Bengali Verb Subcategorization Frame Acquisition - A Baseline Model

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

Acquisition of verb subcategorization frames is important as verbs generally take different types of relevant arguments associated with each phrase in a sentence in comparison to other parts of speech categories. This paper presents the acquisition of different subcategorization frames for a Bengali verb *Kara* (*do*). It generates compound verbs in Bengali when combined with various noun phrases. The main hypothesis here is that the subcategorization frames for a Bengali verb are same with the subcategorization frames for its equivalent English verb with an identical sense tag. Syntax plays the main role in the acquisition of Bengali verb subcategorization frames. The output frames of the Bengali verbs have been compared with the frames of the equivalent English verbs identified using a Bengali-English bilingual lexicon. The flexible ordering of different phrases, additional attachment of optional phrases in Bengali sentences make this frames acquisition task challenging. This system has demonstrated precision and recall values of 77.11% and 88.23% respectively on a test set of 100 sentences.

1 Introduction

A subcategorization frame is a statement of what types of syntactic arguments a verb (or an adjective) takes, such as objects, infinitives, that-clauses, participial clauses, and subcategorized prepositional phrases (Manning,1993). The verb phrase in a sentence usually takes various types of subcategorization frames compared to phrases of other types and hence the acquisition of such frames for verbs are really challenging.

A subcategorization dictionary obtained automatically from corpora can be updated quickly and easily as different usages develop. Several large, manually developed subcategorization lexicons are available for English, e.g. the COMLEX Syntax (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). It is a hierarchical domain-independent, broad-coverage verb lexicon with mappings to other lexical resources such as WordNet (Miller, 1990), XTAG (XTAG Research Group, 2001) and FrameNet (Baker *et al.*, 1998). But, there is no existing subcategorization lexicon available for Bengali language. The subcategorization 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 evaluation of the acquired subcategorization frames are difficult but crucial tasks. The main difference between English and Bengali sentence is the variation in the ordering of various phrases. The pivotal hypothesis here is that the subcategorization frames obtained for a Bengali verb are same with the subcategorization frames that may be acquired for its equivalent verb with an identical sense tag in English.

The present work deals with the acquisition of verb subcategorization frames of a verb *kara* (do) from a Bengali newspaper corpus. This verb generates various types of compound verbs in combination with other preceding noun phrases in Bengali. The sentences containing these types of compound verb entries have been retrieved from the Bengali corpus. The Bengali verb subcategorization frame acquisition task has been carried out for the ten most frequent compound verbs that contain *kara* (do) as a component. The number of occurrences of other compound verbs

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 76–83,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

is negligible in the corpus. So, for evaluation purpose, we have not considered those verbs. Each of the ten Bengali compound verbs has been searched in the Bengali-English bilingual lexicon[1] and the equivalent English verb meanings with its synonyms have been identified and retrieved. All possible subcategorization frames for each of the English synonyms of the Bengali verb have been acquired from the English VerbNet[2]. These frames have been mapped to the Bengali sentences that contain the compound verb. Evaluation results with a test set of 100 sentences show the effectiveness of the model with precision, recall and F-Measure values of 77.11%, 88.23% and 79.24% respectively. There are some frames that have not been identified due to their absence in the corpus. Linguists have suggested that these frames do appear in Bengali and hence can be acquired.

The rest of the paper is organized as follows. Section 2 gives the description of the related works carried out in this area. Section 3 describes the framework for the acquisition of subcategorization frames for ten compound Bengali verbs. Evaluation results of the system are discussed in section 4. Finally section 5 concludes the paper.

2 Related Work

One of the early works for identifying verbs that resulted in extremely low yields for subcategorization frame acquisition is described in (Brent, 1991). A rule based system for automatically acquiring six verb subcategorization frames and their frequencies from a large corpus is mentioned in (Ushioda *et al.*, 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 ((McCarthy, 2001), (Korhonen, 2002), (Stevenson and Merlo, 1999) and (Walde, 1998)).

A 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. The work on subcategorization frame acquisition of Japanese verbs using breadth-first algorithm is described in (Muraki *et al.*, 1997).

3 System Outline

We have developed several modules for the acquisition of verb subcategorization frames from the Bengali newspaper corpus. The modules consist of POS tagging and chunking, Identification and Selection of Verbs, English Verb Determination, Frames Acquisition from VerbNet and Bengali Verb Subcategorization Frame Acquisition.

3.1 POS Tagging and Chunking

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 1500 sentences have been POS tagged using a Maximum Entropy based POS tagger (Ekbal *et al.*, 2008). The POS tagger was developed with a tagset of 26 POS tags[3], defined for the Indian languages. The POS tagger demonstrated an accuracy of 88.2%. We have also developed a rule-based chunker to chunk the POS tagged data with an overall accuracy of 89.4%.

3.2 Identification and Selection of Verbs

Our previous work (Das *et.al.*, 2009) on the acquisition of Bengali subcategorization frames from the same Bengali news corpus was carried out for the most frequent verb "দেখা" (*dekha*) (see) in that corpus. The next highest frequent verb in this corpus is "করা" (*kara*) (do) which is a special verb in Bengali. However to the best of our knowledge, no frame acquisition task has been carried out yet for this Bengali verb. The single occurrence of "করা" (*kara*) as a main verb in a sentence takes completely different subcategorization frames in comparison with the acquired frames for the compound verbs consisting of "করা" (*kara*) as a component. Hence, we have

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

[2] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

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

concentrated our focus to acquire subcategorization frames for the Bengali verb “করা” (*kara*).

For this purpose, we have manually analyzed the tagged and chunked data to identify the word “করা” (*kara*) that are tagged as main verb (VM) and belong to the verb group chunk (VG) in the corpus. The preceding noun phrase of “করা” (*kara*) generally produces completely different verbs in Bengali (e.g. [তৈরি করা (*tairi*(NN) *kara*(VM))(*make*)], [ব্যবহার করা (*byabahar* (NN) *kara*(VM))(*use*)] etc.).

Bengali, like any other Indian languages, is morphologically very rich. Different suffixes may be attached to a verb depending on the various features such as Tense, Aspect, and Person. The Bengali stemmer uses a suffix list to identify the stem form of the verb “করা” (*kara*). Another table stores the stem form and the corresponding root form. Stemming process has correctly identified 234 occurrences of the verb “করা” (*kara*) from its 241 occurrences in the corpus with an accuracy of 97.09%. The sentences where the verb “করা” (*kara*) appears in any inflected form but has been tagged as main verb (VM) have been retrieved. These sentences have been considered for fine-grained analysis of verb subcategorization frames. It is expected that the corpus will have adequate number of occurrences for each subcategorization frame of the verb. The passive occurrences of “করা” (*kara*) such as “করানো” (*karano*), করিয়ে (*kariye*) have been filtered out and the sentences containing the passive entries of “করা” have not been considered in the present work.

The compound verb phrases with pattern such as {[XXX] (NN) [*kara*] (VM)} have been identified and retrieved from the Bengali POS tagged and chunked corpus. It has been observed that most of these compound verb phrases are individually different verbs in Bengali. Around 182 various kinds of verbs have been identified. Certain typical and distinct occurrences of “করা” (*kara*) have also been identified. But, linguistic verification shows that these typical verbs are formed by attaching the verb “করা” (*kara*) to an adjective or an adverb word, like ঝকঝক করা (*jhakjhak kara*) , তকতক করা (*taktak kara*), শীত করা (*sheet kara*) etc. Such types of around 48 verb entries have been identified and filtered out from the corpus. The rest 134 distinct types of Bengali compound verbs (CV) with “করা” (*kara*) as a component have been considered as target verbs for analysis.

We have identified the frequencies of these verbs in the corpus. It has to be mentioned that only a few verbs have an adequate number of sentences in the corpus. For this reason, only the top ten compound verbs that have the largest number of occurrences in the corpus have been selected. Table 1 represents the top 10 different Bengali compound verbs and their frequencies obtained from the corpus.

Bengali Verbs	**Freq.**
তৈরি করা (*tairi kara*) (make)	23
ব্যবহার করা (*byabahar kara*) (use)	18
বাস করা (*bas kara*) (live)	17
কাজ করা (*kaj kara*) (work)	15
সংগ্রহ করা (*sangraha kara*) (collect)	13
বন্ধ করা (*bandha kara*) (shut)	13
চিত্‌কার করা (*chitkar kara*) (shout)	3
ভুল করা (*bhul kara*) (mistake)	3
জিজ্ঞাসা করা (*jigyasa kara*) (ask)	3
পর্যবেক্ষণ করা (*parjabekkhan kara*) (observe)	3

Table 1. Top 10 Bengali Compound Verbs and their frequencies obtained from the corpus

3.3 English Verb Determination

The verb subcategorization frames for the equivalent English verbs (in 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 root forms of the target verbs appearing in different inflected forms in the Bengali corpus have been identified by the process described in section 3.2. The determination of equivalent English verbs has been carried out using a Bengali-English bilingual lexicon. We have used the available Bengali-English bilingual dictionary that has been formatted for the text processing tasks. Various syntactical representations of a word entry in the lexicon have been analyzed to identify its synonyms and meanings. The example of an entry in the bilingual lexicon for our target verb “করা” (*kara*) is given as follows.

```
<করা [karā] v to do, to per-
form, to accomplish, to exe-
cute (কাজ করা); to build, to
make (তৈরি করা) ;.....>
```

But, the various distinct verbs, with “করা” (*kara*) as a component have individual separate entries in the bilingual dictionary. We have identified the equivalent English verbs from those Bengali verb entries in the dictionary. For example,

```
<তৈরি করা v. to build, to
make; …>
<ব্যবহার করা v. to apply, to
use; to behave; to treat (a
person), to behave towards;
…>
<কাজ করা v. to work; to
serve; to be effective ;…>
```

Different synonyms for a verb having the same sense are separated using "," and different senses are separated using ";" in the lexicon. The synonyms including different senses of the target verb have been extracted from the lexicon. 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 selected for Bengali verb "ব্যবহার করা" (*byabahar kara*) and have been categorized as two different SVS for the Bengali verb "ব্যবহার করা". Two synonyms (*make, build*) for the Bengali verb "তৈরি করা" (*tairi kara*) are thus present in the same SVS. Now, the task is to acquire all the possible existing frames for each member of the SVS from the VerbNet. The "করা" (*kara*) verb may also appear in passive form in Bengali sentences. For example,

```
 রামকে          কাজ
(Ramke)NNP   (kaj)NN
 করানো          হয়েছিল
(karano)VM   (hayechilo)VAUX
```

The corresponding dictionary entry for the passive form of "করা" (*kara*) is as follows. But in this work, we have concentrated only on those sentences where "করা" (*kara*) appears in active form.

```
<করানো [karānō] v to cause to
do or perform or accomplish
or execute or build or
make…>
```

3.4 Frames Acquisition from VerbNet

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 is stored in XML file format. The Bengali verb তৈরি করা (*tairi kora*) (make) has no direct class in VerbNet. The verb "make" and its synonymous verb "build" are members of one of the subclasses of the build-26.1 class and "make" is also a member of the dub-29.3 class. A snapshot of XML file for the build-26.1 class is given below.

```
.....
<VNCLASS ID="build-26.1"
.....<SUBCLASSES>
   <VNSUBCLASS ID="build-26.1-1">
<MEMBERS>
    <MEMBER name="build"
wn="build%2:36:00"/>
    <MEMBER name="make"
wn="make%2:36:01 make%2:36:05
.....
make%2:42:13 make%2:36:10"/>
.....
</MEMBERS>
.....
<FRAME>
    <DESCRIPTION descriptionNum-
ber="3.9" primary="NP-PP" secon-
dary="Asset-PP" xtag=""/>
<EXAMPLES>
    <EXAMPLE> The contractor
builds houses for $100,000.
    </EXAMPLE>
    .....
</EXAMPLES>
.....</FRAME>
.....
```

The verbs in VerbNet that take same type of subcategorization frames are stored in the <MEMBER> tag and the possible primary and secondary subcategorization frames are kept in <DESCRIPTION> tag with proper English examples for each frame. The example for each of the subcategorization frames for the English verb "make" has been given in the "build-26.1-1" subclass of the "build-26.1" class in the VerbNet. The sentence tagged within <EXAMPLE>.. </EXAMPLE> shows that after the occurrence of the verb "build/make", one noun phrase (NP) and one prepositional phrase (PP) have occurred as the arguments of the verb. The frame corresponding to this sentence has been described as the primary frame "NP-PP" in the frame description <DESCRIPTION> tag.

Sense wise separated SVS members occupy the membership of same class or subclass in VerbNet. It has been observed that the verbs "*build*" and "*make*" are members of the same SVS (extracted from the Bengali-English bilingual dictionary) and they are also members of the same subclass build-26.1-1. Therefore, both of the verbs take same subcategorization frames.

SVS (VerbNet classes)	**Primary and *Secondary* Frames for a SVS**
Make (build-26.1-1) Build (build-26.1-1)	NP-PP, NP, NP-NP, NP-NP-PP, *Asset-PP* *Asset-Subject*
Use (use-105, consume-66, fit-54.3) Apply (use-105)	NP-ADVP, NP-PP, NP-TO-INF-VC, Basic Transitive, NP-ING-SC, Location Subject Alternation, NP-PP *for-PP, Location-PP*
Behave (masquerade-29.6, 29.6-1)	PP, Basic Transitive *as-PP, like-PP, in-PP*

Table 2. The SVS members and their subcategorization frames for the corresponding Bengali verbs তৈরি করা (*tairi kara*) and ব্যবহার করা (*byabahar kara*)

The xml files of VerbNet have been preprocessed to build up a general list that contains all members (verbs) and their possible subcategorization frames (primary as well as secondary) information. This preprocessed list is searched to acquire the subcategorization frames for each member of the SVS of the ten Bengali verbs (identified in section 3.3). As the verbs are classified according to their semantics in the VerbNet, the frames for the particular Bengali verb are assumed to be similar to the frames obtained for the members of its SVS. It has also been observed that the same verb with a different sense can belong to a separate class in the VerbNet. For example, the acquired frames (primary and secondary) for each member of the SVS of the target verbs ("ব্যবহার করা" and "তৈরি করা") have been shown in Table 2. In this way, all possible subcategorization frames for each member of a SVS have been extracted from the generalized search list for our ten target verbs.

3.5 Bengali Verb Subcategorization Frames Acquisition

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.

The syntax of "NP-PP" frame for a Bengali sentence has been acquired by identifying the target verb followed by a NP chunk and a PREP chunk. The sentences containing prepositional frame "PP" do not appear in the Bengali corpus, as there is no concept of preposition in Bengali. But, when we compare the sentences containing postpositional markers, i.e. PREP (postpositions) as a probable argument of the verb, the system gives the desired output.

যার থেকে হাতপাখা
(jar)PRP (theke)PREP (hat-pakha)NN
আর আচ্ছাদন তৈরি
(ar)CC (achhadon)QF (toiri)NN
করেছিলেন ম্যাক্স
(korechilen)VM (Max)NN

All the frames of a SVS corresponding to a Bengali verb have been considered. The Bengali verb "ব্যবহার করা" (*byabahar kara*) in the following sentence has taken the frame "ADVP-PRED" (the word with RB tag) from a different SVS.

কর্মচারীরা
(karmachari ra)NN
বন্ধুত্বপূর্ণ
(bondhuttwapurno)RB
ব্যবহার করেন
(byabahar)NN (karen)VM

Another form of "ADVP-PRED" frame has been obtained by considering the Bengali meaning of the corresponding English adverbial phrase. "There" is an adverbial phrase taken by the "live" verb in English. The corresponding representation in the equivalent Bengali verb is ওখানেই (*okhanei*) as shown in the following sentence. Hence, the frame has been identified.

ওখানেই বাস
(*okhanei*)RB (*bas*)NN
করতে হবে
(*karte*)VM (*habe*)VAUX

The NNPC (Compound proper noun), NNP (Proper noun), NNC (Compound common noun) and NN (Common noun) POS tags help to determine the subjects, objects as well as the locative information related to the verb. In simple sentences the occurrence of these POS tags preceded by the PRP (Pronoun) or NNPC tags and followed by the verb gives similar frame syntax for "Basic Transitive" frame of the VerbNet. Only the components like subject, object and a single verb in Bengali as well as in English sentence can be signified as simple "Basic Transitive" frame.

সে রকম
(*se*)PRP NP((*rakam*)NN
ডিজাইনের কাজ করে
(*designer*)NN) (*kaj*)NN (*kare*)VM

The following example shows that the frame identified from the sentence is also a "transitive frame" and the secondary frame component is a "material object" for that sentence.

একটি ব্যাগেজ
(*ekti*)QC (*bagaze*)NNP
সংগ্রহ করলাম
VGNF((*sangroho*)NN (*korlam*)VM)

The PREP (postposition) followed by a NP phrase and the target verb gives similar syntax for a NP-PP frame but it has been noticed that the secondary frame here can be a component of "Location-PP".

সেতু থেকে
(*setu*)NNP (*theke*)PREP
নানা উদ্ভিদ
NP((*nana*)JJ (*udvid*)NN))
প্রজাতি পর্যবেক্ষণ
(*projati*)JJ (*porjobekkhon*)NN
করলাম
(*korlam*)VM

The sentences where the determiner (DEM) and a NP chunk follow the target verb the sequence (Target verb DEM NP) is considered as the frame of sentential complement "S" for that target verb.

রাম চিৎকার
(*Ram*)NNP (*chitkar*)(NN)
করল যে সে
(*korlo*)VM(*je*)(DEM) (*se*)(PRP)
আর কখনও
(*ar*)CC (*kokhono*)NN
আসবে না
(*asbe*)VM (*na*)NEG

The presence of JJ (Adjective) generally does not play any role in the acquisition process of verb subcategorization frames. There are some frames that did not have any instance in our corpus. Such frames are "Asset-PP", "After-PP", "Location Subject Alternation" and "NP-TO-INF-VC" etc. A close linguistic analysis shows that these frames can also be acquired from the Bengali sentences. They have not occurred in the corpus that has been considered for the analysis in the present work.

4 Evaluation

The set of acquired subcategorization frames or the frame lexicon can be evaluated against a gold standard corpus obtained either through manual analysis of corpus data 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 also no existing dictionary for Bengali that contains subcategorization frames, manual analysis from corpus data is the only method for evaluation. The chunked sentences that contain the ten most frequent verbs have been evaluated manually to prepare the gold standard data.

We have identified 45 different kinds of verbs in the corpus. A detailed statistics of the verb "করা" (*kara*) is presented in Table 3. During the Bengali verb subcategorization frame acquisition process, it has been observed that the simple sentences contain most of the frames that the 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. There are only three occurrences of "করা" (*kara*) as auxiliary in the corpus. These are chunking errors as the verb "করা" (*kara*) does not occur as auxiliary verb.

The verb subcategorization frames acquisition process is evaluated using type precision (the percentage of subcategorization frame types that the system proposes are correct according to the gold standard), type recall (the percentage of subcategorization frame types in the gold standard that the system proposes) and F-measure:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

The system has been evaluated with 100 gold standard test sentences containing ten most frequent verbs and the evaluation results are shown in Table 4. The recall of the system shows a satisfactory performance in producing Bengali verb subcategorization frames but the precision value requires more improvement.

Information	Freq.
Number of sentences in the corpus	1500
Number of different verbs in the corpus	45
Number of inflected forms of the verb “করা” in the corpus	49
Total number of occurrences of the verb “করা” (before stemming) in the corpus	241
Total number of occurrences of the verb “করা” (after stemming) in the corpus	234
Number of sentences where “করা” occurs as a Main Verb (VM)	206
Number of sentences where “করা” occurs as a Simple Main Verb (SVM)	2
Number of sentences where “করা” occurs as a Compound Main Verb (CVM)	204
Number of sentences where “করা” occurs as a Passive Verb (করানো)(done)	25
Number of sentences where “করা” occurs as a Auxiliary Verb (VAUX)	3
Number of simple sentences where “করা” occurs as a Simple Main Verb (SVM)	0
Number of simple sentences where “করা” occurs as a Compound Main Verb (CVM)	127

Table 3. The frequency information of the verb “করা” (*kara*) acquired from the corpus

Measures	Results
Recall	88.23%
Precision	71.11%
F-Measure	79.24

Table 4. The Precision, Recall and F-Measure values of the system

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. The number of different frames acquired for these ten verbs is shown in Table 5.

Bengali Verbs	Subcategory Frames	No. of Frames
তৈরি করা (*toiri kora*)	NP-PP NP-NP	15 3
ব্যবহার করা (*babohar kora*)	NP-ADVP NP-PP NP-ING-SC NP-PP Location-PP	1 2 1 1 1
বাস করা (*bas kora*)	Basic Transitive PP ADVP-PRED	12 1 1
কাজ করা (*kaj kora*)	PP NP-PP	1 11
সংগ্রহ করা (*sangroho kora*)	Transitive (Material obj) PP	1 2
বন্ধ করা (*bondho kora*)	Basic Transitive NP-PP	1 1
চিত্কার করা (*chitkar kora*)	S PP	1 1
ভুল করা (*bhul kora*)	Nil	0
জিজ্ঞাসা করা (*jigyasa kora*)	BT	1
পর্যবেক্ষণ করা (*porjobekkhon kora*)	Transitive (Location-PP) NP-PP	1 1

Table 5. The frequencies of different frames acquired from corpus

5 Conclusion

The acquisition of subcategorization frames for more number of verbs and clustering them will help us to build a verb lexicon for Bengali language. We need to find out Bengali verb subcategorization frames that may not be supported for the corresponding English verb with identical sense.

There is no restriction for domain dependency in this system. For the free-phrase-order languages like Bengali, the overall performance can be increased by proper assumptions, rules and implementation procedures. Verb morphological information, synonymous sets and their possible subcategorization frames are all important information to develop a full-fledged parser for Bengali. This 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. Columbus, Ohio.

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.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, Cambridge, UK.

Diana McCarthy. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. University of Sussex.

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.

Suzanne Stevenson, Paola Merlo. 1999. Automatic Verb Classification using Distributions of Grammatical Features. *EACL-99*, 45-52. Norge.

Sabine Schulte im Walde. 1998. *Automatic Semantic Classification of Verbs According to Their Alternation Behavior.* Master's thesis, Stuttgart.

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. *ANLP-ACL*, 356-363. Washington, D.C.

XTAG Research Group. 2001. A lexicalized tree adjoining grammar for English. *IRCS.* University of Pennsylvania.