Acquiring Verb Subcategorization Frames in Bengali from Corpora

Dipankar Das, Asif Ekbal, and Sivaji Bandyopadhyay

Department of Computer Science and Engineering,
Jadavpur University, Kolkata, India
dipankar.dipnil2005@gmail.com, asif.ekbal@gmail.com,
sivaji_cse_ju@yahoo.com

Abstract. Subcategorization frames acquisition of a phrase can be described as a mechanism to extract different types of relevant arguments that are associated with that phrase in a sentence. This paper presents the acquisition of different subcategory frames for a specific Bengali verb that has been identified from POS tagged and chunked data prepared from raw Bengali news corpus. Syntax plays the main role in the acquisition process and not the semantics like thematic roles. The output frames of the verb have been compared with the frames of its English verb that has been identified using bilingual lexicon. The frames for the English verb have been extracted using Verbnet. This system has demonstrated precision and recall values of 85.21% and 83.94% respectively on a test set of 1500 sentences.

Keywords: Subcategorization, Frames, Acquisition, Target Verb, Bilingual Lexicon, Synonymous Verb Set (SVS), VerbNet, Evaluation.

1 Introduction

Subcategorization refers to certain kinds of relations between words and phrases in a sentence. A subcategorization frame is a statement of what types of syntactic arguments a verb (or adjective) takes, such as objects, infinitives, that-clauses, participial clauses, and subcategorized prepositional phrases [1]. Several large, manually developed subcategorized lexicons are available for English, e.g. the COMLEX Syntax [2] and the ANLT [3] dictionaries. VerbNet (VN) [4] is the largest on-line verb lexicon with explicitly stated syntactic and semantic information based on Levin's verb classification [5]. It is a hierarchical domain-independent, broad-coverage verb lexicon with mappings to other lexical resources.

The collection of different subcategorization frames for other parts-of-speech phrases is also an important task. But the verb phrase of a language usually takes various types of subcategorization frames compared to other parts of speech phrases. There are several reasons for the importance of subcategorization frames for the development of a parser in that language. There are many languages that have no existing parser. So the information acquired from the subcategorization frames can improve the parsing process. Apart from parsing and dictionary preparation, we can use

W. Li and D. Mollá-Aliod (Eds.): ICCPOL 2009, LNAI 5459, pp. 386–393, 2009.

the acquired subcategorization frames for Question-Answering system to retrieve predictable components of a sentence. The acquired subcategorization frames can also be used for phrase alignment in a parallel sentence level corpus to be utilized for training a statistical machine translation system.

The description of a system developed for automatically acquiring six verb subcategorization frames and their frequencies from a large corpus using rules is mentioned in [7]. Development of a mechanism for resolving verb class ambiguities using subcategrization is reported in [6]. All of these works deal with English. A cross-lingual work on learning verb-argument structure for Czech language is described in [8]. In [9], the method consists of different subcategorization issues that may be considered for the purpose of machine aided translation system for Indian languages.

The present work deals with the acquisition of verb subcategorization frames of a specific verb from a Bengali newspaper corpus. The subcategorization of verbs is an essential issue in parsing for the free phrase order languages such as Bengali that has no existing parser. In this paper, we have developed a system for acquiring subcategorization frames for a specific Bengali verb that occurs most frequently in the Bengali news corpus. Using a Bengali–English bilingual lexicon [10], the English verb meanings with its synonyms have been identified for the Bengali verb. All possible acquired frames for each of the English synonyms for the Bengali verb have been acquired from the VerbNet and these frames have been mapped to the Bengali sentences that contain the verb tagged as main verb and auxiliary. It has been experimentally shown that the accuracy of the frame acquisition process can be significantly improved by considering the occurrences of various argument phrases and relevant POS tags before and after the verb in a sentence. Evaluation results with a test set of 1500 sentences show the effectiveness of the proposed model with precision and recall values of 85.21% and 83.94% respectively.

The rest of the paper is organized as follows. Section 2 describes the framework for the acquisition of subcategory frames for a specific Bengali verb. Evaluation results of the system are discussed in section 3. Finally section 4 concludes the paper.

2 Verb Subcategorization Frames Acquisition from Corpus

We developed several modules for the acquisition of subcategorization frames from the Bengali newspaper corpus. The modules consist of POS tagged corpus preparation, Verb identification and selection, English verb determination, VerbNet frames acquisition and Bengali Verb Subcategorization Frame Acquisition.

2.1 POS Tagged Corpus Preparation

In this work, we have used a Bengali news corpus [11] developed from the web-archive 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 [12]. The POS tagger was developed with a tagset of 26 POS tags (http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf), 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 for identifying phrase level information during the acquisition process.

2.2 Verb Identification and Selection

We have partially analyzed the tagged and chunked data to identify the words that are tagged as main verb (VM). The identified main verbs have been enlisted and passed through a stemming process to identify their root forms. Bengali, like any other Indian languages, is morphologically very rich. Different suffixes may be attached to a verb depending on the different features such as Tense, Aspect, and Person. The stemmer uses a suffix list to identify the stem form. Another table stores the stem form and the corresponding root form for each verb.

The specific root verb that occurs most frequently in any inflected form (as main verb with VM POS tag) is taken up 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. We have chosen the verb (*dekha*) (see) from the corpus that has the largest number of occurrences in the corpus.

The specific root verb, selected in the way as described above, has been considered as the target verb in our work. The sentences containing the target verb including their inflected forms are collected. We have also collected the sentences where the target verb appears as an auxiliary verb (VAUX). These sentences are kept separate for further analysis as these sentences are good candidates for subcategorization frames acquisition.

(*ami*)	(*tader*)	(*aamgulo*)	(*khete*)(VM)	(*dekhechi*)(VAUX)
I	them	mangoes	eating	have seen

2.3 English Verb Determination

The verb subcategorization frames for the equivalent English verbs (in the same sense) are the initial set of verb subcategorization frames that are considered as valid for the Bengali verb. This initial set of verb subcategorization frames are validated in the POS tagged corpus. The process described in section 2.2 already identifies the root form of the target verb. To determine equivalent English verbs, we have used the available Bengali-English bilingual dictionary that has been preprocessed for our task. Various syntactical representations of a word entry in the lexicon are analyzed to identify its synonyms and meanings with respect to verb only. The example of an entry in the bilingual lexicon for our target verb (*dekha*) is given as follows.

[*dēkhā*] v to see, to notice; to look; a. seen, noticed adv. in imitation of.

The above lexicon entry for the Bengali word shows that it can have three different POS tags: verb (v), adjective (a) and adverb (adv). We are interested with those entries that appear with the verb POS. Different synonyms for the 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 (*see*, *notice*) and synonym with other sense (*look*) are selected for Bengali verb (*dekha*). Now, the task is to acquire all the existing possible frames for each member of the SVS from the VerbNet.

2.4 VerbNet Frames Acquisition

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 contain the verbs with their possible subcategory frames and membership information is stored in XML file format.

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 Bengali verb (*dekha*) (identified in section 2.3). 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 members of the SVS also occur in separate classes of the VerbNet depending on their senses. The acquired frames (primary and secondary) for each member of the SVS of the verb (*dekha*) have been enlisted based on their occurrences in the VerbNet classes as shown in Table 1. Stimulus is used by verbs of perception for events or objects that elicit some response from an experiencer. This role usually imposes no restrictions. Theme is used for participants in a location or undergoing a change of location. The prepositional phrases that occur with any of these two senses, i.e., theme and stimulus, are defined as the secondary frames.

Table 1. The members and their subcategorization frames extracted from the Verbnet for the corresponding Bengali verb (*dekha*)

SVS (VerbNet classes)	Primary Frames	Secondary Frames
See (see-30.1) Notice (see-30.1-1)	Basic Transitive, S, Attribute Object Possessor-Attribute Factoring Alternation, HOW-S, WHAT-S, NP-INF-OC, NP-ING-OC, POSSING, PP	Stimulus-PP
Look (peer-30.3)	PP	Theme-PP

2.5 Bengali Verb Subcategorization Frame 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 NNPC (Compound proper noun), NNP (Proper noun), NNC (Compound common noun) and NN (Common noun) tags help to determine the subjects, objects as well as the locative information related to verb.

In simple sentences the occurrence of the NNPC, NNP, NNC or NN tags preceded by the PRP (Pronoun) NNP, NNC, NN or NNPC tags and followed by the verb gives similar frame syntax for “Basic Transitive” frame of the VerbNet.

(*ami*)(PRP)	(*kakatua*)(NNP)	(*dekhi*)(VM)
I	parrot	see

The syntax of "WHAT-S" frame for a Bengali sentence has been acquired by identifying the sentential complement part of the verb (*dekha*). The target verb followed by a NP chunk that consists of another main verb and WQ tag (question word) helps to identify the "WHAT-S" kind of frames.

(*ami*)(PRP)	(*dekhlam*)(VM)	(NP)((*tara*) (PP)	(*ki*)(WQ)	(*korche*)(VM))
I	saw	they	what	did

In order to acquire the frame of "NP-ING-OC", we have created the list of possible Bengali inflections that can appear for the English "-ING" inflection. These inflections usually occur in sentences made up of compound verbs with conjunctive participle form (-e) and infinitive form (-te). If the last word of the phrase contains any of these inflections followed by the target verb then it gives a similar description of the VerbNet frame "NP-ING-OC".

(*ami*)(PRP)	(NP) ((*tader*)	(*haste*))	(*dekhechi*)
I	them	laughing	have seen

The presence of JJ (Adjective) generally does not play any role in the acquisition process of verb subcategorization frames. Some frames like "Attribute Object Possessor-Attribute Factoring Alternation", "HOW-S", "Theme-PP" and "Stimulus-PP" did not have any instance in our corpus. A close linguistic analysis shows that these frames can also be acquired from the Bengali sentences.

3 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 sentences retrieved from the chunker (with an accuracy of 89.4%) have been evaluated manually to extract the sentences that are fully correct. These sentences have been considered as our gold standard data for evaluation of subcategorization frames.

A detailed statistics of the verb (*dekha*) is presented in Table 2. Stemming process has correctly identified 276 occurrences of the verb (*dekha*) from its 284 occurrences in the corpus with an accuracy of 97.18%. During the Bengali verb subcategorization frame acquisition process, it has been observed that the simple sentences contain most of the frames that its English verb form usually takes from 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.

The verb subcategorization frames acquisition process is evaluated using type precision, type recall and F-measure. The results have been shown in Table 3. The evaluation

Table 2. The frequency information of the verb (dekha) acquired from the corpus

No. of sentences in the corpus	1500
No. of different verbs in the corpus	45
No. of inflected forms of the verb (*dekha*) in the corpus	22
Total no. of occurrences of the verb (*dekha*) (before stemming) in the corpus	284
No. of sentences where (*dekha*) occurs as a verb	276
No. of sentences where (*dekha*) occurs as a main verb (VM)	251
No. of sentences where (*dekha*) occurs as an auxiliary verb(VAUX)	25
No. of simple sentences where (*dekha*) occurs as a verb	139
No. of simple sentences where (*dekha*) occurs as a main verb (VM)	125
No. of simple sentences where (*dekha*) occurs as an auxiliary verb (VAUX)	14

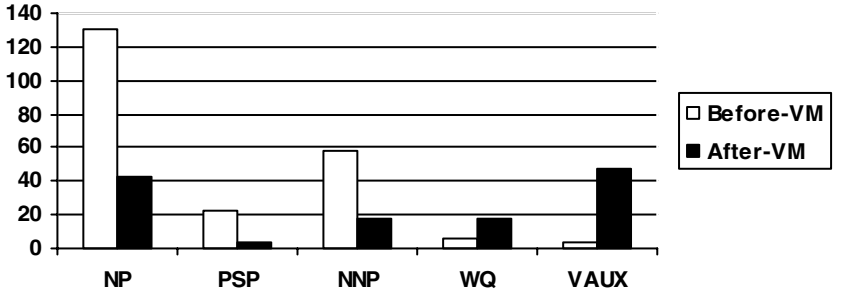

Fig. 1. Frequency of occurrence of different phrases before and after the main verb (VM)

Table 3. Average precision, recall and F-measure for 276 different sentences evaluated against the manual analysis of corpus data

Measure	Stage-1	Stage-2
Recall	83.05%	83.94%
Precision	78.50%	85.21%
F-Measure	80.71	84.57

process has been carried out in two stages. In Stage-1, we have retrieved whatever result we acquired from the corpus and have identified the frames keeping the phrases and their orderings intact. In Stage-2, we have drawn histogram type charts of different phrases as well as tags that appear as frames for a verb before and after of its occurrences in the sentence. This chart is shown in Figure 1. Based on these values, various rules have been applied. Results show that the recall value is not changed so much but there is an appreciable change in the precision values after considering the occurrences of different chunks and tags to select the arguments of the verb.

Number of different types of frames acquired is shown in Table 4. The result shows a satisfactory performance of the system.

Table 4. The frequencies of different frames acquired from corpus

Subcategory Frames	No. of occurrences in the corpus
Basic Transitive	82
S (Sentential Complements)	5
WHAT-S	3
NP-INF-OC	10
NP-ING-OC	2
POSSING	6
PP	12

4 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. There is no restriction for domain dependency in this system. For the free word order languages like Bengali, 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 Q & A system.

References

1. Manning, C.D.: Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In: Proceedings of the 31st Meeting of the ACL, pp. 235–242. ACL, Columbus (1993)
2. Grishman, R., Macleod, C., Meyers, A.: Comlex syntax: building a computational lexicon. In: Proceedings of the International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, pp. 268–272 (1994)
3. Boguraev, B.K., Briscoe, E.J.: Large lexicons for natural language processing utilizing the grammar coding system of the Longman Dictionary of Contemporary English. Computational Linguistics 13(4), 219–240 (1987)
4. Kipper-Schuler, K.: VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA (June 2005)
5. Levin, B.: English Verb Classes and Alternation: A Preliminary Investigation. The University of Chicago Press (1993)
6. Ushioda, A., Evans, D.A., Gibson, T., Waibel, A.: The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora. In: Boguraev, B., Pustejovsky, J. (eds.) Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text, Columbus, Ohio, pp. 95–106 (1993)
7. Lapata, M., Brew, C.: Using subcategorization to resolve verb class ambiguity. In: Fung, P., Zhou, J. (eds.) Proceedings of WVLC/EMNLP, pp. 266–274 (1999)
8. Sarkar, A., Zeman, D.: Automatic extraction of subcategorization frames for czech. In: Proceedings of COLING 2000 (2000)

9. Samantaray, S.D.: A Data mining approach for resolving cases of Multiple Parsing in Machine Aided Translation of Indian Languages. In: International Conference on Information Technology. IEEE Press, Los Alamitos (2007)
10. Samsad Bengali to English Dictionary, http://home.uchicago.edu/~cbs2/banglainstruction.html
11. Ekbal, A., Bandyopadhyay, S.: A Web-based Bengali News Corpus for Named Entity Recognition. Language Resources and Evaluation (LRE) Journal 42(2), 173–182 (2008)
12. Ekbal, A., Haque, R., Bandyopadhyay, S.: Maximum Entropy Based Bengali Part of Speech Tagging. In: Gelbukh, A. (ed.) Advances in Natural Language Processing and Applications, Research in Computing Science (RCS) Journal, vol. 33, pp. 67–78 (2008)