Sādhanā (2019) 44:168
https://doi.org/10.1007/s12046-019-1149-2

Word Sense Disambiguation in Bengali language using unsupervised methodology with modifications

ALOK RANJAN PAL[1,*] and DIGANTA SAHA[2]

[1]Department of Computer Science and Engineering, College of Engineering and Management, Kolaghat, India
[2]Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
e-mail: chhaandasik@gmail.com; neruda0101@yahoo.com

MS received 21 March 2017; revised 1 June 2018; accepted 14 May 2019; published online 27 June 2019

Abstract. In this work, Word Sense Disambiguation (WSD) in Bengali language is implemented using unsupervised methodology. In the first phase of this experiment, sentence clustering is performed using Maximum Entropy method and the clusters are labelled with their innate senses by manual intervention, as these sense-tagged clusters could be used as sense inventories for further experiment. In the next phase, when a test data comes to be disambiguated, the Cosine Similarity Measure is used to find the closeness of that test data with the initially sense-tagged clusters. The minimum distance of that test data from a particular sense-tagged cluster assigns the same sense to the test data as that of the cluster it is assigned with. This strategy is considered as the baseline strategy, which produces 35% accurate result in WSD task. Next, two extensions are adopted over this baseline strategy: (a) Principal Component Analysis (PCA) over the feature vector, which produces 52% accuracy in WSD task and (b) Context Expansion of the sentences using Bengali WordNet coupled with PCA, which produces 61% accuracy in WSD task. The data sets that are used in this work are obtained from the Bengali corpus, developed under the Technology Development for the Indian Languages (TDIL) project of the Government of India, and the lexical knowledge base (i.e., the Bengali WordNet) used in the work is developed at the Indian Statistical Institute, Kolkata, under the Indradhanush Project of the DeitY, Government of India. The challenges and the pitfalls of this work are also described in detail in the pre-conclusion section.

Keywords. Natural language processing; word sense disambiguation; principal component analysis; context expansion.

1. Introduction

Word Sense Disambiguation (WSD) [1, 2] is one of the major tasks in the field of Natural Language Processing (NLP). There are so many words in every language that carry different senses in different contexts. For example, the word "Bank" has different meanings in different contexts, such as "Financial institution", "River side sloppy land", "Reservoir" etc. These words are called ambiguous words. Human brain has some inborn capability to distinguish these senses. However, an automated system depends on some sets of rules for the sense finding. There are three major strategies used in this domain: (a) supervised methodology, (b) knowledge-based methodology and (c) unsupervised methodology.

In supervised methodologies [3–5], a few previously created training sets are used for learning purpose. When a test data comes for sense evaluation, these training sets are used by the system for sense finding.

The knowledge-based methodologies [6, 7] use online dictionaries or thesauri as a sense inventory. The mostly used online semantic dictionary is WordNet.

The unsupervised methodologies [8–12] do not classify the instances; rather, they cluster the instances. This methodology consists of two sub-tasks. First, the sentences are clustered using any clustering algorithm, and these clusters are labelled with their innate senses by manual intervention, as they can be used as sense inventories for further experiment. Next, any distance-based similarity measuring technique is used to find the similarity of a new test data with these sense-tagged inventories. The minimum distance between a test data and a sense-tagged inventory represents the sense of that test data.

In this work, WSD is implemented in the following way: first, sentence clustering is performed using Maximum Entropy (ME) method. The sentence clusters are tagged with relevant senses by manual intervention. Next, Cosine Similarity Measure is used as a distance-based similarity measuring technique. Using this baseline model, the accuracy of WSD achieved is around 35%.

*For correspondence

In this work, two extensions are adopted over the baseline model: (a) Principal Component Analysis (PCA) on the feature vector, which produces 52% accurate result in WSD task, and (b) Context Expansion of the sentences using Bengali WordNet followed by PCA, which produces 61% accuracy in WSD task.

The data sets that are used in this work are obtained from the Bengali corpus, developed under the Technology Development for the Indian Languages (TDIL) project of the Government of India, and the lexical knowledge base (i.e., the Bengali WordNet) used in this work for Context Expansion is developed at the Indian Statistical Institute, Kolkata, under the Indradhanush Project of the DeitY, Government of India. The challenges and the pitfalls of this work are described in the end of this report.

2. Survey

A brief survey in the field of WSD is presented in this section. First, the state-of-the-art performance is presented; next the works in WSD in Asian languages followed by the Indian languages are described.

2.1 *State-of-the-art performance*

The performances of the state-of-the-art WSD systems are presented here. First, the WSD systems were developed using homographs. The accuracy of performance of those systems was above 95% based on very little input knowledge. For example, in 1995, Yarowsky proposed a semi-supervised approach and evaluated on 12 words (96.5%). In 2001, Stevenson and Wilks used a POS-tagged data and other knowledge sources on all words using Longman Dictionary of Contemporary English. Their proposed system achieved an accuracy of 94.7%.

In Senseval-1 (Kilgarriff and Palmer, 2000) evaluation exercise, the best accuracy achieved is 77% on the English lexical sample task, which is just below the level of human performance (80%), estimated by inter-tagger agreement; however, human replicability is estimated at 95%. In 2001, the scores in Senseval-2 (Edmonds and Cotton, 2001) were lower as the task was more difficult, because it was based on the finer grained senses of WordNet. The best accuracy on the English lexical sample task in Senseval-2 was 64% (to an inter-tagger agreement of 86%). Before Senseval-2 exercise, there was a debate on whether the knowledge-based approach was better or the machine-learning-based approach. However, Senseval-2 showed that supervised approaches had the best performance. The performance of the unsupervised system on the English lexical sample task is 40%, which is below the most frequent-sense baseline of 48%, but better than the random baseline of 16%.

In 2004, in Senseval-3 (Mihalcea and Edmonds, 2004) evaluation exercise, the top systems on the English lexical sample task performed at human level according to inter-tagger agreement. The 10 top systems (all supervised) performed 71.8–72.9% correct disambiguation compared with an inter-tagger agreement of 67%. The best unsupervised system overcame the most-frequent-sense baseline, achieving 66% accuracy. The score on the all-word task was lower than in Senseval-2, probably because of the more difficult text. Senseval-3 also brought the complete domination of supervised approaches over the pure knowledge-based approaches.

2.2 *WSD in Asian languages, as well as in Indian languages*

Various works in WSD are implemented in English and other European languages, but very few works are established in Indian languages due to large varieties in morphological inflections and lack of different sense inventories, machine-readable dictionaries and knowledge resources, etc., which are required by the WSD algorithms. The works in different Asian as well as in Indian languages are described in the next sections.

2.2a *Manipuri*: Richard Singh and K. Ghosh proposed an algorithm for Manipuri language in 2013 [13]. In this work, a 5-gram window is formed using the target word and its context words to form the context information. From this contextual information, the actual sense of the focused word is disambiguated. In the work, positional feature is used because of the lack of other relevant morphological features.

2.2b *Malayalam*: Haroon [14] made the first attempt for an automatic WSD in Malayalam language. The author used the knowledge-based approach. One approach is based on a hand-devised knowledge source and the other is based on the conceptual density using Malayalam WordNet. In the first approach the author used the Lesk and Walker algorithm and in the second method, he used the conceptual density-based algorithm, where the semantic relatedness between the words is considered. The semantic relatedness between the words is calculated based on the path, depth and information content of the words in the WordNet.

2.2c *Punjabi*: Kumar and Khanna [15] have proposed a WSD algorithm for resolving the ambiguity of an ambiguous word from a text document in Punjabi language. The authors used a modified Lesk algorithm for WSD. Two hypotheses were considered in this approach. The first one is based on the words that appear together in a sentence, and the final sense is assigned to the target word that is closest according to the neighbouring words. The second approach is based on the related senses, which are identified by finding the overlapping words in their definitions.

2.2d *Assamese*: Sarmah and Sarma [16] have proposed a supervised WSD system based on decision tree. The system

consists of four modules: (a) preprocessing of raw data, (b) sense inventory preparation, (c) feature selection and (d) constructing the decision tree. The algorithm produced an average *F*-measure of 0.611 in the 10-fold cross-validation evaluation strategy when this was tested on 10 Assamese ambiguous words.

Kalita and Barman [17] have proposed a WSD system to disambiguate the Assamese nouns and adjectives. The authors propose a model based on Walker algorithm, which uses the subject category or domain to determine the actual sense of the words. The system produced the precision and recall of 86.66 and 61.09, respectively, on random sentences collected from the internet.

2.2e *Hindi*: Pushpak Bhattacharyya and group [18] have proposed the first WSD system for Hindi nouns using the WordNet. Accuracy of the system ranges from 40% to 70% for various documents like Agriculture, Science, Sociology, etc.

Vishwarkarma and Vishwarkarma [19] have proposed a graph-based algorithm for WSD in Hindi. The authors used a graph-based model based on the similarities among word senses. The authors claim an accuracy of 65.17% in WSD task.

Satyendar Singh and group have proposed a WSD system based on the Leacock–Chodorow semantic relatedness measure [20]. The algorithm is tested on the data set consisting of 20 Hindi polysemous nouns, obtaining the average precision and recall of 60.65% and 57.11%, respectively.

Yadav and Vishwarkarma [21] have proposed a WSD system for Hindi nouns based on mining association rules. The authors claim an average precision of 72% in sense finding.

Gaurav Tomar and group have proposed a WSD system based on word clusters, obtained using Probabilistic Latent Semantic Analysis (PLSA) [22]. The authors tested this method on English and Hindi data sets and achieved the accuracies of 83% and 74%, respectively.

Kumari and Singh [23] have proposed an algorithm for WSD for Hindi nouns using genetic algorithm (GA). The authors applied this algorithm on a list of 12 nouns and achieved a recall of 91.6%.

Devendra K Tayal and group have proposed an approach based on Hyperspace Analogue to Language (HAL) to disambiguate the polysemous Hindi words [24]. The authors claim an accuracy of 79.16% in WSD task for the Hindi words.

2.2f *Nepali*: Roy *et al* [25] have proposed a semantic graph-based algorithm for WSD in Nepali language. This algorithm combines the lexical overlap and conceptual distance-based strategy. The authors carried out the experiment on a data set of 912 nouns and 751 adjectives. The overlap-based approach produced an accuracy for WSD of 54% for nouns and 42% for adjectives, and the conceptual distance-based method produced an accuracy for WSD of 62% for nouns and 58% for adjectives.

2.2g *Myanmar*: Aung *et al* [26] have proposed a WSD system based on Naive Bayes classification to disambiguate the nouns and verbs in Myanmar language. The system was evaluated on 60 ambiguous nouns and 100 ambiguous verbs, which produced 89% precision, 92% recall and 90% *F* score in WSD task.

2.2h *Arabic*: A. Zouaghi and group proposed an algorithm for WSD in Arabic language using Lesk algorithm. The authors claim a precision of 59% using the traditional Lesk algorithm, whereas the modified Lesk algorithm produced a precision of 67% in WSD task.

Mohamed M El-Gamml and M Waleed Fakhr proposed a WSD system for Arabic language using the Support Vector Machine (SVM) classifier following the Levenshetin Distance measuring algorithm to determine the similarity between the words. They compared the performance of their proposed model with the supervised and unsupervised algorithms like Naive Bayes classifier and Latent Semantic Analysis with *K*-mean clustering.

Merhben *et al* [27] have proposed a hybrid approach for WSD in Arabic language. In this experiment, the authors use Latent Semantic Analysis, Harman, Croft and Okapi methods for information retrieval, and finally, the Lesk algorithm is developed for sense disambiguation. In this experiment, the test instances are collected from the web corpus. The authors conduct the experiment on the data set containing 10 ambiguous words, and the accuracy in WSD task is claimed as 73%. Bouhriz *et al* [28] proposed a semi-supervised method for Arabic WSD. The authors claim the precision of 83% in WSD task.

Merhbene *et al* [29] have proposed an algorithm for WSD in Arabic language based on two context information. First, the information that is extracted from the local context of the word to be disambiguated and second, the global context that is extracted from the full text. The authors claim a precision of 74% in WSD task.

2.2i *Bengali*: Although several works on WSD in Bengali language are in progress at different research organizations in India and Bangladesh, only a few of them are available in the web.

Das and Sarkar [30] have presented one WSD system for Bengali to get correct lexical choice for Bengali–Hindi machine translation. The authors used an unsupervised graph-based method to find the sense clusters. Following this strategy, the authors used a vector-space-based approach to map the sense clusters to the Hindi translation of the target word and thus the actual sense of an ambiguous word was predicted from this mapping operation.

Pandit and Naskar [31] have proposed a memory-based approach for WSD in Bengali using *k*-NN method with an accuracy of 71%.

Afsana Haque and Mohammed Moshiul Hoque proposed a dictionary-based approach for sense disambiguation of Bengali noun, adjective and verbs. The authors claim that the proposed system can disambiguate the ambiguous Bengali words with 82.40% accuracy for some selected sentences.

Nazah *et al* [32] have proposed an algorithm for WSD in Bengali language based on Naive Bayes classifier and Artificial Neural Network (ANN). The authors claim an accuracy of 82% in WSD task for some selective Bengali ambiguous words.

3. Proposed unsupervised approach for WSD

Unsupervised approaches perform the WSD task through two sub-tasks: first, sentence clustering, which is used for grouping the sentences into several clusters. Next, these clusters are tagged with relevant senses according to their innate senses and finally, the similarity measuring, which is used for finding the minimum distance between a test data and the sense-tagged clusters. The minimum distance of a test data with a particular sense-tagged cluster assigns the same sense to the test data as that of the cluster it is assigned with.

In this work, sentence clustering is performed using ME method and the similarity is measured using Cosine Similarity Measuring technique.

3.1 *Flowchart of the baseline method*

The flowchart in figure 1 depicts the baseline strategy developed in this work.

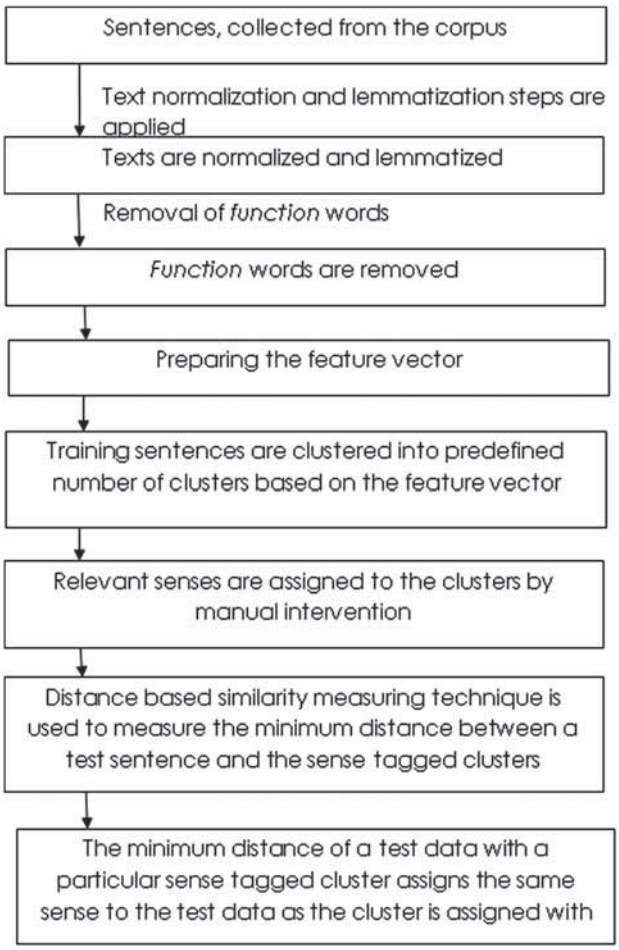

Figure 1. Flowchart of the baseline procedure.

3.2 *Text normalization*

Text normalization is the task of converting a real life text into a uniform representation. The texts retrieved from the TDIL Bengali corpus[1] are not normalized properly. Hence, a set of text normalization steps are executed to transform the texts into machine-readable form. The steps include (a) removal of different punctuation symbols, multiple spaces and new lines, (b) conversion of all the fonts into a single unicode-compatible font ("Vrinda" is used throughout the work), (c) taking into account the different sentence termination symbols (especially the "dāri" symbol, used in Bengali language), etc.

Figures 2 and 3 present a sample non-normalized text and a normalized text, respectively.

[1]The TDIL Bengali corpus is obtained from the Linguistic Research Unit Department, ISI, Kolkata.

3.3 *Text lemmatization*

To increase the lexical coverage of the data sets, all the texts have been lemmatized before this work. In this experiment, the texts are lemmatized using a Bengali lemmatizer tool, developed in a project at CSE Department, Jadavpur University, Kolkata. A sample data after lemmatization task is presented in figure 4.

In our earlier experiment [33], it's observed that as the Bengali words are morphologically very much complex, the percentage of accuracy in WSD increases from 80% to 85% on the same data set when it runs in a non-lemmatized environment and lemmatized environment, respectively.

Hence, in this work, all the experiments are carried out in lemmatized environment.

3.4 *Function word selection*

In Bengali language, there is no specific distinction between the function word and content word; rather it depends on the nature of the experiment. According to

<BTR১২><Literature><Travelogue><১৯৮৩><Book.><স্মৃতিচি><চিন্তামণ><১২১০১>

স্মৃতিচিহ্নিত। চিন্তামণি কর। ভ্রমণকাহিনী।
এঁদের পড়াবার রীতি দেখে মুগ্ধ হলাম। শিক্ষক ছাত্রকে জিজ্ঞাসা করলেন ,
` জো সাঁত্ ' , মানে কি ? ছাত্র চুপ করে দাঁড়িয়ে রইল। সে জানে , ` জো '-র
অর্থ ` আমি ' কিন্তু ` সাঁত্ ' কি জানে না। শিক্ষক অমনি গুন গুন করে গেয়ে উঠে
বললেন , ` সাঁত্ ' । ছাত্র বুঝল , ` আমি গান গাই ' । এমনি সোজোসুজি প্রাসঙ্গিকভাবে
শিক্ষা দেওয়ায় বাড়িতে বিশেষ না খেটেও তাড়াতাড়ি ভাষাটা অভ্যাস হয়ে যায়।
প্রত্যেকটি একটি আন্তর্জাতিক সম্মিলনী করে , শিক্ষার মধ্য দিয়ে সর্বজাতির
মিলন সাধন করেছে এই বিদ্যালয়টি। সামান্য গুটি কয়েক ফরাসী কথা আর
বাকিটা হাত-মুখ নেড়ে ছাত্রদের পরস্পরকে জানবার কি আকুল আগ্রহ ! আমার আসনের
পাশে একটি চেকোশ্লোভাকিয়ান মেয়ে বসত। যে-দিন হিটলার চেকের স্বাধীনতা
চোরের মত সিঁদ দিয়ে চুরি করলে , সে সন্ধ্যায় মেয়েটি ক্লাসে এল না। পরের দিন
অতি গম্ভীরভাবে ক্লাসে এল। প্রফেসার বললেন , " মাদময়জেল তোমার দেশটা
চুরি গেলো ! " সে তখুনি কান্নার উচ্ছ্বাসে ফুঁপিয়ে উঠল ! দেখলাম প্রত্যেকটি
ছাত্রছাত্রীর তার দিকে সহানুভূতির সজল চাহনি। মেয়েটি বলল , " যদি তারা যুদ্ধ

Figure 2. Partial view of a sample non-normalized text.

স্মৃতিচিহ্নিত।চিন্তামণি কর।ভ্রমণকাহিনী।এঁদের পড়াবার রীতি দেখে মুগ্ধ হলাম।শিক্ষক ছাত্রকে জিজ্ঞাসা
করলেন জো সাঁত্ মানে কি।ছাত্র চুপ করে দাঁড়িয়ে রইল।সে জানে জো র অর্থ আমি কিন্তু সাঁত্ কি জানে
না।শিক্ষক অমনি গুন গুন করে গেয়ে উঠে বললেন সাঁত্।ছাত্র বুঝল আমি গান গাই।এমনি সোজোসুজি
প্রাসঙ্গিকভাবে শিক্ষা দেওয়ায় বাড়িতে বিশেষ না খেটেও তাড়াতাড়ি ভাষাটা অভ্যাস হয়ে যায়।প্রত্যেকটি
একটি আন্তর্জাতিক সম্মিলনী করে শিক্ষার মধ্য দিয়ে সর্বজাতির মিলন সাধন করেছে এই
বিদ্যালয়টি।সামান্য গুটি কয়েক ফরাসী কথা আর বাকিটা হাত মুখ নেড়ে ছাত্রদের পরস্পরকে জানবার
কি আকুল আগ্রহ!আমার আসনের পাশে একটি চেকোশ্লোভাকিয়ান মেয়ে বসত।যে দিন হিটলার চেকের
স্বাধীনতা চোরের মত সিঁদ দিয়ে চুরি করলে সে সন্ধ্যায় মেয়েটি ক্লাসে এল না।পরের দিন অতি গম্ভীরভাবে
ক্লাসে এল।প্রফেসার বললেন মাদময়জেল তোমার দেশটা চুরি গেলো।সে তখুনি কান্নার উচ্ছ্বাসে ফুঁপিয়ে
উঠল।দেখলাম প্রত্যেকটি ছাত্রছাত্রীর তার দিকে সহানুভূতির সজল চাহনি।মেয়েটি বলল যদি তারা যুদ্ধ
করে হেরে যেত তা হলে এত দুঃখের কারণ হতো না।চেক সৈন্যের হাতের অস্ত্র হাতেই রইল একটি
গুলিও কেউ ছুঁড়তে পারলে না।স্বাধীন দেশে দ্বন্দ্ব বিক্রম যাদের জীবনের প্রধান অঙ্গ তাদের বীর্য্যকে
কৌশলে অপমানিত করার জ্বালা কতখানি যে তারা অনুভব করে বহুদিন ধরে পরাধীন আমরা তা বুঝতে

Figure 3. Partial view of a sample normalized text.

গর্ভবতী/গর্ভ/noun মহিলাদের/মহিলা/noun চামড়ায়/চামড়া/noun টান/টান/noun ধরে/ধরা/verb
পরে/পরে/adj সেখানে/সেখানে/prn কিছু/কিছু/adj দাগ/দাগ/noun পড়ে/পড়া/verb
যায়/যাওয়া/verb,ভীষণ/ভীষণ/adj ক্লান্তি/ক্লান্তি/noun আর/আর/abay অবসন্নতা/অবসন্ন/adj
ঘিরে/ঘেরা/verb ধরে/ধরা/verb তাকে/তাকে/prn,প্রোজেক্টোরন/প্রোজেক্টোরন/noun
হরমোনের/হরমোন/noun প্রভাবে/প্রভাবে/anu অনেক/অনেক/adj সময়/সময়/noun পায়ের/পা/noun
শিরায়/শিরা/noun টান/টান/noun ধরে/ধরা/verb ফুলে/ফোলা/adj যায়/যাওয়া/verb,অভাব/অভাব/noun
আর/আর/abay অশিক্ষার/অশিক্ষা/noun স্তুপের/স্তুপ/noun উপর/উপর/adj বসে/বসা/verb
থাকলে/থাকা/verb নিরুপণী/নিরুপণ/noun শক্তিতে/শক্তি/noun এরকমই/এরকম/prn ঘুণ/ঘুণ/noun
ধরে/ধরা/verb,ফরম্যাল/ফরম্যাল/noun ফ্লোরাইড/ফ্লোরাইড/noun দাঁতের/দাঁত/non
এনামেলকে/এনামেল/noun য়্যাসিড/য়্যাসিড/noun থেকে/থেকে/anu সুরক্ষিত/সুরক্ষা/noun
রাখে/রাখা/verb যার/যার/prn ফলে/ফল/noun দাঁতের/দাঁত/noun যন্ত্রণাদায়ক/যন্ত্রণা/noun
ক্ষয়/ক্ষয়/noun রোগ/রোগ/noun ধরে/ধরা/verb না/না/abay,পোপ/পোপ/noun ও/ও/abay
রোমান/রোম/noun সম্রাটের/সম্রাট/noun মধ্যেকার/মধ্যে/adj এই/এই/adj সম্পর্কে/সম্পর্ক/noun

Figure 4. A sample text after lemmatization.

theoretical linguistics, all the Bengali words carry relevant senses in specific cases. However, in computational linguistics, to keep the size of the data set within a manageable length, a few less informative words are considered as function words. In this work, the Bengali words, except noun, verb, adjective and adverb (adverbs are also called as a type of adjective in Bengali), are identified and discarded as function words.

3.5 *Feature selection*

Feature selection task performs an important role in clustering operation.

In this experiment, during the clustering task, initially all the distinct words (vocabulary) present in the text were considered as the features for clustering operation. Thus, the length of the feature vector became around 2000–3500 according to the length of the data sets. Unfortunately, my system [Processor: Intel(R) Core(TM) i7-4510U CPU @ 2.00 GHz 2.60 GHz; RAM: 8.00 GB; System type: 64-bit OS] could not handle this length of the feature vector for clustering operation because the size of the feature space (number of sentences v/s feature-vector array length) became too large to handle. For example, when 500 sentences have to be clustered w.r.t. the feature vector of length 3500, the size of the array is [500 $\times$ 3500]. During the mathematical calculations on this extremely large array, my system failed to perform the clustering task.

To resolve this problem, term frequency (TF) of each distinct word in the text is calculated and the features are arranged in decreasing order w.r.t. their TFs. After this, to prepare the feature vector within a manageable length, pruning is applied on the feature vector from the bottom of the list. As a result, the less occurring features are removed from the list and the length of the feature vector becomes shorter. During the experiment, the features having TF of 1, 2, 3 and 4 are gradually pruned and the length of the feature vector becomes gradually shorter. Finally, after removing the features having TF up to 4, the length of the feature vector becomes around 120–160, which is manageable by the system. Thus, the threshold of pruning is set to 4.

3.6 *Selection of ambiguous word*

As stated earlier, according to theoretical linguistics, all the Bengali words carry multiple senses based on the contexts. A standard Bengali lexical dictionary cites the major senses of the words, but the available Bengali machine-readable dictionary (WordNet) covers a smaller subset of that sense domain. Moreover, the TDIL Bengali corpus contains the commonly used ambiguous words with their mostly used senses.

Hence, the system has to follow an effective methodology to select the ambiguous words for the experiment. Using a separate program, it is calculated that the TDIL Bengali text corpus contains totally 3589220 words in inflected and non-inflected forms; among them, 199245 numbers of words are distinct in nature (vocabulary). Using a separate program, the TF of each distinct word is also

না	34041
করে	29081
এই	28373
ও	28172
হয়	22844
এবং	20291
যে	20028
থেকে	16291
আর	14948
তার	14830
?	14686
করা	12774
কিন্তু	12587
বা	12469
যায়	11818
হয়ে	11427
এক	11117
সঙ্গে	11097
কোন	10874
জন্য	10024

Figure 5. The top most occurrences of the words in the corpus.

calculated. Figure 5 presents the top most occurrences of the words in the corpus.

Although theoretically almost every Bengali word carries multiple senses in different contexts [34, 35], in computational field only those words are considered for experiment that are present in the corpus with some needful numbers of occurrences.

3.7 *Selection of senses for the ambiguous words for evaluation*

In reality, most of the senses of the Bengali words are covered by the Bengali lexical dictionary, whereas the online semantic dictionary (Bengali WordNet) contains a subset of the overall senses. Also, the corpus contains only those senses for an ambiguous word that are commonly used in different contexts. In the experiment, only those senses are taken into account that are present in the corpus in some needful number of sentences. The threshold value for the number of sentences carrying a particular sense is set as 20 (as at least 10 sentences could be used for learning purpose and remaining 10 for test purpose). Algorithm-1 presents the sense selection procedure.

Algorithm 1: Sense-Selection

input : Sentences from the corpus containing a particular ambiguous word.
output : Selected senses for that word.

1 Sentences are classified programmatically through context word analysis.
2 Manual intervention is implied to rectify the errors in classification task.
3 Sense inventory is prepared for that particular ambiguous word with the help of Bengali lexical dictionary (Sansad Bānglā Abhidhān) and the Bengali online semantic dictionary (WordNet).
4 All the classes are tagged with their relevant senses.
5 The sense-tagged classes are arranged in decreasing order according to the number of member sentences.
6 Only those classes (senses) are considered for the experiment which contains at least 20 numbers of member sentences.
7 End.

3.8 *Result and corresponding evaluation*

First of all, the total instance sentences are clustered individually, using almost every clustering algorithm available in the weka-3-6-13 tool. However, the clustering results are not acceptable in all the cases. For example, the simple *K*-mean clustering algorithm available in this tool nearly fails to cluster the instances. Some of the other algorithms could not cluster the instances according to the given (pre-defined) number of clusters; a few of them produced even empty clusters.

The ME method performed the clustering task up to a certain level of expectation. In this method, the number of desired clusters is predefined, which is the same as the number of different senses considered for evaluation (see table 1). Next, the derived sentence clusters are labelled with their innate senses by manual intervention.

In the next phase, Cosine Similarity Measure is used as a distance-based similarity measuring technique to find the closeness of a test data with all the sense-tagged clusters. The minimum distance of a test data with a sense-tagged cluster assigns the same sense to the test data as that of the cluster it is labelled with.

During the experiment, approximately half of the data were used for preparation of knowledge base through

Table 1. WSD result using the baseline approach.

Word (no. of sentences)	Number of clusters (number of senses considered for evaluation)	Accuracy (%)
ঘণ্টা: ghantā (498)	2 (বাদ্য যন্ত্র: bādya yantra: bell, ষাট মিনিট: sāt minute: sixty minutes)	38
ঘর: ghar (410)	3 (সংসার: sansār: leading family life, গৃহ : griha: home, বংশ: bansha: family back ground)	43
পড়া: parā (234)	2 (পতিত হওয়া: patita haoyā: fall, পড়াশুনা করা: parāshunā karā: read)	62
জল: jal (423)	4 (বারি: bāri: water, অশ্রু: ashru: tear, জিভে জল : jive jal: saliva, ঘটনা প্রবাহ : ghatanā prabāha: flow of incidents)	24
সময়: samay (353)	2 (ক্ষণ: kshan: time ,সময় কাল:samay kāl: in time)	33
শব্দ: shabda (343)	2 (ধ্বণি: dhanee: sound,অক্ষর: akshar: word)	16
পাতা: pātā (217)	4 (গাছের পাতা: gācher pātā: leaf, পৃষ্ঠা : prishthā: page,অক্ষি পল্লব: akshi pallab: eye leaf, বিছানো: bichāno: unfold)	28
Sentence total: 2478		35

clustering and sense tagging, and remaining data were used for testing purpose.

The accuracy of the result is evaluated programmatically by comparing the sense-derived test data to an ideal result, prepared earlier by the help of a standard Bengali lexical dictionary (Sansad Bānglā Abhidhān).

This baseline model is tested on 7 mostly used Bengali ambiguous words, and the accuracy in WSD task achieved is 35%.

The final result of WSD is presented in the form of percentage of accuracy" instead of precision, recall and *F*-measure, because the system labels a sense tag to each and every sentence either correctly or wrongly.

4. Extensions on the baseline methodology

Two measures are adopted for betterment of accuracy:

(a) PCA on the feature vector and
(b) Context Expansion of the sentences using the Bengali WordNet

4.1 *PCA on the feature vector*

PCA is used in this model to filter the principal components among the features. In the first phase of execution, the entire vocabulary was considered as the feature vector. However, unfortunately the length of the feature vector (according to the size of the data sets, this length varies from 2000 to 3500 approximately) was out of the computational power of the available system (system specification is mentioned in section 3.5). Hence, reduction of length of the feature vector, while preserving the principal components, became an obvious issue. To deal with this problem, PCA is incorporated within this model. During execution of the system, the PCA module available in weka-3-6-13 tool is used to sort out the principal components from the feature vector. This tool selects the principal components and eliminates the least important features from the feature vector with the help of Ranker's algorithm, which is an in-built algorithm used for this task in this tool. Using this tool, the length of the feature vector was reduced to 120–160 approximately from its original size, which was in thousand-scale initially.

Result and corresponding evaluation

Now, the same data sets, used in the baseline methodology are used for sense evaluation using the reduced feature vector and the overall accuracy of WSD is increased to 52% (see table 2).

The accuracy of the result is evaluated programmatically by comparing the sense derived test data to an ideal result, prepared earlier by the help of a standard Bengali lexical dictionary (Sansad Bānglā Abhidhān).

The final result of WSD is presented in the form of percentage of accuracy" instead of precision, recall and *F*-measure, because the system labels a sense tag to every sentence either correctly or wrongly.

4.2 *Context Expansion using the Bengali WordNet*

Although the accuracy of result is increased using PCA, it is not up to the level of expectation. The reason, observed through close observation, is the lack of lexical match between the words of the sentences and the features in the feature vector.

Table 2. WSD result using PCA.

Word (no. of sentences)	Number of clusters (number of senses considered for evaluation)	Accuracy using PCA (%)
ঘণ্টা: ghantā (498)	2 (বাদ্য যন্ত্র: bādya yantra: bell, ষাট মিনিট: sāt minute: sixty minutes)	45
ঘর: ghar (410)	3 (সংসার: sansār: leading family life, গৃহ : griha: home, বংশ: bansha: family back ground)	43
পড়া: parā (234)	2 (পতিত হওয়া: patita haoyā: fall, পড়াশুনা করা: parāshunā karā: read)	78
জল: jal (423)	4 (বারি: bāri: water, অশ্রু: ashru: tear, জিভে জল : jive jal: saliva, ঘটনা প্রবাহ : ghatanā prabāha: flow of incidents)	42
সময়: samay (353)	2 (ক্ষণ: kshan: time ,সময় কাল:samay kāl: in time)	60
শব্দ: shabda (343)	2 (ধ্বণি: dhanee: sound,অক্ষর: akshar: word)	52
পাতা: pātā (217)	4 (গাছের পাতা: gācher pātā: leaf, পৃষ্ঠা : prishthā: page,অক্ষি পল্লব: akshi pallab: eye leaf, বিছানো: bichāno: unfold)	47
Sentence total: 2478		52

To overcome this problem, the same strategy is considered as in the knowledge-based approach, that is Context Expansion of the sentences using Bengali WordNet. In this method, context of every sentence is expanded by handling the meaningful words of the sentences and their synonymous words from the WordNet (see figure 6).

The sizes of the synsets of different words present in the WordNet are different. Although many commonly used Bengali words are not present in this WordNet (see section 5.6), still the system uses the available knowledge from this dictionary for sense expansion. Sense definitions of a sample word are given in table 3.

The Bengali WordNet

The Bengali WordNet is an online semantic dictionary, used for obtaining the semantic information of the Bengali words (Dash 2012). It provides different information about the Bengali words and also gives the relationship/relationships that exists/exist between the words. The Bengali WordNet is developed at the Indian Statistical Institute, Kolkata, under the Indradhanush Project of the DeitY, Government of India. In this WordNet, a user can search a Bengali word and get its meaning. In addition, it gives the grammatical category, namely noun, verb, adjective or adverb, of the word being searched. It is noted that a word may appear in more than one grammatical categories and a particular grammatical category can have multiple senses. The WordNet also provides information for these categories and all senses for the word being searched.

Apart from the category for each sense, the following set of information for a Bengali word is present in the WordNet:

(a) meaning of the word,
(b) example of use of the word,
(c) synonyms (words with similar meanings),
(d) part-of-speech,
(e) ontology (hierarchical semantic representation) and
(f) semantic and lexical relations.

At present the Bengali WordNet contains 36534 words covering all major lexical categories, namely noun, verb, adjective and adverb.

Result and corresponding evaluation

Now, the same data sets used in previous two experiments are used in this phase and the overall accuracy is increased to 61% (see table 4).

The accuracy of the result is evaluated programmatically by comparing the sense derived test data to an ideal result, prepared earlier by the help of a standard Bengali lexical dictionary (Sansad Bānglā Abhidhān).

5. A few close observations

A lot of challenges appeared in every phase of the experiments.

5.1 *Wide range of morphological inflections*

The wide range of morphological inflections of Bengali words is a major factor in this experiment. This range is too large in real life data that it is quite impossible to track all the inflections computationally. For example, in English the word “eat” has only five morphological forms: “eat”, “ate”, “eaten”, “eating” and “eats”. However, in Bengali language, this word has more than 150 morphological inflections including calit (colloquial) and sādhu (chaste) form, such as খাই (khāi), খাও (khāo), খাস (khās), খায় (khāy), খাচ্ছি (khācchi), খান (khān), খাচ্ছিস (khācchis), খাচ্ছেন (khācchen), খাচ্ছ (khāccha), খাচ্ছে (khācche), খেয়েছি (kheyechi), খাচ্ছি (khācchi), খেয়েছ (kheyecha), etc.

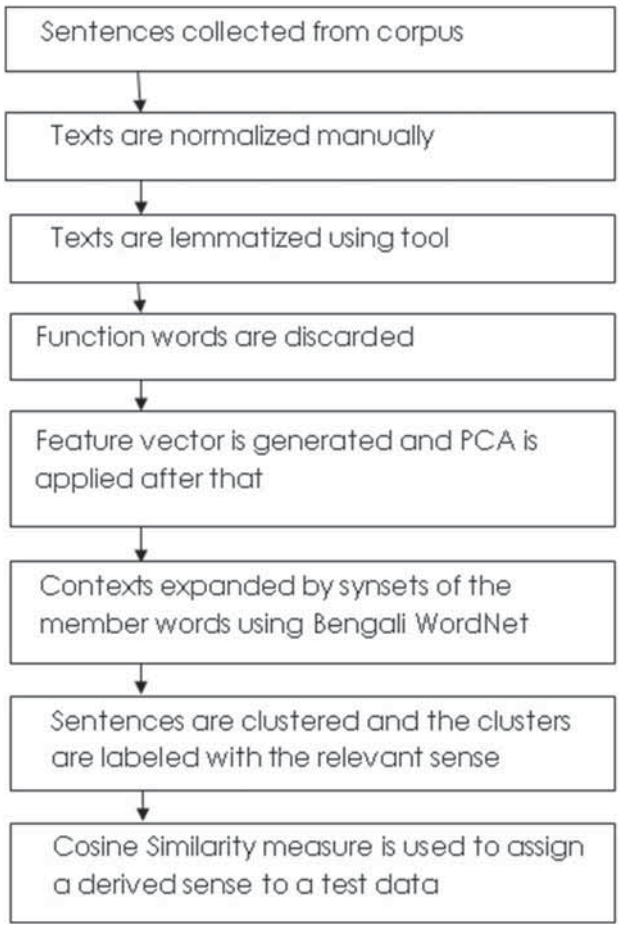

Figure 6. Flowchart of the Context Expansion approach.

Further, the nominal and adjectival morphology in Bengali is lighter compared with verbal morphology. In general, nouns are inflected according to seven grammatical cases (nominative, accusative, instrumental, ablative, genitive, locative and vocative), two numbers (singular and plural), a few determiners like -টা(-ṭā), -টি(-ṭi), -খানা (-khānā) and -খানি (-khāni) and a few emphatic markers like -ই (-i) and -ও (-o), etc. The adjectives, on the other hand, are normally inflected with some primary and secondary adjectival suffixes denoting degree, quality, quantity and similar other attributes. As a result, to build up a complete and robust system for WSD, considering all types of morphologically derived forms with lexical information and semantic relations, is a real challenge in Bengali language.

5.2 *Vast semantic variety*

The vast semantic varieties of Bengali words are also a big challenge in this research field. Examples are the following.

5.2a *Same sense with no contextual similarity*: A few sentences are encountered that carry similar sense but there is no similarity among contextual words. For example

"যে খালের কথা তোমাকে বলেছি তাতে একফোঁটাও জল নেই। (je khāler kathā tomāke bolechi tāte ekfontāo jal nei.)"

"এই রশ্মি জীবাণু নাশক এর দ্বারা জল জীবাণু মুক্ত করা হয়। (ei rashmi jeebānu nāshak er dwārā jal jeebānu mukta karā hay.)"

"এক অণু গ্লিসেরল তিন অণু স্টিয়ারিক য়্যাসিড এক অণু গ্লিসেরল ট্রাইস্টিয়ারেট তিন অণু জল।

(ek anu gliseral tin anu stiyārik yāsid ek anu gliseral trāistiyāret tin anu jal.)"

"রামকৃষ্ণদেব বলতেন আগুনের তাপে জল গরম হয়ে ফুটে উঠলে আলু পটলগুলো সব ওপর নীচ করতে থাকে ছোট ছেলেরা ভাবে আলু পটলগুলো লাফাচ্ছে।

(rāmkrishnadeb balten āguner tāpe jal garam hoye fute uthle ālu patalgulo sab opar neec korte thāke choto chelerā vābe ālu patalgulo lāfācche.)" etc.

Establishing semantic relation among these sentences through the contextual words is a big challenge.

5.2b *Same contextual words with different senses*: This is just the opposite situation of the previous issue. There are several sentences that carry dissimilar meanings through their similar contextual words. For example: "সেই যুগে মানুষ ছিল যাযাবর প্রকৃতির। (sei yuge mānus chila yāyābar prakritir.)" and

ভূপর্যটক কলম্বাস ছিলেন যাযাবর প্রকৃতির মানুষ। (vuparyatak kalambās chilen yāyābar prakritir mānus.)"

"সেইযুগে মানুষ গুহার অভ্যন্তরে প্রাচীর গাত্রে দৈনন্দিন শিকারের হিসাব ও নিহত জীবজন্তুর সংখ্যা প্রস্তরখণ্ডের সাহায্যে খোদাই করিয়া রাখিত।

(seiyuge mānus guhār avyantare prāceer gātre dainandin shikārer hisāb o nihata jeebjantur sankhyā prastarkhander sāhāyye khodāi kariyā rākhita.)" and

"ওড়িশার ময়ূরভঞ্জ এলাকায় কিছু শ্রেণীর মানুষ আছেন যাঁরা বংশপরম্পরায় পাথরের মুর্তি খোদাই করে বাজারে বিক্রি করে সংসার চালান।

(orishār mayurvanj elākāy kichu mānus āchen ynārā banshaparamparāy pātharer murti khodāi kare bājāre bikri kore sansār cālān.)" etc.

These sentences are composed of similar key words but they carry different senses individually.

5.2c *Presence of contextual words in a single sentence carrying different senses*: A few sentences are encountered where multiple-sense-carrying keywords are present in a single sentence to denote a single sense as a whole. For example:

"পান্ডুলিপির ধূসর পাতায় তাঁর আত্মজীবনী আজও এতটাই জীবন্ত যে একবার পড়তে শুরু করলে চোখের পাতা পড়েনা।

(pāndulipir dhoosar pātāy tnār ātmajeebanee ājo etotāi jeebanta ye ekbār parte shuru karle cokher pātā prenā.)" In this sentence, while disambiguating the word "পাতা: pātā", the word "পান্ডুলিপি: pāndulipi" is a contextual word for the sense "

Table 3. Sense definitions of a sample word "মাথা" from the existing Bengali WordNet.

Synset ID	Synonyms	Gloss	Statement ex.
604	মাথা, মস্তক, ললাট: (māthā, mastak, lalāt)	মাথার উপরের এবং সামনের অংশ: (māthār uparer eban sāmner ansha)	রামের মাথায় আভা বিচ্ছুরিত হচ্ছে: (rāmer māthāy āvā bicchurita hocche)
4758	মাথা, মুণ্ড, :(māthā, munda)	শরীরে গলার সামনের বা উপরের সেই গোলাকার অংশ যেখানে চোখ,কান, নাক,মুখ ইত্যাদি অঙ্গ থাকে এবং যার মধ্যে মস্তিষ্ক থাকে: (sharire galār sāmner bā uparer sei golākār ansha jekhāne chokh, kān, nāk, mukh ityādi anga thāke eban jār madhye mastiska thāke)	মাথায় আঘাত লাগার ফলে মানুষের প্রাণও যেতে পারে/কালী মায়ের গলায় মুণ্ডের মালা সুশোভিত থাকে: (māthāy āghāt lāgār fale mānuser prān o jete pāre/kālee māyer galāy munder mālā sushovita thāke)
10988	মাথা:(māthā)	শরীরের সেই অংশ যার মধ্যে মস্তিষ্ক থাকে: (shareerer sei ansha yār madhye mastiska thāke)	মোহনের মাথায় চুল নেই: (mohaner māthāy chul nei)
16852	মাথা: (māthā)	নৌকা বা জলযানের অগ্রভাগ : (noukā bā jalayāner agravāg)	তিনি বিশ্রাম নেওয়ার জন্য নৌকার মাথায় গিয়ে বসলেন: (tini bishrām neoyār janya noukār māthāy giye boslen)
16993	মাথা: (māthā)	কোনো উঁচু ভবন, মহল প্রভৃতির শিখর: (kono unchu vaban, mahal pravritir shikhar)	যে বাড়ির মাথায় চিল বানানো রয়েছে আমি সেখানেই থাকি: (ye bārir māthāy chil bānāno rayeche āmi sekhānei thāki)

পৃষ্ঠা: pristhā", and the word "ধূসর : dhoosar"" is a contextual word for the sense "পৃষ্ঠা: pristhā", as well as "গাছের পাতা: gācher pātā", and "চোখ : cokh" is a contextual word for the sense "অক্ষি পল্লব: akshi pallab".

5.2d *Sentence with sense anomaly*: A few sentences are encountered where it appears quite impossible to tag a particular sense even by human judgment. For example: "সে পরীক্ষার চারিদিকে এত সংযমের বেষ্টন যে সামান্য মানুষ তেমন উপভোগ লাভ করিবার সহিষ্ণুতা সঞ্চয় করিতে পারে না। (se pareekshār cāridike eta sanyamer bestan ye sāmānya mānus teman upavog lāv karibār sahisnutāsanchay karite pāre nā.)" "তন্ত্র বলে সে কথা গুরুমুখ করিয়া শুনিতে হয়। (tantra bale se kathā gurumukh kariyā shunite hay.)"

5.3 *Very large sentence*

Some sentences are too large that they carry large amount of irrelevant information in it. For example: (keha bā dui kāne ānul cāpiyā jhup jhup kariyā drutabege katakgulo dub pāriyā caliyā yāyita, keha bā dub nā diyā gāmchāy jal tuliyā ghana ghana māthāy dhālite thākita, keha bā jaler upari vāger malinatā erāibār janya bārbār dui hāte jal kātāiyā lai laiyā hathāt eksamay dhnā kariyā dub pārita, keha bā uparer sniri haitei binā voomikāy sashabde jaler madhye jhāmp diyā pariyā ātmasamarpan karita, keha bā jaler madhye nāmite nāmite ek nishwāse katakguli shlok āorāiyā laita, keha bā bysta konomate snān sāriyā laiyā bāri jāibār janya utsuk, kāhāro bā bystatā leshmātra nāi dheere susthe snān kariyā jap kariyā gā muchiyā kāpar chāriyā knocātā dui tinbār jhāriyā bāgān haite kichu bā ful tuliyā mridumanda dodul gatite snānsnigdha shareerer ārāmtike bāyute bikeerna karate karate bārir dike tāhār yātrā.)"

> "কেহ বা দুই কানে আঙুল চাপিয়া ঝুপ ঝুপ করিয়া দ্রুতবেগে কতকগুলো ডুব পাড়িয়া চলিয়া যাইত, কেহ বা ডুব না দিয়া গামছায় জল তুলিয়া ঘন ঘন মাথায় ঢালিতে থাকিত, কেহ বা জলের উপরিভাগের মলিনতা এড়াইবার জন্য বারবার দুই হাতে জল কাটাইয়া লইয়া হঠাত্ একসময়ে ধাঁ করিয়া ডুব পাড়িত, কেহ বা উপরের সিঁড়ি হইতেই বিনা ভূমিকায় সশব্দে জলের মধ্যে ঝাঁপ দিয়া পড়িয়া আত্মসমর্পণ করিত, কেহ বা জলের মধ্যে নামিতে নামিতে এক নিশ্বাসে কতকগুলি শ্লোক আওড়াইয়া লইত, কেহ বা ব্যস্ত কোনোমতে স্নান সারিয়া লইয়া বাড়ি যাইবার জন্য উত্সুক, কাহারো বা ব্যস্ততা লেশমাত্র নাই ধীরেসুস্থে স্নান করিয়া জপ করিয়া গা মুছিয়া কাপড় ছাড়িয়া কোঁচাটা দুই তিনবার ঝাড়িয়া বাগান হইতে কিছু বা ফুল তুলিয়া মৃদুমন্দ দোদুল গতিতে স্নানস্নিগ্ধ শরীরের আরামটিকে বায়ুতে বিকীর্ণ করিতে করিতে বাড়ির দিকে তাহার যাত্রা।

5.4 *Very short sentence*

Some sentences are very short in length. As a result, the system could not retrieve sufficient information from them. For example:

"নে জল আন। (ne jal ān.)"

"বাকি রইল একমাত্র মানুষ। (bāki raila ekmātra mānus.)"

5.5 *Spelling error*

In a few cases, spelling errors in the words are the obstacles during execution of the system.

Table 4. WSD result using Context Expansion and PCA.

Word (no. of sentences)	Number of clusters (number of senses considered for evaluation)	Accuracy (%)
ঘণ্টা: ghantā (498)	2 (বাদ্য যন্ত্র: bādya yantra: bell, ষাট মিনিট: sāt minute: sixty minutes)	63
ঘর: ghar (410)	3 (সংসার: sansār: leading family life, গৃহ : griha: home, বংশ: bansha: family back ground)	64
পড়া: parā (234)	2 (পতিত হওয়া: patita haoyā: fall, পড়াশুনা করা: parāshunā karā: read)	81
জল: jal (423)	4 (বারি: bāri: water, অশ্রু: ashru: tear, জিভে জল : jive jal: saliva, ঘটনা প্রবাহ : ghatanā prabāha: flow of incidents)	46
সময়: samay (353)	2 (ক্ষণ: kshan: time ,সময় কাল:samay kāl: in time)	65
শব্দ: shabda (343)	2 (ধ্বণি: dhanee: sound,অক্ষর: akshar: word)	66
পাতা: pātā (217)	4 (গাছের পাতা: gācher pātā: leaf, পৃষ্ঠা : prishthā: page,অক্ষি পল্লব: akshi pallab: eye leaf, বিছানো: bichāno: unfold)	40
Sentence total: 2478		61

Table 5. Unknown sense definitions in Bengali WordNet.

Index no. in WordNet	POS	Gloss	Example sentence	Word
27588	Noun	সাধনা রূপে উপলব্ধ সেই সময়াবধি যা কারও নিয়ন্ত্রণে আছে: (Sādhanā roope upalabdha sei samayābadhi yā kāro niyantrane āche)	আমার খাবার খাওয়ার সময় নেই: (āmār khābār samay nei)	সময় : (Samay)
33958	Noun	যজ্ঞ বা হোমের সেই সময় যখন বৈদিক স্তোত্রের পাঠ করা হয়: (Yajna bā homer sei samay yakhan baidik stotrer pāth karā hay)	সময়ের পর সবাই হোমের সামগ্রী হোম কুন্ডে ঢেলে দিল: (Samayar par sabāi homer sāmagree hom kunde dhele dila)	সময় : (Samay)

The wrong use of ‘শ: sh’, “ষ: s”’, “স: s”’ ; “ি : i”’ “ী : ee”’; “ু : u”, “ূ oo”’; “ত: t”’ and “ৎ : t”’ and different typographical mistakes in the words are the major issues in this aspect. These errors can be managed easily in a manual system; however, in an automated system, these spelling errors directly affect the performance of the system.

5.6 *Scarcity of information in WordNet*

The Bengali WordNet is in developing phase, so it is not a complete reference for semantic information of the Bengali words.

(a) The different sense definitions of the common Bengali ambiguous words are missing in this dictionary, such as “মানুষ: mānus (single sense available), “পড়া: parā” (absent), etc., and a few ambiguous inflected forms such as “নীচে: neece”, “ধরে: dhare”, “ফলে: fale”, “মনে: mane”, etc. are also absent in this dictionary.

b) Some sense definitions are found in the WordNet that are absent in the standard lexical dictionary, as well as those unknown to the linguistic experts (see table 5).

c) Some common relations among the senses of the words are not established (properly/not at all) in this online dictionary, such as hypernymy, hyponymy, holonymy, meronymy, antonymy, etc.

5.7 *Usefulness of function words in Bengali*

Handling the function words and the content words in Bengali is one of the toughest jobs. To bring the size of the data sets to some manageable length, a few function words are removed from the data sets. As there is no fixed boundary between the function words and content words, the keywords with primary part-of-speech (noun, verb, adjective and adverb) are considered as content words. However, the diversity in senses of the Bengali words is too large that a few part-of-speeches like অব্যয়: indeclinable, অনুসর্গ: postposition, সাহায্যকারী ক্রিয়া: auxiliary verb, etc. are used as the function words in a few cases, as well as content words in a few cases. For example, the words “হওয়া: haoyā”, “করা: karā”, etc. are generally used as auxiliary verbs, but when those words are used as a part of a compound word, they are used as content words, such as

"মানুষ হওয়া: mānus karā", "হাত করা: hāt karā", etc. The word "কাছে: kāche" is used in different sentences with three different part-of-speeches, such as "বিশেষ্য: তার কাছ থেকে এনেছি noun: tār kāch theke enechi", "ক্রিয়া বিশেষণ: কাছেপিঠে তখন লোক ছিল না adverb: kāchepithe takhan lok chila nā", "অব্যয়: বিদ্যার কাছে অর্থ মূল্যহীন and indeclinable: bidyār kāche artha moolyaheen". Hence, handling the function word and content word in Bengali might be a separate research work.

6. Conclusion and future work

In this work, WSD in Bengali language is presented using unsupervised methodology. First, the ME method is used as a baseline clustering algorithm. Next, two extensions, PCA over the feature vector and Context Expansion of the sentences using WordNet, are implemented, and Cosine Similarity Measure is used as a distance-based similarity measuring technique.

Although, the accuracy of result is increased due to these two extensions, one obvious obstacle still remains in this methodology. As the size of the test instances is scaled down to sentence level (smaller context) from document level (larger context), two issues appear at the time of clustering; first, the TF of a feature in a sentence becomes very small, which plays an important role in the mathematical calculation in clustering task, and second, the intra-cluster relations among the features have not been established properly, which has a great impact on the accuracy of output.

Finally, through a close observation it is also noticed that, although the collocating words of a keyword have multiple meanings in the WordNet, associated with related synsets, glosses and example sentences, they do not participate in lexical overlap as their sense domains are different.

References

[1] Ide N and Véronis J 1998 Word sense disambiguation: the state of the art. *Computational Linguistics* 24(1): 1–40

[2] Navigli R 2009 Word sense disambiguation: a survey. *ACM Computing Surveys* 41(2): 1–69

[3] Sanderson M 1994 Word sense disambiguation and information retrieval. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'94*, July 03–06, Dublin, Ireland, Springer, New York, pp. 142–151

[4] Mihalcea R and Moldovan D 2000 An iterative approach to word sense disambiguation. In: *Proceedings of FLAIRS*, Orlando, FL, pp. 219–223

[5] Sanderson M 1994 Word sense disambiguation and information retrieval. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'94*, Dublin, Ireland, pp. 142–151

[6] Banerjee S and Pedersen T 2002 An adapted Lesk algorithm for word sense disambiguation using WordNet. In: *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 136–145

[7] Lesk M 1986 Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: *Proceedings of SIGDOC '86, the 5th Annual International Conference on Systems Documentation*, Toronto, Ontario, Canada, pp. 24–26

[8] Seo H, Chung H, Rim H, Myaeng S H and Kim S 2004 Unsupervised word sense disambiguation using WordNet relatives. *Computer Speech and Language* 18(3): 253–273

[9] Martin W T and Berlanga L R 2012 A clustering-based approach for unsupervised word sense disambiguation. In: *Procesamiento del Lenguaje Natural*, Revista no 49, pp 49–56

[10] Heyan H, Zhizhuo Y and Ping J 2011 Unsupervised word sense disambiguation using neighborhood knowledge. In: *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pp. 333–342

[11] Niu C, Li W, Srihari R K, Li H and Crist L 2004 Context clustering for word sense disambiguation based on modeling pairwise context similarities. In: *Proceedings of SENSEVAL-3, Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain

[12] Jurafsky D and Martin J H 2000 *Speech and language processing*. ISBN 81-7808-594-1, Pearson Education (Singapore) Pte. Ltd. Indian Branch, Delhi 110092, India

[13] Singh R L, Ghosh K, Nongmeikapam K and Bandyopadhyay S 2014 A decision tree based word sense disambiguation system in Manipuri language. *Advanced Computing: An International Journal* 5(4): 17–22

[14] Haroon R P 2010 Malayalam word sense disambiguation. In: *Proceedings of the 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*

[15] Kumar R and Khanna R 2011 Natural language engineering: the study of word sense disambiguation in Punjabi. *Research Cell: An International Journal of Engineering Sciences* 1: 230–238

[16] Sarmah J and Sarma S K 2016 Decision tree based word sense disambiguation for Assamese. *International Journal of Computer Applications* 141: 42–48

[17] Kalita P and Barman A K 2015 Implementation of Walker algorithm in word sense disambiguation for Assamese language. In: *Proceedings of the International Symposium on Advanced Computing and Communication (ISACC)*, pp. 136–140

[18] Shahid H and Preeti Y 2014 Study of Hindi word sense disambiguation based on Hindi WorldNet. *International Journal for Research in Applied Science and Engineering Technology* 2(5): 390–395

[19] Vishwarkarma S and Vishwarkarma C 2012 A graph-based approach to word sense disambiguation for Hindi language. *International Journal of Scientific Research Engineering & Technology* 1(5): 313–318

[20] Singh S 2013 Hindi word sense disambiguation using semantic relatedness measure. In: *Proceedings of the*

International Workshop on Multi-disciplinary Trends in Artificial Intelligence, pp. 247–256

[21] Yadav P and Vishwarkarma S 2013 Mining association rules based approach to word sense disambiguation for Hindi language. *International Journal of Emerging Technology and Advanced Engineering* 3(5): 470–473

[22] Tomar G S *et al* 2013 Probabilistic latent semantic analysis for unsupervised word sense disambiguation. *International Journal of Computer Science Issues* 10(5): 127–133

[23] Kumari S and Singh P 2013 Optimized word sense disambiguation in Hindi using genetic algorithm. *International Journal of Research in Computer and Communication Technology* 2(7): 445–449

[24] Tayal D K 2015 Word sense disambiguation in Hindi language using hyperspace analogue to language and fuzzy-*C* means clustering. In: *Proceedings of the International Conference on Natural Language Processing (ICON)*

[25] Roy A, Sarkar S and Purkayastha B S 2014 Knowledge based approaches to Nepali word sense disambiguation. *International Journal on Natural Language Computing* 3(3): 51–63

[26] Aung N T, Soe K M and Thein N L 2011 A word sense disambiguation system using Naive Bayes algorithm for Myanmar language. *International Journal of Scientific & Engineering Research* 2(9): 1–7

[27] Merhben L, Zouaghi A and Zrigui M 2010 Ambiguous Arabic words disambiguation. In: *Proceedings of the 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pp. 157–164

[28] Bouhriz N, Benabbou F and Lahmar E H B 2016 Word sense disambiguation approach for Arabic text. *International Journal of Advanced Computer Science and Applications* 7(4): 381–385

[29] Merhbene L, Zouaghi A and Zrigui M 2013 A semi-supervised method for Arabic word sense disambiguation using a weighted directed graph. In: *Proceedings of the International Joint Conference on Natural Language Processing*, pp. 1027–1031

[30] Das A and Sarkar S 2013 Word sense disambiguation in Bengali applied to Bengali–Hindi machine translation. In: *Proceedings of the 10th International Conference on Natural Language Processing (ICON)*, Noida, India

[31] Pandit R and Naskar S K 2015 A memory based approach to word sense disambiguation in Bangla using *k*-NN method. In: *Proceedings of the 2nd IEEE International Conference on Recent Trends in Information Systems (ReTIS)*, pp. 383–386

[32] Nazah S, Hoque M M and Hossain R 2017 Word sense disambiguation of Bangla sentences using statistical approach. In: *Proceedings of the 3rd International Conference on Electrical Information and Communication Technology (EICT)*, pp. 1–6

[33] Pal A R, Saha D, Naskar S and Dash N S 2015 Word sense disambiguation in Bengali: a lemmatized system increases the accuracy of the result. In: *Proceedings of the 2nd IEEE International Conference on Recent Trends in Information Systems (ReTIS)*, pp. 342–346

[34] Dash N S 1999 Corpus oriented Bangla language processing. *Jadavpur Journal of Philosophy* 11(1): 1–28

[35] Dash N S and Chaudhuri B B 2001 A corpus based study of the Bangla language. *Indian Journal of Linguistics* 20: 19–40