Classification of Verbs - Towards Developing a Bengali Verb Subcategorization Lexicon

Somnath Banerjee **Dipankar Das** **Sivaji Bandyopadhyay**
Department of Computer Science & Engineering
Jadavpur University, Kolkata-700032, India
s.banerjee1980@gmail.com, dipankar.dipnil2005@gmail.com,
sivaji_cse_ju@yahoo.com

Abstract

The acquisition of subcategorization frames for verbs for any language is generally carried out either manually or automatically. The subcategorization lexicon is an important resource for any language and more so for languages such as Bengali that is morphologically rich, free phrase order and above all has no existing full-fledged parser. This paper presents the classification of Bengali verbs and their synonyms with different subcategorization frames according to their sense-based similarities. Syntax plays the main role in the acquisition of Bengali verb subcategorization frames. The main hypothesis on which the work is based is that the subcategorization frames for a Bengali verb are generally same with the subcategorization frames for its equivalent English verb with an identical sense tag. This hypothesis is reexamined to acquire the newly found subcategorization frames for the synonymous Bengali verbs to create verb classes. The classification of Bengali verbs according to their senses is carried out in a recursive way to create as many possible classes covered by the synonymous verbs that share the same types of subcategorization frames. The verb subcategorization frame acquisition system has demonstrated precision, recall and F-measure values of 74.11%, 70.83% and 72.44% respectively on a test set of 120 sentences.

1 Introduction

Several large, manually developed subcategorization lexicons are available for English, e.g. the COMLEX (Macleod *et al.*, 1994), ACQUILEX (Copestake, 1992) and the ANLT (Briscoe *et al.*, 1987) dictionaries. VerbNet (VN) (Kipper-Schuler, 2005) is the largest online verb lexicon with explicitly stated syntactic and semantic information based on Levin's verb classification (Levin, 1993). On the other hand, the lexicographic research project FrameNet (Baker *et al.*, 1998) and PropBank (Kingsbury and Palmer, 2002) have also generated important resources. But, there is no existing subcategorization lexicon available for the less privileged and less computerized Indian languages, e.g., Bengali.

The diverse characteristics of different Indian languages make the subcategorization frame acquisition task more challenging. Apart from other parts of speeches, the subcategorization information of verbs is an essential issue in parsing for the free phrase order languages such as Bengali. As there is no such existing parser available in Bengali, the acquisition as well as the evaluation of the acquired subcategorization frames is a difficult task. The main difference between English and Bengali sentences is the variation regarding the order of phrases.

(Das *et al.*, 2009) identified the subcategorization frames for the Bengali verb *dekha* (see) and (Banerjee *et al.*, 2009) dealt on the subcategorization frame acquisition task for ten different compound verbs that contain *kara* (do) as a component. The pivotal hypothesis in these two works is that the subcategorization frames obtained for a Bengali verb are generally same with the subcategorization frames that may be acquired for its equivalent verb with an identical sense tag in English. In this present task, these above-mentioned eleven Bengali verbs have been considered as the basic set of key verbs to construct different hierarchical sense based classes. The additional hypothesis in this work is that the synonyms of these key verbs containing the same sense share the same subcategorization frames and are classified into the same class with their key verb but the verbs containing multiple senses occupy multiple classes.

The number of different sense based English equivalent synonymous verb groups of a Bengali key verb is extracted from the Bengali to English bilingual dictionary[1] and are termed as Key Synonymous Verb Sets (KSVS). To accomplish the objectives, each class containing the key verb has been formed primarily based on the KSVSs. For each of the eleven key verbs, the synonyms that have been extracted from the Bengali-to-Bengali synonyms thesaurus (Mukhopadhyay, 2007) are termed as the member verbs for that corresponding key verb. Each Bengali synonym is searched in the Bengali to English bilingual dictionary to extract their English equivalent synonyms and the synonyms belonging to the same sense are termed as Member Synonymous Verb Set (MSVS). We have mapped the elements of each MSVS of a Bengali member verb to the elements of each KSVS of its corresponding key Bengali verb. If there is at least one element that belongs to both MSVS and KSVS, then that concerned member verb is included in the key verb class formed for that corresponding KSVS of the key verb.

Each member verb of a key verb class is passed through the subcategorization frame acquisition process (Das and Bandyopadhyay, 2009) carried out on the Bengali news corpus (Ekbal and Bandyopadhyay, 2008) with the help of English VerbNet. The newly acquired subcategorization frames have been evaluated manually with already existing subcategorization frames for the key verb class. This experiment is carried out to identify any valid subcategorization frames not available in the existing frames in that class. It has been observed that all the verbs in a key verb class share the same subcategorization frames as the members of their English equivalent Synonymous Verb Set (SVS) occupy in English VerbNet class.

As there is no WordNet (Miller, 1990) available in Bengali, the main problem here is to acquire the verb synonyms containing same and different senses to build a verb subcategorization lexicon. But, the synonymous member verbs containing same sense identified during the classification process can contribute to make the synsets for Bengali WordNet.

The rest of the paper is organized as follows. Section 2 gives the description of the related work carried out in this area. The classification strategy for the member verbs corresponding to each key verb is specified in Section 3. Section 4 describes the framework for the acquisition of subcategorization frames for the member verbs. Evaluation results of the system are mentioned in Section 5. Finally Section 6 concludes the paper.

2 Related Work

The early works for identifying verbs that resulted in extremely low yields for subcategorization frame acquisition are described in (Brent, 1991). Automatic acquisition strategies of verb subcategorization frames and their frequencies from large corpora are mentioned in ((Ushioda *et al.*, 1993) and (Manning, 1993)). An open class vocabulary of 35,000 words was analyzed manually in (Briscoe and Carroll, 1997) for subcategorization frames and predicate associations. The result was compared against associations in ANLT and COMLEX. Variations of subcategorization frequencies across corpus type (written vs. spoken) have been studied in (Carroll and Rooth, 1998). A mechanism for resolving verb class ambiguities using subcategorization frames is reported in (Lapata and Brew, 1999). All these works deal with English. Several works on the term classification of verb diathesis roles or the lexical semantics of predicates in natural language have been reported in (Korhonen, 2002).

The work on subcategorization frame acquisition of Japanese verbs using breadth-first algorithm is described in (Muraki *et al.*, 1997). Cross lingual work on learning verb-argument structure for Czech language is described in (Sarkar and Zeman, 2000). (Samantaray, 2007) gives a method of acquiring different subcategorization frames for the purpose of machine aided translation system for Indian languages.

Most of the above works have been done manually. The related cross lingual works have been evaluated with the help of parsers. In this present task, the non-availability of Bengali parser makes the evaluation strategy difficult to design. Though Bengali is a free phrase order language, the chunk level similarity with English phrases helps to construct the basic subcategorization frames for Bengali verbs. The contribution of the present work is that the sense-based classification of Bengali verbs shares same type of subcategorization frames.

[1] http://home.uchicago.edu/~cbs2/banglainstruction.html

3 Classification Strategy for Member Verbs

The classification strategy of member verbs sharing the same subcategorization frames as their key verb is the first step to create the verb classes. Equivalent English verbs of a key verb for different senses are identified from the Bengali to English bilingual dictionary. The sense wise separated English synonymous elements for each key verb constitute different Key Synonymous Verb Sets (KSVS).

The synonyms of the Bengali key verbs have been extracted from the Bengali-to-Bengali synonyms thesaurus. The thesaurus entries for the synonyms of the key verbs দেখা (*dekha*) [see], তৈরি করা (*toiri kara*) [make] and ব্যবহার করা (*bybohar kara*) [use / behave] are shown as follows where "(ক)" indicates the component part "করা" (*kara*) [do]. These entries have been retrieved from the thesaurus to create the set of synonymous member verbs for the Bengali key verb.

< দেখা > তাকানো , চাওয়া, দর্শণ (ক), দৃষ্টিপাত (ক), লক্ষ (ক) ।

< তৈরি (ক) > নির্মাণ (ক), গঠন (ক), প্রস্তুত (ক), সৃষ্টি (ক), স্থাপন (ক), উতপাদন (ক) ।
< ব্যবহার (ক)> প্রয়োগ (ক), আচরণ (ক) ।

The frequencies of the member verbs collected from the Bengali news corpus (Ekbal and Bandyopadhyay, 2008) are listed in Table 1. It has been observed that the key verbs like ভুল করা (*bhul kara*) [mistake], বন্ধ করা (*bondho kara*) [stop] and পর্যবেক্ষণ করা (*porjobekkhon kara*) [observe] have no direct entries for verb synonyms present in the thesaurus. Constructions of verb classes for these key verbs have not been attempted in this present task.

Each Bengali synonym is searched in the Bengali to English bilingual dictionary to extract their English equivalent synonyms of same and different senses. The elements of MSVS, the English equivalent synonyms carrying same sense are checked for mapping to the elements of each KSVS of its corresponding key verb. If there is at least one element belonging to MSVS and KSVS, the Bengali member verb is then placed in the same class with its key verb.

The classification process is recursive in nature. Each phase for classifying a member verb *Xbm* that belongs to Bengali synonyms set *Cbs* of the key verb *Ybk* is described below.

<**Key Verb**>: [<Synonym1 {Freq1}>, <Synonym2 {Freq2}>……]
< **দেখা (*dekha*)** [see]>: [<তাকানো (*takano*) {2}>, < চাওয়া (*chaoa*) {2}>, < দর্শণ করা (*darshan kara*) {9}>, < লক্ষ করা (*lakhya kara*) {1}>, < দৃষ্টিপাত করা (*dristipat kara*) {0}]
< **তৈরি করা (*tairi kara*)** [make] > : [<নির্মান করা (*nirman kara*) {0}>, < প্রস্তুত করা (*prostut kara*){11}> , < সৃষ্টি করা (*sristi kara*) {11}>,< উতপাদন করা (*utpadon kara*) {0} [produce] >, < স্হাপন করা (*sthapon kara*) {6}>,< গঠন করা (*gathan kara*){10}]
< **ব্যবহার করা (*babohar kara*)** [use/behave] >: [< প্রয়োগ করা (*proyog kara*){7} [apply]>, < আচরণ করা (*achoron kara*) {2} [behave]>]
< **বাস করা (*bas kara*)** [live] >: [< বসবাস করা (*basobas kara*) {5}>, < অধিষ্ঠান করা (*adhisthan kara*) {0}>, < অবস্হান করা (*abasthan kara*) {0} >]
< **কাজ করা (*kaj kara*)** [work] >: [< কর্ম করা (*karma kara*) {0} >, < কাজকর্ম করা (*kajkarma kara*) {0} >, < কার্য করা (*karya kara*) {0}>, < পরিশ্রম করা (*porishrom kara*) {2}>, <কাজকাম করা (*kajkam kara*) {0} >, <কাম করা (*kam kara*) {0} >, < কম্ম করা (*kamma kara*) {0}]
< **সংগ্রহ করা (*sangroho kara*)** [collect]>: [< যোগাড় করা (*jogar kara*) {5}>, < আদায় করা (*aday kara*) {4} >, < উশুল করা (*ushul kara*) {0}>]
< **চিৎকার করা (*chitkar kara*)** [shout] >: [< চ্যাঁচামিচি করা (*chanchamechi kara*) {0}>, < চেল্লাচেল্লি করা (*chellachelli kara*) {0}>, < গোলমাল করা (*golmal kara*) {0}>, <গন্ডগোল করা (*gandogol kara*) {0} >, < কোলাহল করা (*kolahal kara*) {0} >, < শোরগোল করা (*Shorgol kara*) {0} >, < হট্টগোল করা (*hattogol kara*) {0}>, < হৈচৈ করা (*haichai kara*) {1}>]
< **জিজ্ঞাসা করা (*jigyasa kara*)** [ask/enquire] >: [< প্রশ্ন করা (*prosno kara*) {6} [ask]>, < জিজ্ঞাসাবাদ করা {1} (*jigyasabad kara*) >, < জেরা করা (*jera kara*) {0} [enquiry]>, < সওয়াল করা (*sawal kara*) {0} >, < জিজ্ঞেস করা (*jigyes kara*) {4}>, < জিগেস করা (*jiges kara*){0}>, < জিগ্যেস করা (*jigyes kara*) {0}>]

Table 1: Frequencies of the member verbs for eight key verbs acquired from the Bengali news corpus

The class corresponding to *Ybk* is *Cbk*. The English equivalent classes *ECk* for the key verb *Ybk* and *ECm* for the member verb *Xbm* are defined as,

ECk = {KSVS1, KSVS2, ….., KSVSq}
ECm = {MSVS1, MSVS2, ….., MSVSp}

If $\exists$ *Xbm* | (*Xbm* € *Cbs*) for i = 1 to p, j = 1 to q (Zsi $\cap$ Zdj) $\neq \phi$
then *Xbm* € *Cbk*
where Zsi € MSVSi and Zdj € KSVSj.

The possible example entries present in the Bengali to English bilingual dictionary for synonymous member verbs প্রয়োগ করা (*prayog kara*) and আচরণ করা (*achoran kara*) and their corresponding key verb ব্যবহার করা (*byabahar kara*) are given as follows.

# *Member Verbs*:

< প্রয়োগ করা v. to employ; ***to apply, to use*** ;>

< আচরণ করা v. ***to behave***; to deal (with), to act (towards); to practice ;>

# *Key Verb:*

< ব্যবহার করা v. to apply, to use; to behave, to treat (a person), to behave towards; …>

In the dictionary, different synonyms for a verb with the same sense are separated using "," and different senses are separated using ";". The synonyms for the different senses of the verbs have been extracted from the dictionary. This yields a resulting set called Synonymous Verb Set (SVS). For example, the English synonyms (*apply*, *use*) and synonym with another sense (*behave*) have been retrieved for the Bengali key verb "ব্যবহার করা " (*byabahar kara*) and have been categorized as two different KSVS for the Bengali key verb. Each separate class has been formed for each KSVS of the key verb. On the other hand, the English equivalents *apply, use* of the Bengali member verb প্রয়োগ করা (*payog kara*) and *behave* of the Bengali member verb আচরণ করা (*achoran kara*) are different MSVSs for these two member verbs. These two MSVSs consist of two different senses belong to two key classes formed by their corresponding key verb. The English synonym *employ* constitutes another MSVS of the Bengali member verb প্রয়োগ করা (*payog kara*). But this MSVS does not belong to any existing classes formed by its key verb, as it is not satisfied by the classification criterion. Table 2 shows the statistics of the number of KSVSs of the eight (8) key verbs and number of MSVSs of eighteen (18) different member verbs acquired from the corpus after the first phase of the recursive process.

<Key Verb> : { < Member Verb [Number of MSVSs] >}	**Number of Verb Classes/ KSVSs for the Key Verb**
<**দেখা** (***dekha***) [see]>: {<তাকানো (*takano*) [1]>, < চাওয়া (*chaoa*) [1]>, < দর্শণ করা (*darshan kara*) [2]>, < লক্ষ করা (*lakhya kara*) [2]>}	3
<**তৈরি করা (*tairi kara*)** [make] > : {< প্রস্তুতকরা (*prostut kara*) [1]> , < সৃষ্টি করা (*sristi kara*) [2]>,< স্হাপন করা (*sthapon kara*) [2]>,< গঠন করা (*gathan kara*) [2]>}	3
<**ব্যবহার করা (*babohar kara*)** [use/behave] >: {< প্রয়োগ করা (*proyog kara*) [2] >, < আচরণ করা (*achoron kara*) [2]>}	2
<**বাস করা (*bas kara*)** [live] >: {< বসবাস করা (*basobas kara*) [1]>}	1
<**কাজ করা (*kaj kara*)** [work] >: {< পরিশ্রম করা (*porishrom kara*) [1]>}	1
<**সংগ্রহ করা (*sangroho kara*)** [collect]>: {< যোগাড় করা (*jogar kara*) [1]>, < আদায় করা (*aday kara*) [2] >}	2
<**চিৎকার করা (*chitkar kara*)** [shout] >: {< হৈচৈ করা (*haichai kara*) [1]>}	1
<**জিজ্ঞাসা করা (*jigyasa kara*)** [ask/enquire] >: {< প্রশ্ন করা (*prosno kara*) [1] >, < জিজ্ঞাসাবাদ করা (*jigyasabad kara* [2]) >, < জিজ্ঞেস করা (*jigyes kara*) [1]>}	2

Table 2: Number of KSVSs and MSVSs of the member verbs after first phase of the recursive process.

If any MSVS of a member verb remains unclassified, then that member verb is passed through the present classification process considering it as a key verb. But, the process will not be repeated for the synonym entry present in the Bengali synonyms thesaurus if that synonym entry has already been attempted in the classification process as a key verb. The recursive process terminates when no MSVS of a member verb is left unclassified. It has been observed that the preliminary separation of the Bengali member verbs into different verb classes follows the same classification as their English equivalent verbs present in English VerbNet.

4 Subcategorization Frames Acquisition FrameWork

The subcategorization frames acquisition task for the ten key verbs has been reported in the previous work (Banerjee *et al*., 2009). The subcategorization frames acquisition task is conducted separately for each member verb of a class except the key verb as the acquisition has already been done for the key verbs. This task has been carried out for two reasons. The first reason is to extract any frame that may exist in the corpus and verify the membership of this frame to its corresponding class. The second motive is to classify the newly acquired subcategorization frames into the existing key classes according to the closeness related to their frame sharing properties.

We have developed several modules for the acquisition of verb subcategorization frames for the member verbs from the Bengali newspaper corpus. The modules consist of POS tagging and chunking, identification and selection of verbs, English verb determination, frame acquisition from VerbNet and mapping of the acquired Bengali verb subcategorization frames to their English equivalent VerbNet frames.

We have used a Bengali news corpus (Ekbal and Bandyopadhyay, 2008) developed from the web-archives of a widely read Bengali newspaper. A portion of the Bengali news corpus containing 14000 sentences have been POS tagged using a Maximum Entropy based POS tagger (Ekbal *et al.*, 2008). The POS tagger is developed with a tagset of 26 POS tags[2], defined for the Indian languages. The POS tagger demonstrated an accuracy of 88.2%. We have developed a rule-based chunker to chunk the POS tagged data with an overall accuracy of 89.4%.

To identify the member verbs from the tagged and chunked corpus, the data are analyzed to identify the words that are tagged as main verb (VM) and belong to the verb group chunk (VG) in the corpus. For the compound member verbs containing "করা" (*kara*) with pattern such as {[XXX] (NN) [*kara*] (VM)} have been identified and retrieved from the Bengali POS tagged and chunked corpus (e.g. [(*prayog*(NN) *kara*(VM))(apply)], [(*byabahar* (NN) *kara*(VM))(behave)] etc.).

The verb subcategorization frames for the equivalent English verbs (sharing the same sense) of a Bengali verb are the initial set of verb subcategorization frames that have been considered as valid for that Bengali verb. The different inflected forms in which the member verbs appear in the Bengali corpus have been identified accordingly. Different suffixes may be attached to a verb depending on the various features such as Tense, Aspect, and Person. A Bengali stemmer with an accuracy of 97.09% that uses a suffix list to identify the stem form of the member verbs has been developed. Another table stores the stem form and the corresponding root form.

The determination of equivalent English verbs has been carried out using a Bengali to English bilingual dictionary. The Bengali to English dictionary entry as mentioned in Section 3 for each verb has been analyzed to identify its synonyms and meanings to construct the SVS of that verb.

VerbNet associates the semantics of a verb with its syntactic frames and combines traditional lexical semantic information such as *thematic roles* and *semantic predicates*, with syntactic frames and *selectional restrictions*. Verb entries in the same VerbNet class share common syntactic frames, and thus they are believed to have the same syntactic behavior. The VerbNet files containing the verbs with their possible subcategorization frames and membership information are stored in XML file format. The XML files of VerbNet have been preprocessed to build up a general list that contains all verbs, their classes and possible subcategorization frames (primary as well as secondary). This preprocessed list is searched to acquire the subcategorization frames for each SVS of the Bengali verb.

The acquired VerbNet frames have been mapped to the Bengali verb subcategorization frames by considering the position of the verb as well as its general co-existing nature with other phrases in Bengali sentences (Das *et al.*, 2009).

```
   ম্যাক্স          যার           থেকে
(Max)NN (jar) PRP (theke) PSP
     হাতপাখা
   (NP(Hatpakha) NN  )
     প্রস্তুত               করেছিলেন
   (prostut)NN (korechilen) VM
```

For example, the syntax of "NP-PP" frame for a Bengali sentence has been acquired by identifying the target member verb followed by a NP chunk and a PSP chunk. The above sentence containing prepositional frame "PP" does not appear in the Bengali corpus, as there is no concept of preposition in Bengali. But, when we

[2] http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

compare these types of sentences containing postpositional markers, i.e. PSP (postpositions) as a probable argument of the verb, the system gives the desired output.

There are some frames that did not have any instance in our corpus. A close linguistic analysis shows that these frames can also be acquired from the Bengali sentence. It has been observed that sense wise separated SVS members consist of English equivalent synonyms of a Bengali verb occupy the membership of same class or subclass in VerbNet. The example Bengali verb class for the key verb “দেখা” (*dekha*) [see] is as follows,

```
   <?xml version="1.0" encoding="UTF-8"?>
   <VNCLASS ID= দেখা.xml
   <MEMBERS>
   <MEMBER name="দেখা">
   <MEMBER name="লক্ষ্যকরা  "
   <MEMBER name="দর্শণকরা  "
   <MEMBER name="চাওয়া"
   <MEMBER name="তাকানো  "
   </MEMBERS>
   <FRAMES>
   <FRAME name=Basic-Transitive</FRAME>
   <SYN>NP-NP-V</SYN>
   <Example>আমি কাকাতুয়া দেখি</Example>
   <FRAME name=S</FRAME>
   <SYN>NP-V-যে-(NP-NP-V) </SYN>
   <Example>আমি দেখলাম যে রাম ঐ কাজটি করছে
</Example>
   </FRAMES>……
```

5 Evaluation Results

The set of acquired subcategorization frames or the frame lexicon can be evaluated against a gold standard corpus obtained either through manual analysis or from subcategorization frame entries in a large dictionary or from the output of the parser made for that language.

As there is no parser available for the Bengali and no existing dictionary for Bengali containing subcategorization frames, manual analysis of the system output with the gold standard corpus data is the only method for evaluation. The gold standard data has been prepared manually from the chunked sentences that contain the member verbs. The verb subcategorization frames acquisition process is evaluated using type precision (*tp*) (the percentage of subcategorization frame types that the system proposes are correct according to the gold standard), type recall (*tr*) (the percentage of subcategorization frame types in the gold standard that the system proposes) and F-measure as

$$[2*(tp)*(tr)]/ [(tp) + (tr)].$$

The classification of acquired subcategorization frames for the eighteen member verbs have been carried out accordingly. But, our main objective is to explore the newly found subcategorization frames identified for the member verbs and their classification into the respective classes. Identification of such valid frames in case of Bengali and their presence in the appropriate classes have been conducted to improve the recall and precision values as well. The system has been evaluated with 120 gold standard test sentences containing the eighteen member verbs and the evaluation results are shown in Table 3.

Measures	**Results**
Recall	70.83%
Precision	74.11%
F-Measure	72.44%

Table 3. The Precision, Recall and F-Measure values of the system for the acquired eighteen (18) member verbs

A detailed statistics of the verbs is presented in Table 4. During the Bengali verb subcategorization frame acquisition process, it has been observed that simple sentences generally contain most of the frames as their corresponding English verb form usually takes in VerbNet. Analysis of a simple Bengali sentence to identify the verb subcategorization frames is easier in the absence of a parser than analyzing complex and compound sentences.

It has been noticed that the absence of other frames in the Bengali corpus is due to the free phrase ordering characteristics of Bengali Language. The proper alignment of the phrases is needed to cope up with this language specific problem. It can help to accelerate the task of disambiguating the arguments from the adjuncts with sufficient accuracy. The number of different frames acquired for these ten verbs is shown in Table 5. Two types of newly found valid frames have been extracted for the two Bengali verb classes. The ‘*’ in Table 5 indicates the new valid frame type as identified by the member verb. These frames have been included in their corresponding key classes and the verification is done manually.

Information	**Freq.**
Number of sentences in the corpus	14000
Number of key verbs considered in the present task	11
Number of key verb entries available in the Bengali synonyms thesaurus to construct main classes	8
Number of member verbs identified from the Bengali synonyms thesaurus entries	41
Number of member verbs appeared in the corpus with frequency >0	18
Number of sentences containing member verbs in the corpus	120
Number of KSVSs or Verb Classes of the key verbs after first phase of recursive classification task	15
Number of KSVSs or Verb Classes of the key verbs at the end the recursive classification task	22
Number of subcategorization frames acquired from the chunked gold standard 120 sentences	85
Number of subcategorization frames identified correctly from the acquired 85 sentences	63
Number of newly found subcategorization frames only for Bengali	2

Table 4. The frequency information of the verbs acquired from the corpus

Bengali <Key Verb Class>		
(Member Verb)	**Type of Subcategory Frames**	**No. of Frames**
< দেখা (***dekha***)>[see]		
(তাকানো)	Basic Transitive	(1)
(দর্শণ করা)	Basic Transitive	(2)
(লক্ষ করা)	S (Sentential Complement)	(1)
<তৈরি করা (*toiri kara*) >[make]		
(প্রস্তুত করা)	NP-PP NP-NP Basic Transitive*	(2) (7) (1)
(সৃষ্টি করা)	NP-PP	(9)
(স্হাপন করা)	NP-PP NP-NP Basic Transitive*	(2) (1) (1)
(গঠন করা)	NP-PP NP-NP	 (4) (5)
<ব্যবহার করা (*babohar kara*)>[use/behave]		
(প্রয়োগ করা)	NP-PP NP-NP	(1) (5)
(আচরণ করা)	NP-PP	(2)
<বাস করা (*bas kara*)>[live]		
(বসবাস করা)	Basic Transitive ADVP-PRED For-PP*	(2) (1) (1)
<কাজ করা (*kaj kara*)>[work]		
(পরিশ্রম করা)	NP-PP	(1)
< সংগ্রহ করা (*sangroho kara*)>[collect]		
(যোগাড় করা)	Transitive (Material object)	(1)
(আদায় করা)	PP	(2)
< চিৎকার করা (*chitkar kara*) >[shout]		
(হৈচৈ করা)	S (Sentential Complement)	(1)
<জিজ্ঞাসা করা (*jigyasa kara*)> [ask/enquire]		
(প্রশ্ন করা)	That-S	(3)
(জিজ্ঞাসাবাদ করা)	Basic Transitive	(1)
(জিজ্ঞেস করা)	S-SUBJUNCT Basic Transitive That-S	(4) (1) (3)

Table 5. The frequencies of different frames acquired from corpus

5 Conclusion

The acquisition of subcategorization frames for Bengali verbs and their clustering has helped to build a small verb lexicon for Bengali language. The language specific new frames have been identified in this present task. The sense-based separation of verbs according to syntactical resemblance requires an emphasis on the semantic roles for further exploring the classes towards generalization. For the free-phrase-order languages like Bengali, the error caused in improper argument-adjunct distinction can be reduced and successively the overall performance can be increased with the help of machine learning approaches. Verb morphological information, synonymous sets and their possible subcategorization frames are all important information to develop a full-fledged parser for Bengali. The system can be used for solving alignment problems in Machine Translation for Bengali as well as to identify possible argument selection for Question and Answering systems.

References

Anna Korhonen. 2002. Semantically motivated subcategorization acquisition. *ACL Workshop on Unsupervised Lexical Acquisition*. Philadelphia.

Anoop Sarkar and Daniel Zeman. 2000. Automatic extraction of subcategorization frames for czech. *COLING-2000.*

A. Ekbal and S. Bandyopadhyay. 2008. A Web-based Bengali News Corpus for Named Entity Recognition. *LRE Journal*. Springer.

A.Ekbal, R. Haque and S. Bandyopadhyay. 2008. Maximum Entropy Based Bengali Part of Speech Tagging. *RCS Journal*, (33): 67-78.

Akira Ushioda, David A. Evans, Ted Gibson, Alex Waibel. 1993. The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora. *Workshop on Acquisition of Lexical Knowledge from Text*, 95-106.

Ashoke Mukhopadhyay. 2007 ed. Samsad Samarthasabda Kosh. ISBN 81-85626-09-X

B. K. Boguraev and E. J. Briscoe.1987. Large lexicons for natural language processing utilising the grammar coding system of the Longman Dictionary of Contemporary English. *Computational Linguistics*, 13(4): 219-240.

Christopher D. Manning. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. *31st Meeting of the ACL*, 235-242. Columbus, Ohio.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe.1998. The Berkeley FrameNet project. *COLING/ACL-98*, 86-90. Montreal.

Copestake A.1992. The ACQUILEX LKB: Representation Issues in the Semi-automatic Acquisition of Large Lexicons. *ANLP*. Trento, Italy.

D.Das, A.Ekbal, and S.Bandyopadhyay. 2009. Acquiring Verb Subcategorization Frames in Bengali from Corpora. *ICCPOL-09,* LNAI-5459, 386-393.Hong Kong.

Grishman, R., Macleod, C., and Meyers, A. 1994. Comlex syntax : building a computational lexicon. *COLING-94*, 268-272. Kyoto, Japan.

George A. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235-312.

Glenn Carroll, Mats Rooth. 1998. Valence induction with a head-lexicalized PCFG. *EMNLP*. Granada.

Karin Kipper-Schuler.2005. VerbNet: *A broad-coverage, comprehensive verb lexicon.* Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA.

Kazunori Muraki, Shin'ichiro Kamei, Shinichi Doi.1997. *A Left-to-right Breadth-first Algorithm for. Subcategorization Frame Selection of Japanese Verbs.* TMI.

Levin, B. 1993. English Verb Classes and Alternation: A Preliminary Investigation. The University of Chicago Press.

Michael Brent.1991. Automatic acquisition of subcategorization frames from untagged text. *29th Meeting of the ACL*, 209-214. California.

Maria Lapata, Chris Brew.1999. Using subcategorization to resolve verb class ambiguity. *WVLC/EMNLP*, 266-274.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to PropBank. In *Proceedings of the 3rd InternationalConference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

S.Banerjee, D.Das and S.Bandyopadhyay. 2009. Bengali Verb Subcategorization Frame Acquisition - A Baseline Model. *(ACL-IJCNLP-2009), ALR-7 Workshop,* 76-83, Suntec, Singapore.

S.D. Samantaray.2007. A Data mining approach for resolving cases of Multiple Parsing in Machine Aided Translation of Indian Languages. *ITNG'07 © IEEE.*

Ted Briscoe, John Carroll.1997. Automatic Extraction of Subcategorization from Corpora. *AFNLP-ACL.*