Automatic Extraction of Compound Verbs from Bangla Corpora

Sibansu Mukhopadhyay[1] *Tirthankar Dasgupta*[2] *Manjira Sinha*[2] *Anupam Basu*[2]
(1) Society for Natural Language Technology Research, Kolkata 700091
(2) Indian Institute of Technology Kharagpur, Kharagpur 721302
{sibansu, iamtirthankar, manjira87, anupambas}@gmail.com

ABSTRACT

In this paper we present a rule-based technique for the automatic extraction of Bangla compound verbs from raw text corpora. In our work we have (a) proposed rules through which a system could automatically identify Bangla CVs from texts. These rules will be established on the basis of syntactic interpretation of sentences, (b) we shall explain problems of CV identification subject to the semantics and pragmatics of Bangla language, (c) finally, we have applied these rules on two different Bangla corpuses to extract CVs. The extracted CVs were manually evaluated by linguistic experts where our system and achieved an accuracy of around 70%.

KEYWORDS: COMPOUND VERBS, AUTOMATIC EXTRACTION, VECTOR VERBS

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 153–162,
COLING 2012, Mumbai, December 2012.

1 Introduction

Compound verbs (henceforth CV) are special type of complex predicates consisting of a sequence of two or more verbs acting as a single verb and express a single expression of meaning. However, not all verb sequences are considered as compound verbs. A compound verb consists of a sequence of two verbs, V1 and V2 such that V1 is a common verb with /-e/ [non-finite] inflection marker and V2 is a finite verb that indicates orientation or manner of the action or process expressed by V1 (Dasgupta, 1977). The verb V1 is known as pole and V2 is called as vector. For example, in the sentence রুটিগুলো খেয়ে ফেলো (*/ruTigulo kheYe phela/) "bread-plural-the eat and drop-pres. Imp"* "Eat the breads", the verb sequence "kheYe phela" is an example of CV.

Identification of compound verbs from sentences is useful in many NLP applications including Wordnet development, Information Retrieval, and Machine Translation. However, automatic identification of compound verbs from a given text document is not a trivial task. As mentioned in (Dasgupta, 1977), a sequence of two or more verbs does not always guarantees to be a compound verb. Depending on the context a verb sequence may or may not act as CV. Thus, automatic identification of compound verbs is extremely important and a challenging task.

This paper deals with the rule-based automatic identification of these types of Bangla CV, where V1 is a pole and V2 is a vector. In our work (a) we shall propose rules through which a system could automatically identify Bangla CVs from texts and these rules will be established on the basis of syntactic interpretation of sentences, (b) we shall explain problems of CV identification subject to the semantics and pragmatics of Bangla language, (c) finally we shall make a statistical evaluation of our rules.

The rest of the paper is organized as follows: In section 2 we first perform the linguistic study and the related concepts of the compound verb and different issues related to the automatic extraction of CVs. Section 3 briefly discuss about the different related works done in this area. Section 4 discuss about the different linguistic rules that can be applied to extract CVs from text corpuses. Section 5 presents the experimentations, evaluations and results of our work.

2 Background

An important feature of CV is that the vector verb has no independent meaning. The Vector verb only can affect/support the pole to express some certain pragmatic expression. Linguists call this process of semantic nullification, *process of grammaticalization*. Considering the previous example, if we have a CV like 'খেয়ে ফেলো', we, or any native speaker of Bangla, must not differentiate those two verbs to comprehend the meaning for the each. Bangla native speakers have the sense that the combination 'খেয়ে ফেলো' produces a common meaning which is almost but not really same to the central meaning of the verb 'খা' [*khaa*: 'eat']. And the meaning of the 2nd verb, 'ফেল্' [*phel*: 'drop'] is being bleached out. This second or the vector verb is functionally attached to the pole and grammatically subservient and both the two verbs produce a single meaning.

This is true that each of the verbs of this phrase can be used as a pole or as an independent verb in different contexts. Native speakers are pragmatically competent to understand the phrase duly depending on the certain contexts. If one says in Bangla, "রুটিগুলোর কয়েকটা খেয়ে ফেলে দাও।" (*/ruTigulor kaYekaTaa kheYe phele daao/*), it means "Eat some of the breads and reject the rest of breads." The same combination of the root verbs, "খা" (eat) and "ফেল্", (drop) plays a different role. The interesting thing can be pointed out that there is another one verb 'দাও' is being

attached on the right side of the combination 'খেয়ে ফেলে' and the 'ফেলে' is containing an infinite inflection /-e/. It means 'ফেলে' (phele) is no more playing role of a vector. There the combination of CV is being shifted from 'খা' (eat) and 'ফেল' (drop) to 'ফেল্' (drop) and 'দে' (give). The new CV has a new vector 'দে' (give), which has lost its meaning. There about twenty two verbs are used as vectors which support poles to describe its action in CVs.

2.1 Why Compound Verbs Occur?

Now a psycho-cognitive question arises. Why do the speakers intend to speak a half-hearted semantically bliched out vector verb with a main verb, when she has an option to manage her expression with a pole? We have to say that the poles always do nothing in such cases where speakers need to realize specific genres of daily speech, though it is a question of natural language survey. We have to consider some examples [(8) to (11)], where the poles cannot cover up the specification necessary for the conversation.

(8) তুমি কি কাল টাকাটা অমলের হাতে দিতে পেরেছো?
Expression: "Could you give the money to Amal yesterday?"

(9) তুমি কি কাল টাকাটা অমলের হাতে দিয়ে উঠতে পেরেছো?
Expression: "Could you at last give the money to Amal yesterday? Or something like: "Had you managed your time to give the money to Amal yesterday?"

(10) তুমি কাজটা করেছো।
Expression: You have done the job.

(11) তুমি কাজটা করে ফেলেছো।
Expression: You have finished the job.

(8)-(9) and (10)-(11) are the pairs of sentences, where (9) and (11) have CVs, whereas (8) and (10) have not. The expressions are indicating the differences between CV and non-CV. Some verbs feel lonely. They cannot take the risks of such expressions, which go beyond the physical property of the language. Hook (1974) shows that, a CV can tackle sometimes aspectual or modal expressions in Hindi.
The speech act of vector can be discussed under the area of pragmatics. Ancient Indian tradition of grammar proposed more than one way of understanding meaning of speech. According to such Indian grammatical discourse, native speakers have the potential to under meaning depending on some *lakshans* (indication) of the components. A word or a speech unit has this power of concieving intended meaning (*lakshana shakti*). Vectors also have the power. Let us consider again some examples below.
(12) অমল গান গাইলো।
"Amal sang song."

(13) অমল গান গেয়ে উঠলো।
"Amal started to sing song."

(14) অমল এমন সময় হঠাৎ গান গাইলো।
"At that time, suddenly Amal sang song."

(15) অমল এমন সময় হঠাৎ গান গেয়ে উঠলো।
"At that time, suddenly Amal started to sing the song."

(16) অমল কাল সন্ধ্যেবেলা জলসায় গান গাইবে।
"Amal will sing song in a function tomorrow evening."

(17) *অমল কাল সন্ধ্যেবেলা জলসায় গান গেয়ে উঠবে।
"Amal will start to sing (suddenly) song in a function tomorrow evening."

Now we need to focus on the above sentence (12) which is a simple sentence. Speaker states that "Amal sang song". Sentence (13) has a complex predicate /*geYe uThalo*/. Sentence like (13) expresses that there is a reason for which Amal started to sing song. This sentence also deserves sentential extension with such words like "emana samaYa haThat.h" (at that time suddenly) to relate the reason for which Amal started to sing song [(15)]. Therefore, we see there is a question of appropriateness we have to face regarding understanding the semantics of the vectors. As "*uThalo*" refer to the *sudden* reason behind fact of Amal's singing in past, it cannot be appear in future. That is why the sentence (17) is unacceptable to Bengali speakers. So CV is very specific for its use in the social discourses.

2.2 Challenges in Automatic CV Extraction

Now allow us to turn to the question of identification. We have understood that a couple of verbs (V1+V2) can be considered as a CV when the second verb helps to express some pragmatic specification of the first verb (pole) and when the second verb has no independent meaning. We recognize a CV as we have the pragmatic competence. But how do we *refer* that pragmatic sense which is beyond the physical property? How does a machine understand that these compound components are CV and these are not? After POS-tagger describes a sentence, how can a machine annotate CV, as there are so many possibilities where more than one verb occurs immediately in a syntagmatic order?
This paper tries to reveal such syntactic conditions for the identification of CV without applying pragmatics. And we have targeted to fix certain properties for the CV that a system can easily identify. This has to be said that these conditions will work well to a trained or supervised data but not for the all. However, the easiest way to identify a CV is to mark the vectors in a language first. Let us consider following non-semantic or non-pragmatic conditions to identify a CV in Bangla:

(18) (a) Verb (V1) + verb (V2).
(b) V1 ends with an inflection /-e/ (not –te of course).
(c) V2 is a marked vector.

But these conditions (18) do not properly handle the situations. We have discussed little earlier that all the verb plus verb is not CV. V1 with /-e/-ending is also a common form of infinitive in Bangla. For example, consider (19).

(19) রবি ভাত খেয়ে বলবে।
/*rabi bhaata kheYe balabe*/
"Eating rice Rabi will say."

"*kheYe balabe*" in this sentence (19) is not a CV, though there V1 ends with /-e/. And for (18)/(c), this is said that the vectors, usually, are poles (normal verbs). This is very difficult to

identify a verb as a vector, if we do not have the idea of context or the information about its position in a V+V sequence. If a machine understands that the second position in a V+V is for the vectors and if such a vector, machine finds from the list it may identify that this V+V is a CV. But, even in Bangla, there are many options, where a vector-listed verb plays a role of pole. In those cases, V+V are to be sub-categorized as Pole + Pole, not as Pole + Vector. Table 1 describes a list of 16 vector verbs as proposed by (Paul, 2003).

Table 1: Bangla Vector Verb List

Sl. No.	**Vector Verb**	**Transliteration**	**General Meaning**	**Example**
1	যা	*/yaa/*	Go	সকাল থেকে বৃষ্টি পড়ে যাচ্ছে।
2	আস্	*/aas.h/*	Come	বহুদিন ধরে রবি কাজ ক'রে আসছে।
3	পড়্	*/pa.D.h/*	Fall	রবি আজ সকাল সকাল ঘুম থেকে উঠে পড়ল।
4	ফেল্	*/phel.h/*	Drop	রবি কথাটা বলে ফেলল।
5	দে	*/de/*	Give	রবি বিস্কুটটা খেলো না ফেলে দিলো।
6	নে	*/ne/*	Take	রবি তাড়াতাড়ি হাতের কাজ ক'টা সেরে নিল।
7	মর্	*/mar.h/*	Die	তুমি শ্যামলের জন্য মিথ্যে ভেবে মরছো।
8	বস্	*/bas.h/*	Seat	রবি সকলের সামনে কথাটা বলে বসলো।
9	উঠ্	*/uTh.h/*	Get up	তুমি কি তোমার বাবাকে কথাটা বলে উঠতে পারলে?
10	তুল	*/tul/*	Lift	তুমি কি মানুষকে জাগিয়ে তুলবে ভাবছো?
11	ছাড়্	*/chhaa.D.h/*	Leave	রবি কাজটা করে ছাড়লো।
12	রাখ্	*/raakh.h/*	Put down	মোহর সকাল সকাল কাজ গুছিয়ে রেখেছে।
13	আন্	*/aan.h/*	Bring	কাজটা প্রায় শেষ করে এনেছি।
14	পাঠা	*/paaThaa/*	Send	তিনি বলে পাঠিয়েছেন।

3 Related works

Recent trends in computational linguistics revisit several old issues from hardcore linguistics or traditional grammar. CV is one such issue, natural language scientists have the scope to use it in the computational aspect and experiment through a large linguistic corpus. CV issues in Indian language perspective have been reviewed by many Indian researchers, such as (Alsena, 1991; Abbi, 1991, 1992; Gopalkrishnan and Abbi, 1992; Butt, 1993; Butt, 1995) with a special focus to Hindi (Burton-Page, 1957; Hook, 1974), Urdu (Butt, 1995), Bangla (Sarkar, 1975; Paul, 2003, 2004, 2010), Kashmiri (Kaul, 1985) and Oriya (Mohanty, 1992, 2010).

Paul (2004) has attempted to work on a *constraint-based* and *semantically-grounded* account of Bangla CV within the HPSG (Head-Driven Phrase Structure Grammar) framework (Paul, 2010). An automatic extraction of Hindi CV was presented in (Chakrabarty et al., 2008). They analyses

the Hindi complex predicate system and provides scope of linguist test for identification of Hindi CV. Chakrabarty and Paul, both have conceptualized vector and used an incomplete list of those vectors. An automatic extraction of Bangla complex predicates have been performed by (Das et.al, 2010). The system uses the vectors as proposed in the literature of (Paul, 2003). To the best of our knowledge, this is the only attempt made to extract Bangla CV from the text corpuses.

4 Compound Verb Identification Grammar and Formal Rules

Apart from all and keeping (18) in our mind we can certify some more rules for the identification of CV in Bangla. This section is basically an overall revisit of our entire dialogues. To identify a CV we can follow the following condition:

(I) The common identification of a CV: Verb + Verb [V1 (+ /-e/) + V2 (-so many inflectional endings according to the tense, aspect, modality and so on)].

(II) Noun + Verb combination is not a Compound Verb, it may be considered as Composite Verb (For Example, হাসি পেল, রাগ করলো, ঘুম পেল.)

(III) V1+V2 = CV and v2 has no meaning. V2 is grammatically subservient, i.e., v2 serves or acts in a subordinate capacity and formally attached to the v1. On the other hand v1 is grammatically central. (Hook, 1974; Dasgupta, 1977)

(IV) As V2 plays a role of an essentially subordinate of V1, it cannot even take a modifier.

(V) V1, i.e., a pole must not immediately be followed by a V2, i.e. a vector. For example, কাজটা করে(v1) তুমি ঠিক ফেলবে(v2)। This implies possibilities of such following combinations too:
N+P+N+V = CV
N+P+V+N = CV

(VI) CV collapses if there is an adverb in between V1 and V2. For example, *বইটা পড়ে (v1) তাড়াতাড়ি (adv) ফেল (v2) ।

(VII) If there is a sequence like V+V+V, then first two verbs should be (normally) poles. <v1+v2+v3 = pole + pole + vector>.

(VIII) Direct question to a vector is not allowed. Consider an example,

অমল গল্পের বইটা পড়ে ফেলবে।
Amal story book-sing-the read (inf) drop-future
Amal will finish the story book.

One cannot ask question to the vector of the combination, 'পড়ে ফেলবে';

* অমল কি ফেলবে?
One must ask completely;
অমল কি পড়ে ফেলবে?

This proves CV (V1 + V2) is a single entity and as V2 is subservient, it cannot take a question directly.

(IX) Bangla CV does not allow double negation like English

5 Experimentation and Results

Based on the CV extraction technique discussed in the earlier section, we try to identify Bangla CVs from the Rabindra-Rachanabali [1] and Bankim Rachanabali [2] corpus. The Rabindra-Rachanabali corpus has a collection of 176000 Bangla sentences and the Bankim-Rachanabali corpus has a collection of 34000 sentences. These sentences were POS (part of Speech) tagged using the Bangla POS tagger[3]. From the POS tagged corpus, we have identified all possible sentences containing multiple verbs. We consider the verb + verb combinations within these sentences, as potential CVs. Altogether 26500 sentences containing the potential CVs are identified from the corpus (see Table 1).

Table 1: Corpus Statistics

Total No. of Sentence in Corpus	2,10,000
Total Number of V+V sequence	26500
Total No. of CV Annotated Sentence	6500
Total No. of CVs (identified manually)	895

Table 2: Results of the Compound Verb Extraction Module

	Before POS Modification	After POS Modification
No. of V+V correctly identified as CV by the system (True Positive)	313	427
No. of V+V correctly identified as Not-CV by the system (True Negative)	197	197
No. of V+V falsely identified as CV by the system (False Positive)	201	160
No. of V+V falsely identified as non-CV by the system (False Negative)	184	111
Precision (%)	61	72
Recall (%)	63	79
F-Measure (%)	62	75
Accuracy (%)	57	70

Out of these 26500 Bangla sentences, we have manually annotated 6500 sentences using a Linguist. We then applied the CV extraction rules, mentioned in the previous section. The extracted CVs are then compared with the manually evaluated gold standard data. The summary of the results obtained are depicted in table 2.

[1] www.rabindra-rachanabali.nltr.org
[2] www.bankim-rachanabali.nltr.org
[3] www.nltr.org/downloades/

The result from table 2 implies that, we have a precision of around 61% and a recall value of 63%. The F-measure is coming out to be 62% whereas the overall accuracy of the system comes to be 57%.

The results of table 2 imply that a lot of anomalous verb sequences have been incorrectly identified as CV by the present system. A close observation over the result reveals some interesting findings that may play some crucial role during the extraction of Bangla CV. Our result shows that, whenever a pole is attached with a suffix "-te" the V1+V2 sequence does not belongs to the CV group irrespective of whether V2 belongs to the pre-defined list of vectors. Thus, we have incorporated an additional rule on the system that checks whether a pole verb contains a suffix "-te" along with its vector counterpart. We further observe that, the loss in precision was also caused due to high error in POS tagging. We identified around 30% of errors are generated due to incorrect POS tags. Thus, when measures were taken to eliminate these errors we reached an accuracy of around 70%.

Further, we perform the analysis for each of the individual vector verbs. The result shows that not all vector verbs are equally responsible to form compound verbs. It has been observed that in most of the cases, vector verbs like, তোলা, বেড়ানো, and মরা have a higher tendency of forming compound verbs as compared to vectors like, আসা, থাকা and দেখা. Figure 1 presents the graph that shows the percentage of cases for which different vectors form the compound verb structure.

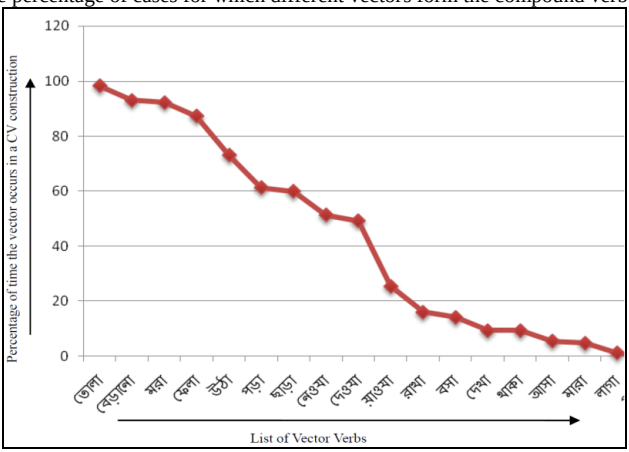

Figure 1: The role of different vector verbs in CV formation

We further classify the vector verbs according to their frequencies. Depending on the frequency, we categorize the list of vectors into the following four different classes:

- Class-I: Frequently occurring vectors with high precision like, ফেলা and উঠা.
- Class-II: Less frequent vectors with high precision like, তোলা and বেড়ানো.
- Class-III: Frequently occurring vectors with low precision like, যাওয়া and আসা
- Class-IV: Less frequent vectors with low precision like, আনা and রাখা.

We have considered the low frequency vectors to be those for which the frequency is below the average frequency. We observe that, the vectors belonging to class-III are very frequent but have

a very low probability of being a CV. Similar effects have been found for vectors belonging to class-IV. We also observed that low frequency vectors have a higher tendency to construct the CV where as high frequency words do not tend to construct CVs.

Conclusion

In the present work we try to automatically extract the Bangla Compound verbs from the literary documents belonging to Rabindra Rachanabali and Bankim Rachanabali. These corpuses have been chosen because of the varied type of text contents. In order to extract the CV we have used the vector verb list as provided by (Paul, 2003). We observe that vector verbs cannot identify the occurrence of compound verbs alone. There are several other features responsible for a verb + verb combination to be a CV. We also saw that, frequencies as well as the type of a vector are very much responsible in order to classify a V+V combination as CV.

In the next stage of our work, we will try to enhance the existing model of CV identification and try to apply the information content measures to identify the degree of compositionality of a given CV.

Acknowledgement

We thank Prof. Probal Dasgupta for providing us useful linguistic insight about the problem and Society for Natural Language Technology Research Kolkata for providing us with the Bangla corpus and partially sponsoring us to conduct the research work.

References

Abbi, Anvita. 1991. Semantics of Explicator Compound Verbs. *In South Asian Languages, Language Sciences*, 13(2): 161-180.

Alsina, Alex. 1996. Complex Predicates: Structure and Theory. *Center for the Study of Language and Information Publications*, Stanford, CA.

Butt, Miriam. 1995. The Structure of Complex Predicates in Urdu. Doctoral Dissertation, Stanford University.

Burton-Page, John. 1957. Compound and conjunct verbs in Hindi. *Bulletin of the School of Oriental and African Studies*, 19: 469-78.

Chakrabarti, Debasri, Mandalia Hemang, Priya Ritwik, Sarma Vaijayanthi, Bhattacharyya Pushpak. 2008. Hindi Compound Verbs and their Automatic Extraction. *International Conference on Computational Linguistics –2008*, pp. 27-30.

Das, D., Pal, S., Mondal, T., Chakraborty, T., Bandyopadhyay, S., "Automatic Extraction of Complex Predicates in Bengali", Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010), pages 37–45, Beijing, August 2010

Dasgupta, Probal. 1977. The internal grammar of Bangla compound verbs. Indian Linguistics 38:2.68-85.

Hook, Peter. 1974. The Compound Verbs in Hindi. *The Michigan Series in South and South-east Asian Language and Linguistics*. The University of Michigan.

Kaul, Vijay Kumar. 1985. The Compound Verb in Kashmiri. Unpublished Ph.D. dissertation. Kurukshetra University.

Mohanty, Panchanan. 2010. WordNets for Indian Languages: Some Issues. *Global WordNet Conference-2010*, pp. 57-64.

Mohanty, Gopabandhu. 1992. The Compound Verbs in Oriya. Ph. D. dissertation, Deccan College Post-Graduate and Research Institute, Pune.

Paul, Soma. 2010. Representing Compound Verbs in Indo WordNet. *Golbal Wordnet Conference- 2010*, pp. 84-91.

Paul, Soma. 2004. An HPSG Account of Bangla Compound Verbs with LKB Implementation. Ph.D dissertation, University of Hyderabad, Hyderabad.

Paul, Soma. 2003. Composition of Compound Verbs in Bangla. *Multi-Verb constructions*. Trondheim Summer School.

Sarkar, Pabitra. 1975. Aspects of Compound Verbs in Bengali. Unpublished M.A. dissertation, Chicago University.