Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 2 (2017) pp. 267-283
© Research India Publications
http://www.ripublication.com

A Knowledge based Methodology for Word Sense Disambiguation for Low Resource Language

Alok Ranjan Pal
Dept. of Computer Science and Engineering
College of Engg. and Mgmt, Kolaghat, India.

Diganta Saha
Dept. of Computer Science and Engineering
Jadavpur University, Kolkata, India.

Antara Pal
Dept. of Computer Science and Engineering
NIT, Durgapur, India.

Abstract

In this work a knowledge based methodology has been proposed for Word Sense Disambiguation for any low resource language. For case study, the module has been tested on data sets in Bengali language. As a knowledge base, Bengali WordNet has been used in this work. As this online dictionary is not a complete knowledge reference, an extension has been adopted to overcome this scarcity of information. First, the traditional Lesk algorithm has been developed as baseline knowledge based approach. But, due to less number of overlap, this baseline method resolute 31% accuracy. Next, an extension of Context Expansion through Synset analysis has been adopted. This extension to the algorithm produced the accuracy around 75%. The data sets have been prepared from the Indian Languages Corpora Initiative (ILCI) Part-of-Speech Tagged Bangla corpus and the Bengali WordNet has been developed by Indian Statistical Institute (ISI), Kolkata. The pitfalls and challenges have been described in this report at last.

Keywords: Natural Language Processing, Word Sense Disambiguation, Knowledge base, WordNet.

INTRODUCTION

In all major languages around the world, there are so many words which carry different meanings according to its use in different contexts. These words are called ambiguous words. For example, in English language, the words “Bank”, “Plant” etc. have different meanings in different contexts. Finding the exact sense of an ambiguous word in a particular context is called Word Sense Disambiguation (WSD) [1-5]. There are three major methodologies for solving this problem, a) Supervised Methodology, b) Knowledge based Methodology and c) Unsupervised Methodology.

In Supervised Methodology [6-25], sense disambiguation is performed with the help of previously tagged learning sets. These learning sets contain related sentences for a particular sense of an ambiguous word. The Supervised methods classify the new test sentences according to the probability distributions, calculated on these learning sets.

In Knowledge based Methodology [26-36], the exact sense of an ambiguous word is resoluted with the help of a dictionary or thesauri.

In Unsupervised Methodology [37-39], the sense disambiguation task is performed in two phases. First, the sentences are clustered using any clustering algorithm and these clusters are tagged with relevant senses with the help of a linguistic expert. Next, a distance based similarity measuring technique is used to find the closeness between the sense-tagged clusters and a new test data. The minimum distance from a sense tagged cluster assigns that sense to that new test data.

In this work, a Knowledge based method has been proposed for Word Sense Disambiguation in Bengali language. In Asian languages like Hindi, Bengali, Tamil, Telugu, Punjabi, Marathi, Malayalam etc., the linguistic resources in computational environment are inadequate, the knowledge bases like online dictionary, and thesauri etc. are not a complete knowledge reference. As a result, the traditional approaches could not produce a satisfactory result. In this work, first the traditional Lesk approach has been developed as a *baseline* method. Due to the obvious reason of scarcity of lexical overlap, the baseline method produced 31% accurate result. As an extension to this baseline approach, context expansion through synset analysis has been adopted. This extended method resolute the test data with 75% accuracy.

In this experiment, test data have been prepared from the Bengali Text Corpus, developed in the Technology Development in Indian Languages (TDIL) project of the Govt. of India and as a knowledge base, Bengali WordNet has been used which is developed by Indian Statistical Institute (ISI), Kolkata.

SURVEY

In this section, a brief survey has been presented on WSD using knowledge based methodology.

LESK Algorithm

This is the first machine readable dictionary based algorithm built for word sense disambiguation. This algorithm depends on the overlap of the dictionary definitions of the words in a sentence. Typical Lesk approach [26, 40] selects a short phrase from the sentence containing an ambiguous word.

Then, dictionary definition (gloss) of each of the senses for ambiguous word is compared with glosses of other words in that particular phrase. An ambiguous word is being assigned with the particular sense, whose gloss has highest frequency (number of words in common) with the glosses of other words of the phrase.

Semantic Similarity

It is said that words that are related, share common context and therefore the appropriate sense is chosen by those meanings, found within smallest semantic distance [41, 42]. This semantic feature is able to provide harmony to whole discourse. Various similarity measures are used to determine how much two words are semantically related. When more than two words are there, this approach also becomes extremely computationally intensive.

Selectional Preferences

Selectional preferences [43-45] find information of the likely relations of word types, and denote common sense using the knowledge source. For example, Modeling-dress, Walk-shoes are the words with semantic relationship. In this approach improper word senses are omitted and only those senses are selected which have harmony with common sense rules.

The basic idea behind this approach is to count how many times this kind of word pair occurs in the corpus with syntactic relation. From this count, senses of words will be identified. There are other methods, which can find this kind of relation among words using conditional probability.

Heuristic Method

In this approach, the heuristics are evaluated from different linguistic properties to find the word sense. Three types of heuristics used as a baseline for estimating WSD system: 1) Most Frequent Sense, 2) One Sense per Discourse and 3) One Sense per Collocation.

The Most Frequent Sense works by finding all likely senses that a word can have and it is basically right that one sense occurs often than the others. One Sense per

Discourse says that a word will preserve its meaning among all its occurrences in a given text. And finally, One Sense per Collocation is same as One Sense per Discourse except it is assumed that words that are nearer provide strong and consistent signals to the sense of a word.

COMPONENTS ASSOCIATED WITH THE WORK

The components associated with the work are described below-

The Bengali Corpus

The ILCI POS Tagged Bangla corpus, developed in Technology Development in Indian Languages (TDIL) project of the Govt. of India has been used in this research work. This corpus contains text samples from 85 text categories or subject domains like Physics, Chemistry, Mathematics, Agriculture, Botany, Child Literature, Mass Media, etc. covering 11,300 A4 pages, 271102 sentences and 3589220 non-tokenized words in their inflected and non-inflected forms. Among these total words there are 199245 tokens (i.e., distinct words) each of which appears in the corpus with different frequency of occurrence. For example while the word মাথা (māthā) “head” occurs 968 times, মাথার (māthār) “of head” occurs 398 times, মাথায় (māthāy) “on head” occurs 729 times followed by other inflected forms like মাথাতে (māthāte) “in head”, মাথাটা (māthāṭā) “the head”, মাথাটি (māthāṭi) “the head”, মাথাগুলো (māthāgulo) “heads”, মাথারা (māthārā) “heads”, মাথাদের (māthāder) “to the heads”, মাথারই (māthāri) “of head itself” with moderate frequency. This corpus is exhaustively used for the proposed work to extract sentences containing a particular ambiguous word.

The Bengali WordNet

The Bengali WordNet is an online semantic dictionary, used for obtaining the semantic information of a Bengali word (Dash 2012). It provides different information about Bengali words and also gives the relationship(s) existing between words. The Bengali WordNet is being developed with the help of tools provided by Indian Institute of Technology (IIT), Bombay. In this WordNet, a user can search for 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 category 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 presented 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),

(f) Semantic and lexical relations.

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

Parameters for Evaluating Performance

The efficiencies of the algorithms have been calculated using the conventional parameters as Precision, Recall, and F-Measure in the following way:

Precision (P) = (number of matched instances according to the human decision)/ (number of instances responded by the system).

Recall (R) = (number of matched instances according to the human decision)/ (total number of instances in the dataset) and

F-Measure = (2*P*R / (P+R)).

During the execution, the systems resolute the entire target words in the datasets either correctly or wrongly. So, the Precision and the Recall values are same.

Proposed Approach

First, in this work, the traditional Lesk algorithm has been developed as the baseline algorithm.

In the typical Lesk algorithm, a short phrase is selected from the sentence containing an ambiguous word. Next, the dictionary definition (gloss) of each of the senses of that ambiguous word is compared with glosses of the other words in that particular phrase. An ambiguous word is assigned with that particular sense, whose gloss has highest number of overlap (number of words in common) with the glosses of other words in that phrase.

Implementation of the Traditional Lesk Algorithm for WSD in Bengali

First of all, the traditional Lesk algorithm has been developed for Bengali language. As a knowledge base, Bengali WordNet has been used and the length of the phrase (size of the window of words) has been taken as 5 (five) including the ambiguous word.

The algorithm has been tested on 9 (nine) ambiguous words. But, the derived accuracy was very poor due to the scarcity of sufficient data in the Bengali WordNet for overlapping operation. This scarcity of overlap has a direct impact on the result.

Pre-Preprocessing

The following preprocessing tasks have been performed in this work-

Selection of Senses of the Ambiguous Words for Evaluation

The **Sense Inventory** for this approach has been considered as the sense definitions, mentioned in the Bengali WordNet.

Table 1: Dictionary definitions of the ambiguous words as defined in the Bengali WordNet

Word(sense1. POS: Gloss; sense2. POS: Gloss; ….. ; senseN. POS: Gloss)
পাতা(a. NOUN: নব উদ্ভূত কোমল পাতা কিশলয় নব পল্লব; b. NOUN: চোখের ওপরের চামড়ার পর্দা যা পরার ফলে তা বন্ধ হয়ে যায়; c. NOUN: কোনও বই বা খাতা ইত্যাদিতে লাগানো বস্তু যার দুদিকেই লেখা হয়; d. VERB: পাতা রয়েছে যা বিছানো)
শব্দ(a. NOUN: আক্ষর বা বর্ণ ইত্যাদি দিয়ে তৈরি আর মুখ দিয়ে উষ্চারিত অথবা লেখা সেই সংকেত; b. NOUN: সেই বস্তু যা শোনা যায় ধ্বনি আওয়াজ স্বর নাদ নিনাদ)
মুখ(a. NOUN: কোনো বস্তুর উপরের বা বাইরের খোলা অংশ; b. NOUN: সেই অঙ্গ যা দিয়ে প্রাণীরা কথা বলে এবং ভোজন করে; c. VERB: প্রায় কোনো সংকেত বিন্দুর নিরিখে কোনো বিশেষ দিকে থাকা)
ফল(a. NOUN: কোনও বিশিষ্ট ঋতুতে ফুল থেকে উত্পন্ন হওয়া শাঁস বা বীজে ভরা বীজকোষ; b. VERB: লাভপ্রদ হওয়া)
মাথা(a. NOUN: শরীরের সেই অংশ যার মধ্যে মস্তিষ্ক থাকে; b. NOUN: কোনো উঁচু ভবন মহল প্রভৃতির শিখর; c. NOUN: এমন কাজ মস্তিষ্কের খুব বেশী শক্তি ব্যয় করে)
পা(a. NOUN: সেই পরিমাণ দূরত্ব যা এক বারে যাওয়া যায়; b. NOUN: ব্যক্তির পায়ের সবথেকে নীচের অংশ যার উপর তিনি খাড়া হন বা যার সাহায্যে চলে; c. VERB: যাওয়ার জন্য পা উঠিয়ে অগ্রসর হওয়া)
সময়(a. NOUN: ইতিহাসে প্রায় নির্দিষ্ট সময়সীমা; b. VERB: সময় ঠিক করা)
নাম(a. VERB: এমন কিছু করা যাতে খ্যাতি বাড়ে; b. NOUN: সেই শব্দ যার দ্বারা কোনো বস্তুব্যক্তি প্রভৃতির বোধ হয় বা তাকে ডাকা হয়; c. NOUN: ঈশ্বরের নামের জপ)
ঘণ্টা(a. NOUN: সময় সূচিত করার জন্য যে ঘণ্টা বাজানো হয়; b. NOUN: ষাঠ মিনিটের সময়)

Text Normalization

The texts collected from the Bengali corpus are not adequately normalized. So, a manual text normalization procedure has been used before the work, as, (a) detachment of punctuation marks like single quote, tilde, double quote, parenthesis, comma, etc. that are attached to the words; (b) conversion of dissimilar fonts into similar one; (c) removal of angular brackets, uneven spaces, broken lines, slashes, etc. from sentences; and (d) identification of sentence terminal markers (i.e., dāri, note of exclamation, and note of interrogation), etc.

Non-functional word removal

In Bengali, the definition of non-functional word varies work to work. No specific rule is there to identify the non-functional words. Still, to reduce the size of the data to some manageable length, few less informative words have been considered as non-functional words. For example, all postpositions, e.g., দিকে (dike) "towards", প্রতি (prati) "per", etc.; conjunctions, e.g., এবং (ebang) "and", কিন্তু (kintu) "but", etc.; interjections, e.g., বা! (bāh) "well", আহা (āhā) "ah!", etc.; pronouns, e.g., আমি (āmi) "I", তুমি (tumi) "you", সে (se) "she" etc.; all articles, e.g., একটি (ekṭi) "one", etc. and proper nouns, e.g., রাম (rām) "Ram", কলকাতা (kalkātā) "Calcutta", etc. have been considered as stop words in this work.

Text Lemmatization

To increase the lexical coverage of the data, texts are lemmatized before the work. In this work, the texts are lemmatized using a Bengali lemmatizer tool. As the lemmatization tool could not produce cent percent accurate result, the system generated lemmatized texts have been manually modified upto certain extent.

After the execution of the traditional Lesk algorithm, the following output has been achieved-

Table 2: Result obtained from the typical Lesk algorithm.

Word	Number of test instance	Number of correctly evaluated instance	Accuracy using typical Lesk
পাতা (pātā)	70	17	24%
শব্দ (shabda)	50	16	32%
মুখ (mukh)	70	23	33%
ফল (fal)	50	14	28%
মাথা (māthā)	35	18	51%

পা (pā)	50	12	24%
সময় (samay)	60	12	20%
নাম (nām)	50	12	24%
ঘন্টা (ghantā)	50	20	40%
total	**485**	**365**	**31%**

The baseline accuracy has been achieved 31% using the typical Lesk approach.

Extension to the Baseline Method

In the extension of the proposed approach, to overcome this bottleneck of scarcity of information in the test sentences and in the WordNet, the following extensions have been adopted -

a) The window size has not been limited to a particular size, rather all the words of the sentence except the stop words, have been considered as the collocating words.

b) Glosses of the collocating words and their synonymous words have also been retrieved from the WordNet, because, the information obtained from the glosses are insufficient for an overlap to be occurred.

c) To increase the number of overlap, the example sentences, given in the sense definition of the words in the WordNet, have also been considered.

In the proposed approach (refer Fig. 1), first the input sentences (say: S_1, S_2, …, S_n) have been retrieved randomly from the Bengali Text Corpus.

The text, retrieved from the Bengali corpus was non-normalized in nature, so it has been normalized manually.

Next, the normalized text is lemmatized. The data set is lemmatized by a Bengali Lemmatizer tool.

After removing the stop words from each sentence (S_i: i=1,…,n), the meaningful words and the synonymous words of the meaningful words have been accumulated with the help of the Bengali WordNet.

Next, the glosses and example sentences of each of these words have been retrieved from the WordNet and concatenated to form a string.

Then, the individual sense-carrying gloss of the ambiguous word has been compared with the string to find the overlap.

The maximum overlap resolute the actual sense of the ambiguous word in that particular context.

Algorithm: Extended-Sense-Overlap

The algorithm of the proposed approach is given below-

Input: Sentences from the Bengali corpus.

Output: Sense of the ambiguous word in that particular sentence.

Step 1: Sentences (S_1, S_2, …, S_n) are collected from the Bengali corpus.

Step 2: Annotation and lemmatization task performed.

Step 3: Repeat step 4 to 12 for each sentence (S_i).

Step 4: Stop words are removed from the sentence.

Step 5: Meaningful words and their synonymous words are accumulated with the help of WordNet.

Step 6: Repeat step 7 for individual word.

Step 7: The gloss and the example sentence of the word are retrieved and concatenated to form a string.

Step 8: End of Step 6 loop.

Step 9: Repeat Step 10 for individual sense-representing gloss of the ambiguous word.

Step 10: Overlap is calculated between individual sense-representing gloss of the ambiguous word and the string, generated at Step 7.

Step 11: End of Step 9 loop.

Step 12: The actual sense of the ambiguous word is represented from the maximum overlap.

Step 13: End loop of Step 3

Step 14: Stop.

Complexity: The complexity of the algorithm is $O(n.s.g)$, where n represents the number of senses of the ambiguous word according to the WordNet; s represents the number of words in the gloss of the ambiguous word; and g represents the number of words in the gloss of the collocating words.

Flow Chart

The flow chart of the proposed approach is given below-

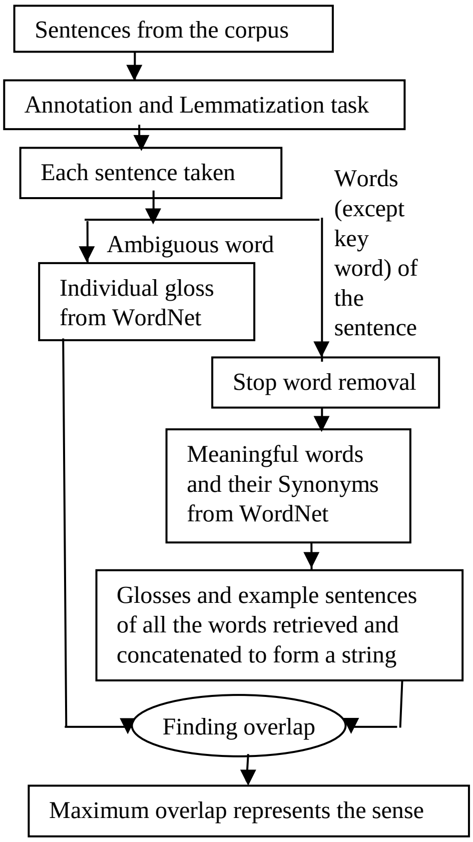

Sentences from the corpus
Annotation and Lemmatization task
Each sentence taken
Ambiguous word
Words (except key word) of the sentence
Individual gloss from WordNet
Stop word removal
Meaningful words and their Synonyms from WordNet
Glosses and example sentences of all the words retrieved and concatenated to form a string
Finding overlap
Maximum overlap represents the sense

Figure 1: Flow chart of the proposed approach

RESULTS AND CORRESPONDING EVALUATIONS

The algorithm has been tested on 485 sentences of 9 ambiguous words. The 9 ambiguous words are selected as, পাতা (pātā), শব্দ (shabda), মুখ (mukh), ফল (fal), মাথা (māthā), পা (pā), সময় (samay), নাম (nām) and ঘন্টা (ghantā).

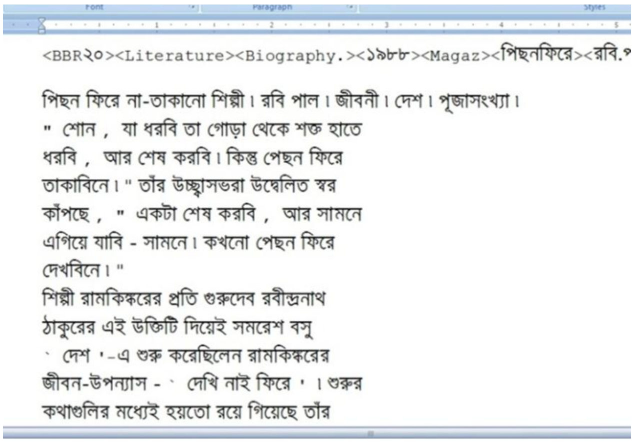
The efficiency of the algorithm has been calculated using Precision, Recall, and F-Measure as mentioned in previously.

The system resolutes the entire target words in the dataset either correctly or wrongly. So, the Precision and the Recall values are same, as P=R=365/485=0.75 and F-Measure= 0.75.

The following steps have been executed to evaluate the system-

A Sample Non-normalized Text

The text, retrieved from the Bengali text corpus is non-normalized in nature (refer Figure 2).



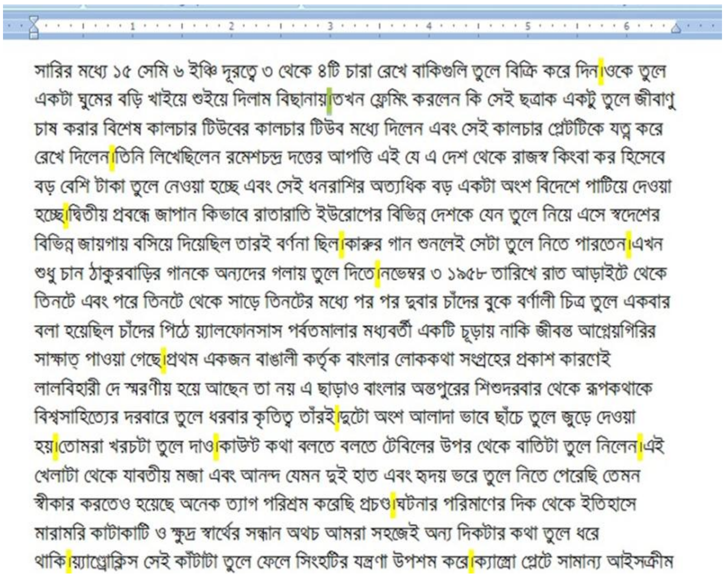
<BBR২০><Literature><Biography.><১৯৮৮><Magaz><পিছনফিরে><রবি.প

পিছন ফিরে না-তাকানো শিল্পী। রবি পাল। জীবনী। দেশ। পূজাসংখ্যা।
" শোন , যা ধরবি তা গোড়া থেকে শক্ত হাতে
ধরবি , আর শেষ করবি। কিন্তু পেছন ফিরে
তাকাবিনে। " তাঁর উচ্ছ্বাসভরা উদ্বেলিত স্বর
কাঁপছে , " একটা শেষ করবি , আর সামনে
এগিয়ে যাবি - সামনে। কখনো পেছন ফিরে
দেখবিনে। "
শিল্পী রামকিঙ্করের প্রতি গুরুদেব রবীন্দ্রনাথ
ঠাকুরের এই উক্তিটি দিয়েই সমরেশ বসু
` দেশ '-এ শুরু করেছিলেন রামকিঙ্করের
জীবন-উপন্যাস - ` দেখি নাই ফিরে ' । শুরুর
কথাগুলির মধ্যেই হয়তো রয়ে গিয়েছে তাঁর

Figure 2: A sample non-normalized text from the Bengali corpus

A Sample Text after Manual Text Normalization

A sample normalized text, generated after the normalization process is given below (refer Figure 3).

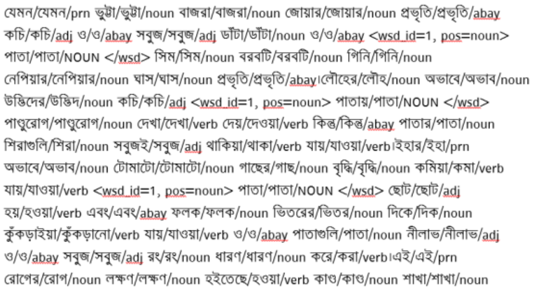


সারির মধ্যে ১৫ সেমি ৬ ইঞ্চি দূরত্বে ৩ থেকে ৪টি চারা রেখে বাকিগুলি তুলে বিক্রি করে দিন।ওকে তুলে
একটা ঘুমের বড়ি খাইয়ে শুইয়ে দিলাম বিছানায়।তখন ফ্লেমিং করলেন কি সেই ছত্রাক একটু তুলে জীবাণু
চাষ করার বিশেষ কালচার টিউবের কালচার টিউব মধ্যে দিলেন এবং সেই কালচার প্লেটটিকে যত্ন করে
রেখে দিলেন।তিনি লিখেছিলেন রমেশচন্দ্র দত্তের আপত্তি এই যে এ দেশ থেকে রাজস্ব কিংবা কর হিসেবে
বড় বেশি টাকা তুলে নেওয়া হচ্ছে এবং সেই ধনরাশির অত্যধিক বড় একটা অংশ বিদেশে পাটিয়ে দেওয়া
হচ্ছে।দ্বিতীয় প্রবন্ধে জাপান কিভাবে রাতারাতি ইউরোপের বিভিন্ন দেশকে যেন তুলে নিয়ে এসে স্বদেশের
বিভিন্ন জায়গায় বসিয়ে দিয়েছিল তারই বর্ণনা ছিল।কারুর গান শুনলেই সেটা তুলে নিতে পারতেন।এখন
শুধু চান ঠাকুরবাড়ির গানকে অন্যদের গলায় তুলে দিতে।নভেম্বর ৩ ১৯৫৮ তারিখে রাত আড়াইটে থেকে
তিনটে এবং পরে তিনটে থেকে সাড়ে তিনটের মধ্যে পর পর দুবার চাঁদের বুকে বর্ণালী চিত্র তুলে একবার
বলা হয়েছিল চাঁদের পিঠে য়্যালফোনসাস পর্বতমালার মধ্যবর্তী একটি চূড়ায় নাকি জীবন্ত আগ্নেয়গিরির
সাক্ষাত্ পাওয়া গেছে।প্রথম একজন বাঙালী কর্তৃক বাংলার লোককথা সংগ্রহের প্রকাশ কারণেই
লালবিহারী দে স্মরণীয় হয়ে আছেন তা নয় এ ছাড়াও বাংলার অন্তপুরের শিশুদরবার থেকে রূপকথাকে
বিশ্বসাহিত্যের দরবারে তুলে ধরবার কৃতিত্ব তাঁরই।দুটো অংশ আলাদা ভাবে ছাঁচে তুলে জুড়ে দেওয়া
হয়।তোমরা খরচটা তুলে দাও।কাউন্ট কথা বলতে বলতে টেবিলের উপর থেকে বাতিটা তুলে নিলেন।এই
খেলাটা থেকে যাবতীয় মজা এবং আনন্দ যেমন দুই হাত এবং হৃদয় ভরে তুলে নিতে পেরেছি তেমন
স্বীকার করতেও হয়েছে অনেক ত্যাগ পরিশ্রম করেছি প্রচণ্ড।ঘটনার পরিমাণের দিক থেকে ইতিহাসে
মারামরি কাটাকাটি ও ক্ষুদ্র স্বার্থের সন্ধান অথচ আমরা সহজেই অন্য দিকটার কথা তুলে ধরে
থাকি।য়্যাগ্রোক্লিস সেই কাঁটাটা তুলে ফেলে সিংহটির যন্ত্রণা উপশম করে।ক্যাস্ত্রো প্লেটে সামান্য আইসক্রীম

Figure 3: A sample normalized text.

A Sample Input Data Set

A sample input file, after the lemmatization procedure is given below (refer Figure 4).

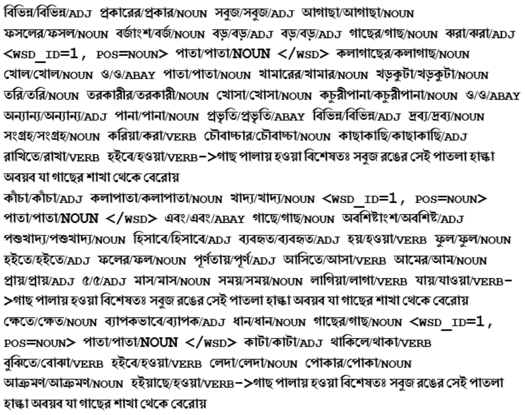


```
যেমন/যেমন/prn ভুট্টা/ভুট্টা/noun বাজরা/বাজরা/noun জোয়ার/জোয়ার/noun প্রভৃতি/প্রভৃতি/abay
কচি/কচি/adj ও/ও/abay সবুজ/সবুজ/adj ডাঁটা/ডাঁটা/noun ও/ও/abay <wsd_id=1, pos=noun>
পাতা/পাতা/NOUN </wsd> সিম/সিম/noun বরবটি/বরবটি/noun গিনি/গিনি/noun
নেপিয়ার/নেপিয়ার/noun ঘাস/ঘাস/noun প্রভৃতি/প্রভৃতি/abay।লৌহের/লৌহ/noun অভাবে/অভাব/noun
উদ্ভিদের/উদ্ভিদ/noun কচি/কচি/adj <wsd_id=1, pos=noun> পাতায়/পাতা/NOUN </wsd>
পাণ্ডুরোগ/পাণ্ডুরোগ/noun দেখা/দেখা/verb দেয়/দেওয়া/verb কিন্তু/কিন্তু/abay পাতার/পাতা/noun
শিরাগুলি/শিরা/noun সবুজই/সবুজ/adj থাকিয়া/থাকা/verb যায়/যাওয়া/verb।ইহার/ইহা/prn
অভাবে/অভাব/noun টোমাটো/টোমাটো/noun গাছের/গাছ/noun বৃদ্ধি/বৃদ্ধি/noun কমিয়া/কমা/verb
যায়/যাওয়া/verb <wsd_id=1, pos=noun> পাতা/পাতা/NOUN </wsd> ছোট/ছোট/adj
হয়/হওয়া/verb এবং/এবং/abay ফলক/ফলক/noun ভিতরের/ভিতর/noun দিকে/দিক/noun
কুঁকড়াইয়া/কুঁকড়ানো/verb যায়/যাওয়া/verb ও/ও/abay পাতাগুলি/পাতা/noun নীলাভ/নীলাভ/adj
ও/ও/abay সবুজ/সবুজ/adj রং/রং/noun ধারণ/ধারণ/noun করে/করা/verb।এই/এই/prn
রোগের/রোগ/noun লক্ষণ/লক্ষণ/noun হইতেছে/হওয়া/verb কাণ্ড/কাণ্ড/noun শাখা/শাখা/noun
```

Figure 4: A sample input text.

A Sample Output File

A sample output file, generated by the system is given below (refer Figure 5).



```
বিভিন্ন/বিভিন্ন/ADJ প্রকারের/প্রকার/NOUN সবুজ/সবুজ/ADJ আগাছা/আগাছা/NOUN
ফসলের/ফসল/NOUN বর্জাংশ/বর্জ/NOUN বড়/বড়/ADJ বড়/বড়/ADJ গাছের/গাছ/NOUN ঝরা/ঝরা/ADJ
<WSD_ID=1, POS=NOUN> পাতা/পাতা/NOUN </WSD> কলাগাছের/কলাগাছ/NOUN
খোল/খোল/NOUN ও/ও/ABAY পাতা/পাতা/NOUN খামারের/খামার/NOUN খড়কুটা/খড়কুটা/NOUN
তরি/তরি/NOUN তরকারীর/তরকারী/NOUN খোসা/খোসা/NOUN কচুরীপানা/কচুরীপানা/NOUN ও/ও/ABAY
অন্যান্য/অন্যান্য/ADJ পানা/পানা/NOUN প্রভৃতি/প্রভৃতি/ABAY বিভিন্ন/বিভিন্ন/ADJ দ্রব্য/দ্রব্য/NOUN
সংগ্রহ/সংগ্রহ/NOUN করিয়া/করা/VERB চৌবাচ্চার/চৌবাচ্চা/NOUN কাছাকাছি/কাছাকাছি/ADJ
রাখিতে/রাখা/VERB হইবে/হওয়া/VERB->গাছ পালায় হওয়া বিশেষতঃ সবুজ রঙের সেই পাতলা হাল্কা
অবয়ব যা গাছের শাখা থেকে বেরোয়
কাঁচা/কাঁচা/ADJ কলাপাতা/কলাপাতা/NOUN খাদ্য/খাদ্য/NOUN <WSD_ID=1, POS=NOUN>
পাতা/পাতা/NOUN </WSD> এবং/এবং/ABAY গাছে/গাছ/NOUN অবশিষ্টাংশ/অবশিষ্ট/ADJ
পশুখাদ্য/পশুখাদ্য/NOUN হিসাবে/হিসাবে/ADJ ব্যবহৃত/ব্যবহৃত/ADJ হয়/হওয়া/VERB ফুল/ফুল/NOUN
হইতে/হইতে/ADJ ফলের/ফল/NOUN পূর্ণতায়/পূর্ণ/ADJ আসিতে/আসা/VERB আমের/আম/NOUN
প্রায়/প্রায়/ADJ ৫/৫/ADJ মাস/মাস/NOUN সময়/সময়/NOUN লাগিয়া/লাগা/VERB যায়/যাওয়া/VERB-
>গাছ পালায় হওয়া বিশেষতঃ সবুজ রঙের সেই পাতলা হাল্কা অবয়ব যা গাছের শাখা থেকে বেরোয়
ক্ষেতে/ক্ষেত/NOUN ব্যাপকভাবে/ব্যাপক/ADJ ধান/ধান/NOUN গাছের/গাছ/NOUN <WSD_ID=1,
POS=NOUN> পাতা/পাতা/NOUN </WSD> কাটা/কাটা/ADJ থাকিলে/থাকা/VERB
বুঝিতে/বোঝা/VERB হইবে/হওয়া/VERB লেদা/লেদা/NOUN পোকার/পোকা/NOUN
আক্রমণ/আক্রমণ/NOUN হইয়াছে/হওয়া/VERB->গাছ পালায় হওয়া বিশেষতঃ সবুজ রঙের সেই পাতলা
হাল্কা অবয়ব যা গাছের শাখা থেকে বেরোয়
```

Figure 5: A sample output text.

The percentage of accuracy obtained from the system is presented below:

Table 3: Overall result on 9 (nine) ambiguous words

Word	Number of test instance	Number of correctly evaluated instance	Accuracy after context expansion
পাতা (pātā)	70	59	84%
শব্দ (shabda)	50	42	84%
মুখ (mukh)	70	51	73%
ফল (fal)	50	39	78%
মাথা (māthā)	35	20	57%

পা (pā)	50	42	84%
সময় (samay)	60	42	70%
নাম (nām)	50	33	66%
ঘণ্টা (ghantā)	50	37	74%
total	**485**	**365**	**75%**

Few Close Observations

Two major bottlenecks have been observed in this methodology-

First one is same as the other approaches, as the vast semantic nature of the Bengali language. The supporting data sets to represent a particular sense are so different that overlap could not occur in few cases.

Secondly, the scarcity of data in the Bengali WordNet is a big issue. As the Bengali WordNet is under development, it is not a complete reference of knowledge base in Bengali language.

And, few other important issues also appeared as obstacles during the implementation of the system, as-

a) Vast semantic nature of Bengali language, b) Very large sentences with many irrelevant contextual words, c) Very short sentences with lack of sufficient information within them, d) Spelling errors of few words, e) Usefulness of few non-functional words in Bengali etc.

CONCLUSION AND FUTURE WORK

In the proposed work, WSD in Bengali language has been presented using knowledge based methodology. Although, the work has been tested on Bengali language, this Context Expansion technique could be used for any low resource language. In the Bengali WordNet, the synset hierarchy is not present which could expand the context in more effective way.

In a close observation, it is noticed that, although the collocating words of a key word have multiple meanings in the WordNet, they don’t participate in lexical overlap as their sense domains are different.

REFERENCES

[1] N. Ide and J. Véronis. “Word Sense Disambiguation: The State of the Art”, ComputationalLinguistics, Vol. 24, No. 1, Pp. 1-40, 1998.

[2] R.S. Cucerzan, C. Schafer and D. Yarowsky, “Combining classifiers for word sense disambiguation,” Natural Language Engineering, Vol. 8, No. 4, Cambridge University Press, Pp. 327-341, 2002.

[3] M. S. Nameh, M. Fakhrahmad and M.Z. Jahromi, “A New Approach to Word Sense Disambiguation Based on Context Similarity,” Proceedings of the World Congress on Engineering, Vol. I, 2011.

[4] W. Xiaojie and Y. Matsumoto, “Chinese word sense disambiguation by combining pseudo training data”, Proceedings of The International Conference on Natural Language Processing and Knowledge Engineering, Pp. 138-143, 2003.

[5] R. Navigli, “Word Sense Disambiguation: a Survey”, ACM Computing Surveys, Vol. 41, No.2, ACM Press, Pp. 1-69, 2009.

[6] W. Xiaojie and Y. Matsumoto, “Chinese word sense disambiguation by combining pseudo training data”, Proceedings of The International Conference on Natural Language Processing and Knowledge Engineering, Pp. 138-143, 2003.

[7] M. Sanderson, “Word Sense Disambiguation and Information Retrieval,” 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, 1994.

[8] Word Sense Disambiguation; Algorithms and Applications, Edited by Eneko Agirre and Philip Edmonds, Springer, VOLUME 33.

[9] H. Seo, H. Chung, H. Rim, S. H. Myaeng and S. Kim, “Unsupervised word sense disambiguation using WordNet relatives,” Computer Speech and Language, Vol. 18, No. 3, Pp. 253-273, 2004.

[10] G. Miller, “WordNet: An on-line lexical database,” International Journal of Lexicography,Vol.3,No. 4, 1991.

[11] S.G. Kolte and S.G. Bhirud, “Word Sense Disambiguation Using WordNet Domains,” First International Conference on Digital Object Identifier, Pp. 1187-1191, 2008.

[12] Y. Liu, P. Scheuermann, X. Li and X. Zhu, “Using WordNet to Disambiguate Word Senses for Text Classification,” Proceedings of the 7th International Conference on Computational Science, Springer Verlag, Pp. 781 – 789, 2007.

[13] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller, "WordNet An on-line Lexical Database," International Journal of Lexicography, 3(4): 235-244, 1990.

[14] G.A. Miller, "WordNet: A Lexical Database," Comm. ACM, Vol. 38, No. 11, Pp. 39-41, 1993.

[15] A.J. Cañas, A. Valerio, J. Lalinde-Pulido, M. Carvalho, and M. Arguedas, "Using WordNet for Word Sense Disambiguation to Support Concept Map Construction," String Processing and Information Retrieval, Pp. 350-359, 2003.

[16] C. Marine, W.U. Dekai, "Word Sense Disambiguation vs. Statistical Machine Translation," Proceedings of the 43rd Annual Meeting of the ACL , Ann Arbor, pages 387–394, June 2005.

[17] http://www.ling.gu.se/~sl/Undervisning/StatMet11/wsd-mt.pdfdate: 14/05/2015

[18] http://nlp.cs.nyu.edu/sk-symposium/note/P-28.pdf date: 14/05/2015

[19] S. C. Yee, T. N. Hwee, C. David, "Word Sense Disambiguation Improves Statistical Machine Translation," Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 33–40, Prague, Czech Republic, June 2007.

[20] R. Mihalcea and D. Moldovan. An iterative approach to word sense disambiguation. In Proceedings of Flairs 2000, pages 219–223, Orlando, FL, May 2000.

[21] S. Christopher, P. O. Michael, T. John, "Word Sense Disambiguation in Information Retrieval Revisited," SIGIR'03, July 28–August 1, 2003, Toronto, Canada, 2003.

[22] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.6828 &rep=rep1&type=pdf date: 14/05/2015

[23] http://www.aclweb.org/anthology/P12-1029 date: 14/05/2015

[24] https://www.comp.nus.edu.sg/~nght/pubs/esair11.pdf date: 14/05/2015

[25] http://cui.unige.ch/isi/reports/2008/CLEF2008-LNCS.pdf date: 14/05/2015

[26] S. Banerjee, T. Pedersen, "An adapted Lesk algorithm for word sense disambiguation using WordNet," In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February, 2002.

[27] M. Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," Proceedings of SIGDOC, 1986.

[28] http://www.dlsi.ua.es/projectes/srim/publicaciones/CICling-2002.pdf date: 14/05/2015

[29] K. Mittal and A. Jain, "Word Sense Disambiguation Method using Semantic Similarity Measures and OWA Operator," ICTACT Journal on Soft Computing: Special Issue on Soft-Computing Theory, Application and Implication in Engineering and Technology, January, 2015, volume: 05, Issue: 02, 2015.

[30] http://www.d.umn.edu/~tpederse/Pubs/cicling2003-3.pdf date: 14/05/2015

[31] http://www.aclweb.org/anthology/U04-1021 date: 14/05/2015

[32] http://www.aclweb.org/anthology/C10-2142 date: 14/05/2015

[33] M.C. Diana, J. Carroll, "Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences," Computational Linguistics, Volume 29, Number 4, pp. 639-654.

[34] Y. Patrick and B. Timothy, "Verb Sense Disambiguation Using Selectional Preferences Extracted with a State-of-the-art Semantic Role Labeler," Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), pages 139–148, 2006.

[35] http://link.springer.com/article/10.1023/A%3A1002674829964#page-1 date: 14/05/2015

[36] S. Parameswarappa, and V.N, Narayana, "Kannada Word Sense Disambiguation Using Decision List," Volume 2, Issue 3, May – June 2013, pp. 272-278, 2013.

[37] http://www.academia.edu/5135515/Decision_List_Algorithm_for_WSD _for_Telugu_NLP

[38] Ted Pedersen, "Word Sense Disambiguation: Algorithms and Applications‖," Edited by- Eneko Agirre, Philip Edmonds, Springer, VOLUME-33.

[39] http://link.springer.com/article/10.1023/A%3A1002674829964#page-1 date: 14/05/2015

[40] Lesk, M., "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", Proceedings of SIGDOC, 1986.

[41] http://www.dlsi.ua.es/projectes/srim/publicaciones/CICling-2002.pdf date: 14/05/2015

[42] http://www.d.umn.edu/~tpederse/Pubs/cicling2003-3.pdfdate: 14/05/2015

[43] http://www.aclweb.org/anthology/U04-1021 date: 14/05/2015

[44] http://www.aclweb.org/anthology/C10-2142 date: 14/05/2015

[45] Diana, M.C., Carroll, J., “Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences”, Computational Linguistics, Volume29,Number 4, pp. 639-654.