Word Sense Disambiguation in Bengali applied to Bengali-Hindi Machine Translation

Ayan Das
Dept. of Computer Science and Engg.
Indian Institute of Technology
Kharagpur, India
ayandas84@gmail.com

Sudeshna Sarkar
Dept. of Computer Science and Engg.
Indian Institute of Technology
Kharagpur, India
shudeshna@gmail.com

Abstract

We have developed a word sense disambiguation(WSD) system for Bengali language and applied the system to get correct lexical choice in Bengali-Hindi machine translation. We are not aware of any existing system for Bengali WSD. Since there is no sense annotated Bengali corpus or sufficient amount of parallel corpus for Bengali-Hindi language pair, we had to use an unsupervised approach. We use a graph based method to find sense clusters in Bengali language. Following this we use a vector space based approach to map these sense clusters to Hindi translations of the target word and to predict translation of the target word in test instance. We used monolingual Bengali and Hindi corpora and the available Bengali and Hindi wordnet and bilingual sense dictionary.

1 Introduction

Word Sense Disambiguation (WSD) is defined as the task of finding the correct sense of a word in a context, when the word has multiple meanings. The identifiation of the correct sense of a word in a context is useful for many applications including machine translation, information extraction and anaphora resolution. The various methods for word sense disambiguation can be broadly classified as knowledge/dictionary based, supervised, semi-supervised and unsupervised approaches. However, there is no strict boundary between these methods and combinations of these sometimes yield good results.

Our aim was to improve Bengali to Hindi machine translation by incorporating word sense disambiguation. Specifically, we started with a basic Bengali-Hindi rule based machine translation system where the polysemous Bengali words are replaced by most frequent translation as learnt from some training data. We are required to find the sense of the word from contextual cues and suggest a suitable translation for the word in Hindi.

Although there is a lot of work in word sense disambiguation and its application in machine translation, there is little work in Indian languages. To the best of our knowledge this is the first attempt to develop a cross-lingual WSD system for Bengali to Hindi machine translation.

The supervised methods of WSD perform much better than the unsupervised systems. However, the supervised systems rely heavily on knowledge and require large volume of sense annotated corpus. Such corpus is not currently available for Bengali. Given these resource constraints, we needed to develop a WSD system for the Bengali language. Since it is expensive and time-consuming to develop sense annotated corpus, we decided to develop an unsupervised WSD system. The resources that we had are unannotated Bengali and Hindi News corpora, a Bengali sense dictionary and Hindi WordNet (Chakrabarti et al., 2002).

We use an unsupervised graph-based clustering approach for sense clustering and compare our method with two existing graph-based approaches, (Navigli and Crisafulli, 2010) and (Jurgens, 2011) for sense clustering.

Most work that use WSD to improve translation is based on bilingual parallel corpora. Since Bengali-Hindi parallel corpus is not available to us, we propose an approach for prediction of correct translation of polysemous Bengali words in a given context, in Bengali to Hindi machine translation using a vector space based model.

The rest of the paper is organized as follows. We discuss some of the related works in section 2. In Sections 3 and 4 we discuss our objective and some state-of-the-art works on graph-based WSD respectively. In Sections 5 and 6 we discuss our works in details. Finally, in Section 7 we give a detailed analysis of our results and Section 8 concludes the paper.

2 Related Work

Unsupervised methods have the potential to overcome the lack of large-scale corpus which has been manually annotated with word senses(Pedersen and Bruce, 1997). The unsupervised approaches can be broadly classifed into three subcategories: context clustering, word clustering and graph-based approaches. The graph based methods have been quite popular since the results are quite promising.

Schütze proposed an unsupervised approach based on the notion of context clustering (Schütze, 1998). The approach is based on the vector space model. Each

word is represented as a vector based on cooccurrence. The word vectors are clustered into groups based on cosine similarity and each group is considered as identifying a sense of the target word.

Lin (1998) proposed another method based on clustering semantically similar words. The construction of a cooccurrence graph based on grammatical relations between words in context was described by Widdows and Dorow (2002). The adjacency matrix of the graph is interpreted as a Markov chain and Markov clustering algorithm is used to identify the word senses. Véronis (2004) proposed a graph based approach called *HyperLex*. Some of the recent work on unsupervised graph-based word sense disambiguation are that of Navigli and Crisafulli (2010) and Jurgens (2011). Navigli and Crisafulli attempts to identify connected components in the local graph, built from the contexts retrieved for ambiguous query word. The graph-based algorithm exploits cycles of size 3 and 4 in the co-occurrence graph of the input query to detect the query words meanings.

Jurgens (2011) aggolmeratively clustered the edges of the cooccurrence graph constructed from the corpus, based on a similarity score between the edges using single-link criteria. The dendrogram was cut at a level at which, a function of the edge density of the clusters is maximum. Lee and Ng (2002) and Voorhees (1993) developed an unsupervised WSD algorithm that uses with other knowledge resources like dictionary.

As already mentioned, a major application of WSD is the improvement of performance of machine translation systems. Brown et al. (1991) developed a method to predict the translation of an ambiguous French word in English based on the assumption that an ambiguous word in French may be translated into different English words depending on the sense of use. Apidianaki (2009) reports an unsupervised method and lexical selection based approach that exploit the results of a data-driven sense induction from parallel corpora for cross-lingual WSD in English-Greek and Greek-to-English machine translation. Chen et al. (2010) proposed a vector space model based algorithm to compute sense similarity between lexical units (words, phrases, rules, etc.) and use it in statistical machine translation. Vintar et al. (2012) reported a series of experiments performed for English-Slovene language pair using UKB, a freely available graph-based WSD system. Apidianaki et al. (2012) proposed a method of integrating semantic information at two stages of translation process of a SMT system.

In Indian Languages, Khapra et. al. introduced an algorithm iWSD (iterative WSD)(Khapra et al., 2011b),(Khapra et al., 2009),(Mitesh M. Khapra and Bhattacharyya, 2010) which uses the WordNet. The monosemous words are initially tagged with their meaning and used as seed set. The senses of words in a sentence are then resolved iteratively ordered by increasing degree of polysemy. Mitesh M. Khapra and Bhattacharyya (2010) applied iWSD algorithm to data from health and tourism domains and used WordNet based features such as *conceptual distance, semantic graph distance*, *belongingness to dominant concept*, *sense distribution* and *corpus co-occurrence* as corpus features.

Khapra et al. (2011b; Khapra et al. (2011a; Mitesh M. Khapra and Sharma (2008) attempted a domain specific bilingual WSD using bilingual bootstrapping. A persistent problem was that, sometimes words onto which the parameters were projected(Mitesh M. Khapra and Bhattacharyya, 2009), were themselves polysemous. An Expectation-Maximization(EM) approach was used to settle on fixed parameter values by *parameter projection* from one language to another at each alternate iteration of the EM approach.

3 Objective

Most work on application of WSD to machine translation is based on statistical machine translation(SMT) framework and use parallel corpus. Our work aims to use WSD in rule-based machine translation system for resource poor language pairs where parallel corpus is not available. From the literature, we observed that the supervised methods perform much better than the unsupervised methods. However, supervised methods require large volume of sense annotated corpus. Due to lack of such sense annotated corpus in Bengali and time and effort involved in doing this we decided to use some unsupervised method for WSD and unsupervised method to map sense clusters in source language to appropriate translation in target language.

4 Some existing approaches to WSD implemented for Bengali

Graph-based approaches have been quite successful in unsupervised word sense disambiguation, so we decided to work on graph-based WSD system for Bengali. We have studied the performance of two successful WSD methods suggested by Navigli and Crisafulli(2010) and Jurgens(2011) in Bengali. These are reported in 4.1 and 4.2 respectively.

4.1 Navigli and Crisafulli's approach

In this approach (Navigli and Crisafulli, 2010), the corpus is queried with a target word and a cooccurrence graph is constructed from the contexts after removal of the target word from the contexts. The main idea behind this approach is that edges in the cooccurrence graph participating in cycles are likely to connect vertices (i.e. words) belonging to the same meaning component. The work focuses on cycles of length 3 (triangle) and 4 (square). The edge weights are equal to the Dice-coefficient of cooccurrence of two words in the retrieved context set and edges with weight below a threshold are removed. Each of the remaining edges are assigned weight equal to the *triangle* and *square* scores and edges with score below a threshold value are removed. Finally, all the connected components of size

greater than a threshold are identified and each such component is assumed to contain words that together indicate a distinct sense of the target word.

On implementing this method in Bengali, we observed that the value of the cut-off (*triangle/square*) scores in the Navigli and Crisafulli (2010) approach varies widely for different target words depending on the volume of data retrieved as a result of querying the corpus. Secondly, the connected components obtained by this method are non-overlapping. Hence, any word can occur in at most a single cluster. However, some words may be indicators of multiple senses of the target polysemous word. Although the overall system performs better with square score, we observed that for some words triangle score performs better.

4.2 Jurgens' approach

Jurgens (2011) proposed a community detection algorithm from a cooccurrence graph constructed from the nouns in the corpus that occur with frequency greater than a threshold. Initially, similarity between each edge pair is computed by a scoring function which equals zero if the edges do not share any vertex and is the ratio of the number of common neighbors and the total number of neighbors of the two vertices apart from the common vertex. Finally, the edges are aggolmeratively clustered by single-link criteria. The construction of the dendrogram is stopped when the sum of the edge density in the clusters is highest.

We observed that the graph becomes excessively large size when the whole corpus was used for the graph construction. It consumes lot of memory and the clustering process becomes extremely slow and time-consuming. To make the algorithm more scalable, we introduced a modification to this approach. Keeping the algorithm and the scoring function the same we executed the algorithm on the cooccurrence graph constructed from the contexts retrieved by querying the corpus with the target word and removing the target word from the contexts, instead of building the graph from the whole corpus.

We observed that some clusters are highly overlapping and the number of clusters formed for a target word are very large. Moreover, it is not possible to control the degree of overlap between the clusters.

5 Our Approach to WSD based on community detection

5.1 Motivation

As discussed in sections 4.1 and 4.2, the approaches have certain disadvantages. Our aim was to find a method that addresses these issues. We wish to find an algorithm that generates relatively less number of clusters and at the same time the degree of overlap among the clusters should not be very high. We decided to use an algorithm that focusses on the edge-density of a graph and identifies clusters within the graph in which the words(vertices) are more strongly connected amongst themselves than with the vertices outside the community. Our intuition was that, a set of words that cooccur very strongly tend to form a clique or a very dense subgraph within the cooccurrence graph where edge density within these subgraphs is much higher than the edge density between any two such clusters i.e., the internal edge density of such a community should be greater than its external edge density. We can map this concept to WSD as, if a word is a stronger indicator of a particular sense of a target ambiguous word then it should have more number of neighbors in a particular community than other communities and addition of that word into the community is expected to increase the fitness function, discussed in 5.2, value of the community. Based on these observations, we propose a community detection based approach.

5.2 Overview

In our approach, we treat WSD as a community detection problem. We extract the contexts containing a target word from the Bengali corpus and build a cooccurrence graph. In this cooccurrence graph, a community detection algorithm is used to detect communities. In order to extract the sense-specific information corresponding to a target word we hypothesize that each community in the co-occurrence graph provides contextual information that indicates a single sense of the target word. In the co-occurrence graph, each vertex represents a word in the corpus and an edge exists between two vertices if the two words co-occur in a sentence (a context). The weight of an edge is equal to the number of contexts in which the two words co-occur. We studied some of the existing community detection algorithms and decided to use Greedy Clique expansion algorithm(GCE)(Lee et al., 2010). The summary of the steps of the approach for word clustering is given in Algorithm 1:

We now describe Lee et al. (2010)'s algorithm. The community detection algorithm is fundamental to finding the contextual information for sense induction. The input to this algorithm are the graph and four parameters : the minimum clique size (k), the scaling parameter of the fitness function α, minimum overlap degree of initial cliques and minimum overlap degree of final communities (ϵ). We used the same fitness function and the seed detection algorithm as used in Lee et al. (2010). We use the notation used in Lee et al. (2010) to explain the the main steps of the algorithm.

The community fitness function is a measure of the degree to which a induced subgraph S of the graph G corresponds to the notion of community. It takes an induced subgraph S of the graph G as input and returns a *real-valued fitness value* as output. The fitness function of a community S is defined by the in terms of $S's$ internal degree k_{in}^{s} and external degree k_{out}^{s}. k_{in}^{s} is equal to twice the number of edges that both start and end in

Glosses of Bengali words on which we tested our system				
Word	Sense 1	Sense 2	Sense 3	Sense 4
আচার(achar)	Ritual	Pickle		
অর্থ(artha)	Money	Meaning		
চাল(chaal)	Rice	Maneuver	Roof	
ডাল(daal)	Branch of a tree	Pulses		
গোলা(gola)	Shell/Cannon ball	Place to store harvest		
জাল(jaal)	Forge	Net	Trap	Network
কেন্দ্র(kendra)	Center	regarding	Central Government	Place intended for some purpose
লক্ষ্য(lakshya)	Aim	Purpose	Observation	
প্রণালী(pranaali)	Recipe	Strait		
রাস্তা(raasta)	Road	Method	Option	

Table 1: Glosses of Bengali words on which we tested our system

Algorithm 1: Steps followed in word clustering

input : Bengali Corpus, query word
output: Bengali word clusters with respect to the query word

1 Query the Bengali corpus with the target word to retireve the sentences containing the target word;
2 From the retrieved sentences, remove the target word and select the nouns with frequency above a threshold value;
3 Build the co-occurrence graph with the words selected in the previous step where edge weight between two nodes is the number of times the two words occur together in a sentence;
4 Set a threshold for the edge weight and remove the edges from the graph with weight below that threshold;
5 Perform community detection on the graph. We use GCE (Lee et al., 2010) implementation from *https://sites.google.com/site/greedycliqueexpansion*;

S and k^s_{out} is the number of edges that have only one edge in S. The community fitness is defined as:

$$F_s = \frac{k^s_{in}}{(k^s_{in} + k^s_{out})^\alpha}$$

where α is the parameter that can be tuned.

Let S be an induced subgraph of graph G which is a seed or core of a community C. In other words, S is embedded in some larger community C, such that all of its nodes are part of C, but not all nodes in C are included in S. S has to be expanded by adding nodes to it until it includes all nodes in C. The technique is summarized as follows:

- For each node v in the boundary of the seed community S, the extent to which inclusion of v increases or decreases the fitness of S, is computed.
- The node with highest fitness value, v_{max}, is selected.
- If v_{max} has a positive fitness value, then add it to S and loop back to step 1. Else, stop and return S.

The algorithm proceeds as follows;

1. Find the seeds (maximal cliques of size atleast equal to k) in the graph.
2. Choose the largest unexpanded seed, create a candidate community C' and continue expanding the seed with a community fitness function F until addition of any node would lower fitness.
3. If C' is within ϵ of any already accepted community C, then C' is discarded as a near duplicate of C. Otherwise, if no near duplicates are found, C' is accepted.
4. Continue to loop back to step 2 until no seed is remaining.

6 Going from WSD to word selection for Bengali to Hindi translation

In Section 5 we looked at methods for finding sense clusters in Bengali. In phase 2, we find a mapping to a Hindi word for each sense cluster in Bengali and predict the translation of a polysemous Bengali word in Hindi in a given context.

The second phase proceeds in two stages. First, the clusters obtained from the cooccurrence graphs using the algorithms described so far are mapped to some Hindi translation of the target Bengali word and second, these tagged clusters in turn are used to suggest a suitable translation of occurrence of the target word in a test context.

From the literature, we observed that most of the works on application of WSD for improving machine translation use parallel corpora or wordnet relations to identify suitable translation of a polysemous word to the target language. Since we do not have Bengali-Hindi parallel corpora, we propose a vector space based approach for using the clusters and a comparable Bengali-Hindi corpora for predicting the correct translation of a polysemous Bengali word to Hindi. The steps are described below.

6.1 Step 1: Construct the reference vectors

For a Bengali target word, the identifiers of all the concepts (synids) in which the word occurs in the Bengali

sense dictionary are extracted. There exists an one-to-one mapping between the concept ids in Bengali sense dictionary and Hindi wordnet. This feature is utilized to get the synsets from the Hindi wordnet corresponding to the sense ids so obtained. All the words in the Hindi synsets are combined to form a set of unique Hindi words, which is expected to be the set of all possible translations of the target Bengali word in Hindi. For every Hindi word in the set, the sentences containing that word are extracted from the Hindi corpus. All such set of sentences are combined to form a bag of words. A set of all words that occur in the bag of words is used to define a vector space where each Hindi word corresponds to a dimension in the vector space. For each word w in the set of possible Hindi translations of the target word, we define a reference vector such that the magnitude in any direction is equal to the tf-idf score of the word corresponding to that dimenson in the contexts containing w.

6.2 Step 2: Labeling the clusters with Hindi candidate words

The steps in tagging the Bengali sense clusters in Hindi are given in Algorithm 2:

Algorithm 2: Labelling the clusters with Hindi words

input : Word clusters corresponding to the Bengali query word, Bengali-Hindi bilingual dictionary
output: Assignment of a Hindi word to each cluster

1 Translate the Bengali clusters obtained from the *community detection* algorithm to Hindi using a bilingual dictionary;
2 Project each cluster onto the vector space defined in 6.1;
3 find the cosine similarity of each cluster with the reference vectors;
4 Label each cluster with the Hindi translation whose reference has the highest cosine similarity with the cluster;

6.3 Step 3: Prediction of translation of target word using the clusters

The steps involved in prediction of Hindi translation of target Bengali word from the clusters are listed in Algorithm 3:

Algorithm 3: Prediction of translation of target word using the clusters

input : Target word, Test sentence, Bengali word clusters tagged with corresponding Hindi word
output: Translation of the target word in the test context

1 Convert each word in the new test context into their root forms;
2 Extract only nouns from the context;
3 Generate vector space such that each of the unique words occurring in any of the Bengali clusters obtained for a target Bengali word is a dimension;
4 Project the clusters onto the vector space to form reference vectors;
5 Project the test instances onto the vector space to form the test vectors;
6 Find the cluster that has the highest cosine similarity with a test vector;
7 Return the label of the selected cluster as the Hindi translation of the target Bengali word in the test context;

6.4 An example of vector space model based approach

We give a small example to show, how the use of vector space approach works. Let the target word be "{চাল}"(chaal). The word is contained in the synsets 6303 - ধানের বীজ থেকে প্রাপ্ত খাদ্য শস্য(dhaner bij theke prapto khadya shasya)[food grain obtained from paddy seeds] and 6132 - প্যাঁচ, কৌশল, চাল(pyanch,koushal,chaal)[tactics,maneuver,move] in the Bengali sense dictionary. When the Hindi wordnet is queried with these two synids it returns the synsets that contains the synsets containing the words चावल(chaval)[rice] and चाल(chaal)[tactics] respectively. Thus, these two words are considered as the possible translations of the Bengali word.

A mapping is available between the synset/concept identifiers in Hindi Wordnet and Bengali sense dictionary. This feature help us to automate the process of finding the set of all possible translations of a Bengali word in Hindi. Given a Bengali word we first searched the Bengali sense dictionary for synset ids of all the synsets which contain the word. We used these sense ids so obtained, to query the Hindi Wordnet for all the corresponding synsets. The set of all the unique Hindi words contained in Hindi synsets corresponding to the synset ids are considered to be the possible Hindi translations of the target Bengali word.

If the mapping between the two wordnets were not available, one would need to do alignment of the two synsets for finding the set of Hindi translations of the Bengali words and vice versa.

When the Hindi Corpus is searched with the two words we get the following sentences. भारत में इस साल चावल का उत्पादन कम है(Bharat mein is sal chaval ka utpadana kama hai)[The yield of rice is less in India this year], चावल से काफी सारे पकवान बनते है(Chavala se kafi sare pakavan bante hai) [A lot of dishes are prepared from rice] and मंत्री का चाल जनता समझ रहा है (Mantri ka chal janata samajh raha hai)[The public is able to look through the political maneuver of the

minister] and let the Bengali test sentence be এই বছর চাল উৎপাদন বড় ভালো হয়েছে(Ei bochor chaal utpadon boro valo hoyechhe).[The yield of rice is very good this year] Translation of the test instance to Hindi is इस साल चावल उत्पादन बड़ा अच्छा हुआ(Is sala chaval ka utpadana bada achha hua).

The Hindi vectors are as given in the Table 3.

The reference vectors 1 and 2 are combined to form the vector for चावल(chaval)[rice] and the vector 3 is the reference vector forचाल(chaal)[tactics] as shown in Table 4. The cosine similarity of the test vector turns out to be greater with reference vector for चावल(chaval)[rice]. Hence, the translation चावल(chaval)[rice] is predicted for চাল(chaal) in the given test instance.

6.5 Analysis of results

We analyzed the data and the results so obtained and found that the definite senses of most of the ambiguous words could be identified from the contextual information to a significantly high degree of accuracy. However, the accuracy of identification of the abstract senses of most of the words was quite low. For example, the word অর্থ (***artha***) in Bengali has the glosses **money** and **meaning**, and the word কেন্দ্র (***kendra***) has the glosses **center**,**regarding**, **central government** and **Place intended for some purpose (health centre or powerplant) etc.** It was observed that the sense **meaning** of the word অর্থ (artha) and the sense **regarding** of the word কেন্দ্র (kendra) could not be identified from the contextual information and the performance was poor for these words.But we found that these senses can be distingushed by looking at their context patterns. There exists definite patterns such as, certain patterns of verbs and inflectional forms of the words that occur in the neighborhood of the target word in a given contexts, can be used for identification of these senses for many of the abstract concepts.

The results of the first phase of the experiment (i.e., the prediction of translation from the clusters) gives mixed results. The poor performance of the system for some words is either due to insufficient training data for a word or because the word has some abstract senses that cannot be captured by the contextual nouns. In order to alleviate the problem due to abstract senses we introduce the rule-based post-processing step, which we discuss in section 6.6.

6.6 Rule-based correction

We identified some of these patterns from the data and defined some rules based on these patterns. These rules were encoded into a program which was executed as a post-processing step to reassign the Hindi translation to the instances of target Bengali word in a given set of test contexts. The postprocessing step was executed on the entire test data since we had to ensure that the corrective step assigns correct translation to the misclassified instances and leaves the correctly classified instances unchanged.

The target words for which rules were defined and the corresponding rules are given below.

অর্থ (artha)

Some of the most frequent words that co-occur with the word or bear similar meaning when used in the "money" sense are : টাকা, ব্যাঙ্ক, পরিমাণ, ঋণ, কমিটি, বাজেট, ক্ষেত্র, সময়, কেন্দ্র, মন্ত্রক, লগ্নি, বাণিজ্য etc.(taka, bank, pariman, rin, committee, budget, kshetra,samay,kendra,mantrak,lagni,banijya etc.)[money, bank, quantity, debt, budget, committee, area, time, center, ministry, investment, trade]

However, when the word is used to indicate the sense ***meaning*** it is difficult to identify the sense from contextual cues. We found that there is no specific set of words that could be used to identify this sense. This is due the wide variation and low frequency of the words that co-occur with the target word when used in the sense ***meaning*** in any context e.g.,

(1) উর্দুতে মালিকা নামের অর্থ রানি .(Urdute malika namer artha rani)[The Urdu word *malika* means queen]

(2) সার প্রয়োগ না করার অর্থ ফলন কমা .(sar proyog na korar artha folon koma)[Not using fertilizer implies reduction in yield]

However, we found that it follows a general pattern of "meaning of *something*" i.e., possessive sense. The word preceding the target word has a -র(-r) inflection attached to the root word and in some cases the qualifying words কোনো(kono)[any], কোনও(konoo)[any], গভীর(gaveer)[deep], নানান(nanan)[multiple], নানাবিধ(nanabidho)[multiple] preceding অর্থ(artha) qualifies it. In such cases, we have to look for possessive inflection in nouns preceding these qualifying words. Hence, we define the rules as follows.

(1) Check the word preceding the word অর্থ(artha), if it is contained in the set of qualifying words then look for the word preceding the qualifying word. Else take the word preceding the qualifying word.

(2) If the selected word has a -র(-r) inflection attached to the root word then tag the word অর্থ(artha) as মতলব(matlab)[meaning].

লক্ষ্য (lakshya)

We found that the distribution of a specific set of words in the neighborhood of the word লক্ষ্য(lakshya), when used in the the first two senses is rather consistent but when used in the sense “observation" it suffers from the same problem as that of the word অর্থ(artha). গণতন্ত্রের মাধ্যমে দেশের নিপীড়িত জনগণের শোষণ মুক্তি ঘটানোই পার্টির লক্ষ্য ।(Bahudaliyo ganotantrer madhyame desher nipirito janoganer shoshan mukti ghotanoi partir lakshya)[The main purpose of the party is the liberation of the oppressed people through democracy] (Purpose)

Accuracy of prediction of Hindi translation of Bengali words by unsupervised graph-based approaches			
Word	Navigli's approach(%)	Jurgens' approach(%)	GCE community detection algorithm(%)
আচার(achar)	46	71	86
অর্থ(artha)	54	35	48
চাল(chaal)	36	51	54
ডাল(daal)	82	66	87
গোলা(gola)	68	37.5	59
জাল(jaal)	96	47.3	54
কেন্দ্র(kendra)	22	22	47.8
লক্ষ্য(lakshya)	16	19.6	41
প্রণালী(pranaali)	75	30	71
রাস্তা(raasta)	43	44	51.5
Average Accuracy	**53.8**	**42.34**	**60**

Table 2: Results of prediction of Hindi translation of Bengali words by unsupervised graph-based approaches

Tokens	Bharat	mein	yaha	sal	chaval	ka	utpadana	kama	hai	se	kafi	sara	pakavan	bana	mantri	chal	janata	samajh	raha
Ref 1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Ref 2	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0
Ref 3	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	1	1	1
Test Bilingual Dict	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Table 3: Sentences projected onto vector space

দুষ্কৃতীদের লক্ষ্য বাপ্পা এক সময় প্রোমোটারি করতেন ।(Dushkritider lakshya Bappa ek samay promotari korten)[Bappa, who has been the target of the anti-socials, was involved in real estate business] (Target)
আতাপুর গ্রামের বাসিন্দা কমল সর্দার প্রথম তালতলা ঘাটের কাছে নদীবাঁধে ফাটল লক্ষ্য করেন ।(Aatapur gramer basinda Kamal Sardar prothom taltola ghater kache nodibandhe fatol lakshya koren)[Kamal Sardar, a resident of Atapur village, first say a breach in the dam near the river bank at Taltala] (Observe)

The patterns identified in this case are as follows; If the word succeeding লক্ষ্য(lakshya) has the root রাখ্(rakh) with any inflection other than -এ(-e) then it implies the sense "observation".

Similarly, if the word following লক্ষ্য(lakshya) has the root কর(kar), then it may imply either "aim" or "observe" depending on usage. The form 'কর + -e' has two orthographic forms;

1. non-finite form.
2. finite form (present, 3rd person).

The non-finite form following "লক্ষ্য(lakshya)" corresponds to the meaning "aim". Hence, to capture the "observe" sense we define the rule as;

- If the root of the word succeeding the word লক্ষ্য(lakshya) has the root কর(kar) and the inflection is other than "-e", then translate লক্ষ্য(lakshya) to "observe".
- If the inflection is "-e" an the part-of-speech is finite verb, then mark the sense as লক্ষ্য(lakshya) as "observe".

We define the rules as; Get the word immediately after the target word. Get its inflectional and root form. If the words follow the patterns described then translate the word to Hindi word देख(dekh).

চাল (chaal)

We observed that in the corpus the "Rice" sense is the dominant sense for the word চাল(chaal). Some of the most frequent words that co-occur with the word or bear similar meaning when used in the ***rice*** sense are : রেশন, জেলা, খাদ্য, দর, বরাদ্দ, বিক্রি, কুইন্টাল, দফতর, টন, দাম, সংগ্রহ, কিলো, কিলোগ্রাম etc.(ration, jela, khadya, dor, boraddo, bikri, quintal, doftor, ton, daam, songroho, kilo, kilogram etc.)[ration, district, food, price, allotment, sell, quintal, office, ton, price, collection, kilo, kilogram] We observed that;
(1) distinct clusters were formed for the sense ***Rice*** and ***Roof*** but we didn't find any cluster that corresponds to the ***maneuver*** of the target word.
(2) although two distinct clusters were formed that contains the indicator words.
for the ***roof*** sense of the target word, no suitable Hindi word was suggested for these clusters.

To address the first problem we identified some patterns in the data for the ***maneuver*** sense of the word চাল(chaal). The identified patterns are as follows;
1. The root forms of the words preceding চাল(chaal) contains the words পাল্টা, মোক্ষম, রাজনীতি, আইন(palta, moksham, rajniti, ain) [return, cogent, politics, law]
2. The root of the word immediately af-

Tokens	Bharat	mein	yaha	sal	chaval	ka	utpadana	kama	hai	se	kafi	sara	pakavan	bana	mantri	chal	janata	samajh	raha
r_1	1	1	1	1	2	1	1	1	2	1	1	1	1	1	0	0	0	0	0
r_2	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	1	1	1
Test Bilingual Dict(t_{bd})	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

r_1 reference vector corresponding to चावल r_2 reference vector corresponding to चाल

Table 4: Final reference and test vectors. r_1 and r_2 are the reference vectors

ter চাল(chaal) contains the words চাল, হিসাব, খেল(chaal,hisab,khel)[move,calculation,play].
3. The words দাবা, জুয়া(daba, juya)[chess, gamble] occurs within a a window of 4 words preceding the target word.

For a given test instance, if atleast one of the conditions is satisfied, then translate the target word to চাল(chaal).

The second problem is due to the absence of any concept(synid) corresponding to the roof sense of the word চাল(chaal) in Hindi wordNet. We observed that some of the clusters in Bengali contains words that clearly indicates the ***roof*** sense of the word. Manually, labelling these clusters with the word ছত(chat)[roof] shows some further improvement in performance.

কেন্দ্র (kendra)

Among the senses listed in Table 1 the sense ***issue*** cannot be clearly identified from the contextual information. However, it was found that the word করে(kore) immediately follows the word কেন্দ্র(kendra) when used in the sense ***issue***.
As a post-processing step, we checked the word immediately following the word কেন্দ্র(kendra) in the context. If the word is করে(kore) we identified its sense as that of ***issue***. We got a substantial improvement in performance after incorporating this corrective step.

6.6.1 Results of translation prediction after rule-based correction

The results of the translation prediction after rule-based correction are summarized in Table 5.

7 Experiment and Result

7.1 Context Refinement

FIRE 2011 [http://www.isical.ac.in/ fire] Bengali and Hindi News corpora were used for the experiments. For construction of the co-occurrence graph the words were converted to their root forms and only the nouns were retained for further processing. We wanted to test our system on a small amount of data. We used the most frequent nouns and the number of the nouns ranges between 150 to 300, depending the volume of the data obtained from the corpus for a target query word. The words in the Hindi corpus were also converted to their root forms and were used in construction of the vector space and the reference vectors for the translation.

7.2 Parameter Selection

In this section, we discuss the various parameters used for the target words in the algorithms.

Parameters for Navigli and Crisafulli's approach: For some of the words for which the cooccurrence graph is relatively sparse, a threshold of 0.00033 was used. However, for other words the cutoff Dice coefficient value ranges from 0.03 to 0.05.

The triangle/square score threshold values ranges between 0.24 to 0.68. We performed the experiments over a range of values and both *triangle* and *square* scores and reported the best results.

Paramerters for our community detection approach: We experimented with the minimum clique size k equal to 3 and 4, since any three nouns that occur in a sentence shall form a three clique. During expansion of a clique, a new node is added to the community, one at a time.

In this algorithm, we can control the degree of overlap among the initial(seed) cliques and final communities. In our experiment we considered the clique overlap values in the range of 0.4 to 0.7, and the community overlap values between 0.4 to 0.8.

We got best results for minimum clique size of 3, initial clique overlap degree range 0.4 to 0.6, community overlap degree values in range 0.6 to 0.7. However, we observed that the value of the parameter α ranges widely from 1.0 to 4.5 depending on the edge density of the graph which in turn is proportional to the number of sentences retrieved for a given target word.

The clusters obtained by this method, are tagged with Hindi translations of the target word and finally the tagged clusters are used to predict the Hindi translation of a target Bengali word in a given test context.

7.3 Result

Due to resource constraint we evaluated our system on 10 Bengali words. For each target word we considered 100 to 120 sentences for evaluation. In phase 1, our community detection method, the works of (Navigli and Crisafulli, 2010) and (Jurgens, 2011) was used to generate the sense clusters. The vector space approach was used in phase 2 for translation prediction for all the three cases of phase 1. In Table 1 we give the list of glosses in which the Bengali polysemous words on which we have tested our system. The results are given in Table 2. In Table 5, we report the improvement in results for some of the words after rule-based

Results of prediction of Hindi translation						
	Accuracy before correction			Accuracy after correction		
Word	Navigli's Approach(%)	Jurgens' Approach(%)	GCE community Detection approach(%)	Navigli's Approach(%)	Jurgens' Approach(%)	GCE community Detection approach(%)
অর্থ	**54**	**35**	**48**	**83**	**70.83**	**83.3**
চাল	**36**	**51**	**54**	**47**	**63**	**70.85**
কেন্দ্র	**22**	**22**	**47.8**	**50.7**	**55.6**	**70**
লক্ষ্য	**16**	**19.6**	**41**	**39**	**47.5**	**67**
Average Accuracy	**32**	**31.9**	**47.7**	**54.93**	**59.23**	**72.79**

Table 5: Results graph-based approaches for prediction of Hindi translation of Bengali words before and after rule-based correction. The accuracy for words অর্থ, চাল, লক্ষ্য and কেন্দ্র improved by rules

correction.

8 Conclusion

We have developed an unsupervised WSD system for Bengali which may be used to improve Bengali-Hindi machine translation. The results of the first phase(graph-based approach followed by vector space model approach) shows that the system performs resonably well even when a comparable corpus is used instead of a parallel corpus. However, for certain categories of words/senses, the performance is poor. In the second phase (Rule-based correction phase) we have shown that for some abstract senses of certain words for which the performance of the first phase was not good enough, the disambiguation can be done by identifying certain patterns in the inflectional forms of the the word itelf or the words in the neighbourhood of the target word. This is an initial attempt to test the effect of manually defined rules on WSD. We need to work further on the generalization of rules for particular classes of words and the effect of the generalization. Future work may study the effect on MT system when this module is integrated into some MT system.

References

Eneko Agirre and Oier Lopez de Lacalle. Ubc-alm: Combining k-nn with svd for wsd.

Eneko Agirre, David Martínez, Oier López de Lacalle, and Aitor Soroa. 2006. Two graph-based algorithms for state-of-the-art wsd. In EMNLP '06, pages 585–593, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marianna Apidianaki, Guillaume Wisniewski, Artem Sokolov, Aurélien Max, and François Yvon. 2012. Wsd for n-best reranking and local language modeling in smt. In SSST-6 '12, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual wsd and lexical selection in translation. In *EACL*, pages 77–85.

A. R. Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. 2011. Harnessing wordnet senses for supervised sentiment classification. In EMNLP '11, pages 1081–1091, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In ACL '91, pages 264–270.

Debasri Chakrabarti,Dipak Kumar Narayan, Prabhakar Pandey, Pushpak Bhattacharyya. 2002. Experiences in Building the Indo WordNet: AWordNet for Hindi. In *Proceedings of theFirstGlobal WordNet Conference, 2002*.

Boxing Chen, George F. Foster, and Roland Kuhn. 2010. Bilingual sense similarity for statistical machine translation. In *ACL*, pages 834–843.

David Jurgens. 2011. Word sense induction by community detection. In TextGraphs-6, pages 24–28, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mitesh M. Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2009. Projecting parameters for multilingual word sense disambiguation. In EMNLP '09, pages 459–467, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mitesh M. Khapra, Salil Joshi, and Pushpak Bhattacharyya. 2011a. It takes two to tango: A bilingual unsupervised approach for estimating sense distributions using expectation maximization. In *IJCNLP*, pages 695–704.

Mitesh M. Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011b. Together we can: bilingual bootstrapping for wsd. In ACL HLT '11, pages 561–569, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yoong Keok Lee and Hwee Tou Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation.

C. Lee, F. Reid, A. McDaid, and N. Hurley. 2010. Detecting highly-overlapping community structure by

greedy clique expansion. In *Workshop - ACM KDD-SNA*.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In COLING '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Piyush Kedia Mitesh M. Khapra, Sapan Shah and Pushpak Bhattacharyya. 2009. Projecting parameters for multilingual word sense disambiguation. In EMNLP '09.

Piyush Kedia, Mitesh M. Khapra, Sapan Shah and Pushpak Bhattacharyya. 2010. Domain-specific word sense disambiguation combining corpus based and wordnet based parameters. *5th International Conference on Global Wordnet (GWC 2010),Mumbai*.

Saurabh Sohoney, Mitesh M. Khapra, Anup Kulkarni and Pushpak Bhattacharyya. July 2010. All words domain adapted wsd: Finding a middle ground between supervision and unsupervision. In ACL '10, Uppsala, Sweden.

Sashank Chauhan, Soumya Nair, Mitesh M. Khapra, Pushpak Bhattacharyya and Aditya Sharma. 2008. Domain specific iterative word sense disambiguation in a multilingual setting. In ICON '08.

Roberto Navigli and Giuseppe Crisafulli. 2010. Inducing word senses to improve web search result clustering. In EMNLP '10, pages 116–126, Cambridge, MA, October. Association for Computational Linguistics.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In ACL '96, pages 40–47.

Ted Pedersen and Rebecca Bruce. 1997. Distinguishing word senses in untagged text. In EMNLP '97, pages 197–207.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123, March.

Jean Véronis. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

Špela Vintar, Darja Fišer, and Aljoša Vrščaj. 2012. Were the clocks striking or surprising?: using wsd to improve mt performance. In the Joint Workshop on ESIRMT andHyTra, EACL 2012, pages 87–92, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ellen M. Voorhees. 1993. Using wordnet to disambiguate word senses for text retrieval. In SIGIR '93, pages 171–180, New York, NY, USA. ACM.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *19th International Conference on Computational Linguistics*, pages 1093–1099.