In2019 Amity International Conference on Artificial Intelligence (AICAI) 2019 Feb 4 (pp. 977-980). IEEE.
[11] Sarmah J, Sarma SK. Decision tree based supervised word sense disambiguation for Assamese. Int. J. Comput. Appl. 2016 May;141(1):42-8.
[12] Sarma SK, Bharali H, Gogoi A, Deka R, Barman A. A structured approach for building Assamese corpus: insights, applications and challenges. InProceedings of the 10th Workshop on Asian Language Resources 2012 Dec (pp. 21-28).
[13] Sarma SK, Gogoi M, Saikia U, Medhi R. Foundation and structure of developing an assamese WordNet. In5th International Conference of the Global WordNet Association (GWC-2010) 2010.
[14] Singh RL, Ghosh K, Nongmeikapam K, Bandyopadhyay S. A decision tree based word sense disambiguation system in Manipuri language. Advanced Computing. 2014 Jul 1;5(4):17.
[15] Walia H, Rana A, Kansal V. A Decision Tree Based Supervised Program Interpretation Technique for Gurmukhi Language. In International Conference on Recent Developments in Science, Engineering and Technology 2019 Nov 15 (pp. 356-365). Springer, Singapore.
[16] Paliouras G, Karkaletsis V, Androutsopoulos I, Spyropoulos CD. Learning rules for large-vocabulary word sense disambiguation: A comparison of various classifiers. In International Conference on Natural Language Processing 2000 Jun 2 (pp. 383-394). Springer, Berlin, Heidelberg.
[17] Abid M, Habib A, Ashraf J, Shahid A. Urdu word sense disambiguation using machine learning approach. Cluster Computing. 2018 Mar;21(1):515-22.
[18] Lawaye AA, Purkayastha BS. Design and Implementation of Part of Speech Tagger for Kashmiri (Doctoral dissertation).
[19] LONE FA. particular context. This paper focuses on the making of the sense based trilingual.
[20] Quinlan JR. C4. 5: programs for machine learning. Elsevier; 2014 Jun 28.