Automatic Extraction of Complex Predicates in Bengali

Dipankar Das      Santanu Pal      Tapabrata Mondal       Tanmoy Chakraborty

Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University
dipankar.dipnil2005@gmail.com,
santanupersonal1@gmail.com,
tapabratamondal@gmail.com, its_tanmoy@yahoo.co.in,
sivaji_cse_ju@yahoo.com

Abstract

This paper presents the automatic extraction of Complex Predicates (CPs) in Bengali with a special focus on compound verbs (Verb + Verb) and conjunct verbs (Noun /Adjective + Verb). The lexical patterns of compound and conjunct verbs are extracted based on the information of shallow morphology and available seed lists of verbs. Lexical scopes of compound and conjunct verbs in consecutive sequence of Complex Predicates (CPs) have been identified. The fine-grained error analysis through confusion matrix highlights some insufficiencies of lexical patterns and the impacts of different constraints that are used to identify the Complex Predicates (CPs). System achieves F-Scores of 75.73%, and 77.92% for compound verbs and 89.90% and 89.66% for conjunct verbs respectively on two types of Bengali corpus.

1 Introduction

Complex Predicates (CPs) contain [verb] + verb (compound verbs) or [noun/adjective/adverb] +verb (conjunct verbs) combinations in South Asian languages (Hook, 1974). To the best of our knowledge, Bengali is not only a language of South Asia but also the sixth popular language in the World\(^1\), second in India and the national language of Bangladesh. The identification of Complex Predicates (CPs) adds values for building lexical resources (e.g. WordNet (Miller et al., 1990; VerbNet (Kipper-Schuler, 2005)), parsing strategies and machine translation systems.

Bengali is less computerized compared to English due to its morphological enrichment. As the identification of Complex Predicates (CPs) requires the knowledge of morphology, the task of automatically extracting the Complex Predicates (CPs) is a challenge. Complex Predicates (CPs) in Bengali consists of two types, compound verbs (CompVs) and conjunct verbs (ConjVs).

The compound verbs (CompVs) (e.g. মেরে ফেলা ‘kill’, বলতে লাগল ‘started saying’) consist of two verbs. The first verb is termed as Full Verb (FV) that is present at surface level either as conjunctive participial form -ে- or the infinitive form -e. The second verb bears the inflection based on Tense, Aspect and Person. The second verbs that are termed as Light Verbs (LV) are polysemous, semantically bleached and confined into some definite candidate seeds (Paul, 2010).

On the other hand, each of the Bengali conjunct verbs (ConjVs) (e.g. যদি করা ‘harpsha

\(^{1}\)http://www.ethnologue.com/ethno_docs/distribution.asp?by=size
kara ‘to depend’, বক্তব্য করা jhakjhak kara ‘to glow’) consists of noun or adjective followed by a Light Verb (LV). The Light Verbs (LVs) bear the appropriate inflections based on Tense, Aspect and Person.

According to the definition of multi-word expressions (MWEs) (Baldwin and Kim, 2010), the absence of conventional meaning of the Light Verbs in Complex Predicates (CPs) entails us to consider the Complex Predicates (CPs) as MWEs (Sinha, 2009). But, there are some typical examples of Complex Predicates (CPs), e.g. তাকা করা dekha kara ‘see-do’ that bear the similar lexical pattern as Full Verb (FV)+ Light Verb (LV) but both of the Full Verb (FV) and Light Verb (LV) loose their conventional meanings and generate a completely different meaning (‘to meet’ in this case).

In addition to that, other types of predicates such as দেনো নেয় niye gelo ‘take-go’ (took and went), দিয়ে দেন diye gelo ‘give-go’ (gave and went) follows the similar lexical patterns FV+LV as of Complex Predicates (CPs) but they are not mono-clausal. Both the Full Verb (FV) and Light Verb (LV) behave like independent syntactic entities and they belong to non-Complex Predicates (non-CPs). The verbs are also termed as Serial Verb (SV) (Mukherjee et al., 2006).

Butt (1993) and Paul (2004) have also mentioned the following criteria that are used to check the validity of complex predicates (CPs) in Bengali. The following cases are the invalid criteria of complex predicates (CPs).

1. Control Construction (CC): লিখে শেষ likhe bollo ‘asked to write’, লিখিতে ঝা করল likhthe bhadhyo korlo ‘forced to write’
   2. Modal Control Construction (MCC): একত হবে jete hobe ‘have to go’ একত হবে khete hobe ‘have to eat’
   3. Passives (Pass) : ধরা পড়ল dhora porlo ‘was caught’, মারা হল mara holo ‘was beaten’

Sometimes, the successive sequence of the Complex Predicates (CPs) presents a problem of deciding the scopes of individual Complex Predicates (CPs) present in that sequence. For example the sequence, উঠে পরো লক্ষমণ uthe pore dekham ‘rise-wear-see’ (rose and saw) seems to contain two Complex Predicates (CPs) (উঠে পরো uthe pore ‘rose’ and পরো লক্ষমণ pore dekham ‘wore and see’). But there is actually one Complex Predicate (CP). The first one উঠে uthe pore ‘rose’ is a compound verb (CompV) as well as a Complex Predicate (CP).

The present task therefore not only aims to extract the Complex Predicates (CPs) containing compound and conjunct verbs but also to resolve the problem of deciding the lexical scopes automatically. The compound verbs (CompVs) and conjunct verbs (ConjVs) are extracted from two separate Bengali corpora based on the morphological information (e.g. participle forms, infinitive forms and inflections) and list of Light Verbs (LVs). As the Light Verbs (LVs) in the compound verbs (CompVs) are limited in number, fifteen predefined verbs (Paul, 2010) are chosen as Light Verbs (LVs) for framing the compound verbs (CompVs). A manually prepared seed list that is used to frame the lexical patterns for conjunct verbs (ConjVs) contains frequently used Light Verbs (LVs).

An automatic method is designed to identify the lexical scopes of compound and conjunct verbs in the long sequences of Complex Predicates (CPs). The identification of lexical scope of the Complex Predicates (CPs) improves the performance of the system as the number of identified Complex Predicates (CPs) increases.

Manual evaluation is carried out on two types of Bengali corpus. The experiments are carried out on 800 development sentences from two corpora but the final evaluation is carried out on 1000 sentences. Overall, the system achieves F-Scores of 75.73%, and 77.92% for compound verbs and 89.90% and 89.66% for conjunct verbs respectively.
The error analysis shows that not only the lexical patterns but also the augmentation of argument structure agreement (Das, 2009), the analysis of Non-MonoClausal Verb (NMCV) or Serial Verb, Control Construction (CC), Modal Control Construction (MCC), Passives (Pass) and Auxiliary Construction (AC) (Butt, 1993; Paul, 2004) are also necessary to identify the Complex Predicates (CPs). The error analysis shows that the system suffers in distinguishing the Complex Predicates (CPs) from the above constraint constructions.

The rest of the paper is organized as follows. Section 2 describes the related work done in this area. The automatic extraction of compound and conjunct verbs is described in Section 3. In Section 4, the identification of lexical scopes of the Complex Predicates (CPs) is mentioned. Section 5 discusses the results of evaluation along with error analysis. Finally, Section 6 concludes the paper.

2 Related Work

The general theory of complex predicate is discussed in Alsina (1996). Several attempts have been organized to identify complex predicates in South Asian languages (Abbi, 1991; Bashir, 1993; Verma, 1993) with a special focus to Hindi (Burton-Page, 1957; Hook, 1974), Urdu (Butt, 1995), Bengali (Sarkar, 1975; Paul, 2004), Kashmiri (Kaul, 1985) and Oriya (Mohanty, 1992). But the automatic extraction of Complex Predicates (CPs) has been carried out for few languages, especially Hindi.

The task described in (Mukherjee et al., 2006) highlights the development of a database based on the hypothesis that an English verb is projected onto a multi-word sequence in Hindi. The simple idea of projecting POS tags across an English-Hindi parallel corpus considers the Complex Predicate types, adjective-verb (AV), noun-verb (NV), adverb-verb (Adv-V), and verb-verb (VV) composites. A similar task (Sinha, 2009) presents a simple method for detecting Complex Predicates of all kinds using a Hindi-English parallel corpus. His simple strategy exploits the fact that Complex Predicate is a multi-word expression with a meaning that is distinct from the meaning of the Light Verb. In contrast, the present task carries the identification of Complex Predicates (CPs) from monolingual Bengali corpus based on morphological information and lexical patterns.

The analysis of V+V complex predicates termed as lexical compound verbs (LCpdVs) and the linguistic tests for their detection in Hindi are described in (Chakrabarti et al., 2008). In addition to compound verbs, the present system also identifies the conjunct verbs in Bengali. But, it was observed that the identification of Hindi conjunct verbs that contain noun in the first slot is puzzling and therefore a sophisticated solution was proposed in (Das, 2009) based on the control agreement strategy with other overtly case marked noun phrases. The present task also agrees with the above problem in identifying conjunct verbs in Bengali although the system satisfactorily identifies the conjunct verbs (ConjVs).

Paul (2003) develops a constraint-based mechanism within HPSG framework for composing Indo-Aryan compound verb constructions with special focus on Bangla (Bengali) compound verb sequences. Postulating semantic relation of compound verbs, another work (Paul, 2009) proposed a solution of providing lexical link between the Full verb and Light Verb to store the Compound Verbs in Indo WordNet without any loss of generalization. To the best of our knowledge, ours is the first attempt at automatic extraction of Complex Predicates (CPs) in Bengali.

3 Identification of Complex Predicates (CPs)

The compound verbs (CompVs) and conjunct verbs (ConjVs) are identified from the shallow parsed result using a lexical pattern matching technique.

3.1 Preparation of Corpora

Two types of Bengali corpus have been considered to carry out the present task. One corpus is collected from a travel and tourism domain and another from an online web archive of Rabindranath Rachanabali 2. Rabindra Rachanabali corpus is a large collection of short stories of Rabindranath Tagore. The for-

2 www.rabindra-rachanabali.nltr.org
The EILMT travel and tourism corpus is obtained from the consortium mode project “Development of English to Indian Languages Machine Translation (EILMT)” System”. The second type of corpus is retrieved from the web archive and pre-processed accordingly. Each of the Bengali corpora contains 400 and 500 development and test sentences respectively.

The sentences are passed through an open source Bengali shallow parser. The shallow parser gives different morphological information (root, lexical category of the root, gender, number, person, case, vibhakti, tam, suffixes etc.) that help in identifying the lexical patterns of Complex Predicates (CPs).

3.2 Extracting Complex Predicates (CPs)

Manual observation shows that the Complex Predicates (CPs) contain the lexical pattern \{XXX (n/adj) [YYY (v)]\} in the shallow parsed sentences where XXX and YYY represent any word. But, the lexical category of the root word of XXX is either noun (n) or adjective (adj) and the lexical category of the root word of YYY is verb (v). The shallow parsed sentences are pre-processed to generate the simplified patterns. An example of similar lexical pattern of the shallow parsed result and its simplified output is shown in Figure 1.

![Figure 1. Example of a pre-processed shallow parsed result.](image_url)

The corresponding lexical categories of the root words অধ্যাপন ‘study’ (e.g. noun for ‘n’) and কর ‘do’ (e.g. verb for ‘v’) are shown in bold face in Figure 1. The following example is of conjunct verb ( ConjV).

The extraction of Bengali compound verbs (CompVs) is straightforward rather than conjunct verbs ( ConjVs). The lexical pattern of compound verb is \{XXX(adj) [YYY (v)]\} where the lexical or basic POS categories of the root words of “XXX” and “YYY” are only verb. If the basic POS tags of the root forms of “XXX” and “YYY” are verbs (v) in shallow parsed sentences, then only the corresponding lexical patterns are considered as the probable candidates of compound verbs (CompVs).

Example 1:

<table>
<thead>
<tr>
<th>ভ্রমণ</th>
<th>কারণ</th>
<th>অভ্যন্তর</th>
<th>একটি</th>
</tr>
</thead>
<tbody>
<tr>
<td>ঘাটা</td>
<td>গ্রাম</td>
<td>মাঠ</td>
<td>খেলা</td>
</tr>
</tbody>
</table>

Example 2:

<table>
<thead>
<tr>
<th>ভ্রমণ</th>
<th>কারণ</th>
<th>অভ্যন্তর</th>
<th>একটি</th>
</tr>
</thead>
<tbody>
<tr>
<td>ঘাটা</td>
<td>গ্রাম</td>
<td>মাঠ</td>
<td>খেলা</td>
</tr>
</tbody>
</table>

Bengali, like any other Indian languages, is morphologically very rich. Different suffixes may be attached to a Light Verb (LVs) (in this case [YYY]) depending on the various features such as Tense, Aspect, and Person.

In case of extracting compound verbs (CompVs), the Light Verbs are identified from a seed list (Paul, 2004). The list of Light Verbs is specified in Table 1. The dictionary forms of the Light Verbs are stored in this list. As the Light Verbs contain different suffixes, the primary task is to identify the root forms of the Light Verbs (LVs) from shallow parsed result. Another table that stores the root forms and the corresponding dictionary forms of the Light Verbs is used in the present task. The table contains a total number of 378 verb entries including Full Verbs (FVs) and Light Verbs (LVs). The dictionary forms of the Light Verbs (LVs) are retrieved from the Table.

On the other hand, the conjunctive participial form -ত্ত্বীয়া -e/iya or the infinitive form -ত্ত্বীয়া -te/ite are attached with the Full Verbs.
(FVs) (in this case [XXX]) in compound verbs (CompVs). इया / iya and इटे / ite are also used for conjunctive participial form -स -े or the infinitive form -ते/-टे respectively in literature. The participial and infinitive forms are checked based on the morphological information (e.g. suffixes of the verb) given in the shallow parsed results. In Example 1, the Full Verb (FV) contains इया / iya suffix. If the dictionary forms of the Light Verbs (LVs) are present in the list of Light Verbs and the Full Verbs (FVs) contain the suffixes of -स/े/- प/े, both verbs are combined to frame the patterns of compound verbs (CompVs).

| aSa ‘come’ | dâRa ‘stand’ |
| rakha ‘keep’ | ana ‘bring’ |
| deoya ‘give’ | pOra ‘fall’ |
| paTha ‘send’ | bERano ‘roam’ |
| neoya ‘take’ | tola ‘lift’ |
| bOsa ‘sit’ | oTha ‘rise’ |
| jaoya ‘go’ | chaRa ‘leave’ |
| phEla ‘drop’ | mOra ‘die’ |

Table 1. List of Light Verbs for compound verbs.

The identification of conjunct verbs (ConjVs) requires the lexical pattern (Noun / Adjective + Light Verb) where a noun or an adjective is followed by a Light Verb (LV). The dictionary forms of the Light Verbs (LVs) that are frequently used as conjunct verbs (ConjVs) are prepared manually. The list of Light Verbs (LVs) is given in Table 2. The detection of Light Verbs (LVs) for conjunct verbs (ConjVs) is similar to the detection of the Light Verbs (LVs) for compound verbs (CompVs) as described earlier in this section. If the basic POS of the root of the first words ([XXX]) is either “noun” or “adj” (n/adj) and the basic POS of the following word ([YYY]) is “verb” (v), the patterns are considered as conjunct verbs (ConjVs). The Example 2 is an example of conjunct verb (ConjV).

For example, अककक करा (jhajhak kara ‘to glow’), भककक करा (taktak ‘to glow’), चुपचाप करा (chupchap kara ‘to silent’) etc are identified as conjunct verbs (ConjVs) where the basic POS of the former word is an adjective (adj) followed by करा kara ‘to do’, a common Light Verb.

| deoya ‘give’ | kara ‘do’ |
| neoya ‘take’ | laga ‘start’ |
| paoya ‘pay’ | kata ‘cut’ |

Table 2. List of Light Verbs for conjunct verbs.

Example 3:
अककक/adj, अककक /JJ/JJP/ (अककक ‘adj’) #
करित/Verb| करित/VM/VGF/ (करि ‘v’ + 5 * ‘-त्त’)

But, the extraction of conjunct verbs (ConjVs) that have a “noun+verb” construction is descriptively and theoretically puzzling (Das, 2009). The identification of lexical patterns is not sufficient to recognize the compound verbs (CompVs). For example, बोि दोिya boi deoya ‘give book’ and राऺा दोिya bharsa deoya ‘to assure’ both contain similar lexical pattern (noun+verb) and same Light Verb दोिya. But, धरसा दोिya bharsa deoya ‘to assure’ is a conjunct verb (ConjV) whereas बोि दोिya boi deoya ‘give book’ is not a conjunct verb (ConjV). Linguistic observation shows that the inclusion of this typical category into conjunct verbs (ConjVs) requires the additional knowledge of syntax and semantics.

In connection to conjunct verbs (ConjVs), (Mohanty, 2010) defines two types of conjunct verbs (ConjVs), synthetic and analytic. A synthetic conjunct verb is one in which both the constituents form an inseparable whole from the semantic point of view or semantically non-compositional in nature. On the other hand, an analytic conjunct verb is semantically compositional. Hence, the identification of conjunct verbs requires knowledge of semantics rather than only the lexical patterns.

It is to be mentioned that sometimes, the negative markers (ना no, नाइ nai) are attached with the Light Verbs दोिya uthona ‘do not get up’, फेलोंa phelona ‘do not throw’. Negative attachments are also considered in the present task while checking the suffixes of Light Verbs (LVs).

4 Identification of Lexical Scope for Complex Predicates (CPs)

The identification of lexical scopes of the Complex Predicates (CPs) from their successive sequences shows that multiple Complex
Predicates (CPs) can occur in a long sequence. An automatic method is employed to identify the Complex Predicates (CPs) along with their lexical scopes. The lexical category or basic POS tags are obtained from the parsed sentences.

If the compound and conjunct verbs occur successively in a sequence, the left most two successive tokens are chosen to construct the Complex Predicate (CP). If successive verbs are present in a sequence and the dictionary form of the second verb reveals that the verb is present in the lists of compound Light Verbs (LV), then that Light Verb (LV) may be a part of a compound verb (CompV). For that reason, the immediate previous word token is chosen and tested for its basic POS in the parsed result. If the basic POS of the previous word is “verb (v)” and any suffixes of either conjunctive participial form -e/iya or -te/ite is attached to the previous verb, the two successive verbs are grouped together to form a compound verb (CompV) and the lexical scope is fixed for the Complex Predicate (CP).

If the previous verb does not contain -e/iya or -te/ite inflections, no compound verb (CompV) is framed with these two verbs. But, the second Light Verb (LV) may be a part of another Complex Predicate (CP). This Light Verb (LV) is now considered as the Full Verb (FV) and its immediate next verb is searched in the list of compound Light Verbs (LVs) and the formation of compound verbs (CompVs) progresses similarly. If the verb is not in the list of compound Light Verbs, the search begins by considering the present verb as Full Verb (FV) and the search goes in a similar way.

The following examples are given to illustrate the formation of compound verbs (CompVs) and find the lexical scopes of the compound verbs (CompVs).

```
I <fell down while walking>.
```

```
Another example is given as follows.
```
Here, “chalte give pore gelam” is a verb group. The two left most verbs chalte give are picked and the dictionary form of the second verb is searched in the list of compound Light Verbs. As the dictionary form (jaoya ‘go’) of the verb giye is present in the list of compound Light Verbs (as shown in Table 1), the immediate previous verb chalte is checked for inflections -e/iya or -te/ite. As the verb chalte contains the inflection -te, the verb group chalte giye is a compound verb (CompV) where giye is a Light Verb and chalte is the Full Verb with inflection (-te). Next verb group, gelam is identified as compound verb (CompV) in a similar way (pore + (-e) gelam (jaoya ‘go’)). Another example is given as follows.

```
I <get up and saw> that you are not here
```

```
Here, uthe pore dekhalam is another verb group. The immediate next verb of uthe is pore that is chosen and its dictionary form is searched in the list of compound Light Verbs (LV) similarly. As the dictionary form (pore) of the verb pore is present in the list of Light Verbs and the verb uthe contains the inflection -e, the consecutive verbs frame a compound verb (CompV) uthe pore where uthe is a Full Verb with inflection -e and pore is a Light Verb. The final verb dekhalam is chosen and as there is no other verb present, the verb dekhalam is excluded from any formation of compound verb (CompV) by considering it as a simple verb.

Similar technique is adopted for identifying the lexical scopes of conjunct verbs (ConjVs). The method seems to be a simple pattern matching technique in a left-to-right fashion but it helps in case of conjunct verbs (ConjVs). As the noun or adjective occur in the first slot of conjunct verbs (ConjVs) construction, the search starts from the point of noun or adjective. If the basic POS of a current token is either “noun” or “adjective” and the dictionary form of the next token with the basic POS “verb (v)” is in the list of conjunct Light Verbs (LVs), then the two successive tokens are
combined to frame the pattern of a conjunct verb (ConjV).

For example, the identification of lexical scope of a conjunct verb (ConjV) from a sequence such as উপার্জন করতে গেলাম ‘uparjon korte gelam’ (went to earn) identifies the conjunct verb (ConjV) উপার্জন korte. There is another verb group করতে গেলাম korte gelam that seems to be a compound verb (CompV) but is excluded by considering গেলাম gelam as a simple verb.

5 Evaluation

The system is tested on 800 development sentences and finally applied on a collection of 500 sentences from each of the two Bengali corpora. As there is no annotated corpus available for evaluating Complex Predicates (CPs), the manual evaluation of total 1000 sentences from the two corpora is carried out in the present task.

The recall, precision and F-Score are considered as the standard metrics for the present evaluation. The extracted Complex Predicates (CPs) contain compound verb (CompV) and conjunct verbs (ConjVs). Hence, the metrics are measured for both types of verbs individually. The separate results for two separate corpora are shown in Table 3 and Table 4 respectively. The results show that the system identifies the Complex Predicates (CPs) satisfactorily from both of the corpus. In case of Compound Verbs (CompVs), the precision value is higher than the recall. The lower recall value of Compound Verbs (CompVs) signifies that the system fails to capture the other instances from overlapping sequences as well as non-Complex predicates (non-CPs).

But, it is observed that the identification of lexical scopes of compound verbs (CompVs) and conjunct verbs (ConjVs) from long sequence of successive Complex Predicates (CPs) increases the number of Complex Predicates (CPs) entries along with compound verbs (CompVs) and conjunct verbs (ConjVs). The figures shown in bold face in Table 3 and Table 4 for the Travel and Tourism corpus and Short Story corpus of Rabindranath Tagore indicates the improvement of identifying lexical scopes of the Complex Predicates (CPs).

In comparison to other similar language such as Hindi (Mukerjee et al., 2006) (the reported precision and recall are 83% and 46% respectively), our results (84.66% precision and 83.67% recall) are higher in case of extracting Complex Predicates (CPs). The reason may be of resolving the lexical scope and handling the morphosyntactic features using shallow parser.

In addition to Non-MonoClausal Verb (NMVC) or Serial Verb, the other criteria (Butt, 1993; Paul, 2004) are used in our present diagnostic tests to identify the complex predicates (CPs). The frequencies of Compound Verb (CompV), Conjunct Verb (ConjV) and the instances of other constraints of non Complex Predicates (non-CPs) are shown in Figure 2. It is observed that the numbers of instances of Conjunct Verb (ConjV), Passives (Pass), Auxiliary Construction (AC) and Non-MonoClausal Verb (NMVC) or Serial Verb are comparatively high than other instances in both of the corpus.

<table>
<thead>
<tr>
<th>EILMT</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound Verb</td>
<td>65.92%</td>
<td>80.11%</td>
<td>72.32%</td>
</tr>
<tr>
<td>(CompV)</td>
<td>70.31%</td>
<td>82.06%</td>
<td>75.73%</td>
</tr>
<tr>
<td>Conjunct Verb</td>
<td>94.65%</td>
<td>80.44%</td>
<td>86.96%</td>
</tr>
<tr>
<td>(ConjV)</td>
<td>96.96%</td>
<td>83.82%</td>
<td>89.90%</td>
</tr>
</tbody>
</table>

Table 3. Recall, Precision and F-Score of the system for acquiring the CompVs and ConjVs from EILMT Travel and Tourism Corpus.

<table>
<thead>
<tr>
<th>Rabindra Rachanabali</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound Verb</td>
<td>68.75%</td>
<td>81.81%</td>
<td>74.71%</td>
</tr>
<tr>
<td>(CompV)</td>
<td>72.22%</td>
<td>84.61%</td>
<td>77.92%</td>
</tr>
<tr>
<td>Conjunct Verb</td>
<td>94.11%</td>
<td>83.92%</td>
<td>88.72%</td>
</tr>
<tr>
<td>(ConjV)</td>
<td>95.23%</td>
<td>84.71%</td>
<td>89.66%</td>
</tr>
</tbody>
</table>

Table 4. Recall, Precision and F-Score of the system for acquiring the CompVs and ConjVs from Rabindra Rachanabali corpus.
The error analysis is conducted on both of the corpus. Considering both corpora as a whole single corpus, the confusion matrix is developed and shown in Table 5. The bold face figures in Table 5 indicate that the percentages of non-Complex Predicates (non-CPs) such as Non-MonoClausal Verbs (NMCV), Passives (Pass) and Auxiliary Construction (AC) that are identified as compound verbs (CompVs). The reason is the frequencies of the non-Complex Predicates (non-CPs) that are reasonably higher in the corpus. In case of conjunct verbs (ConjVs), the Non-MonoClausal Verbs (NMCV) and Auxiliary Construction (AC) occur as conjunct verbs (ConjVs). The system also suffers from clausal detection that is not attempted in the present task. The Passives (Pass) and Auxiliary Construction (AC) requires the knowledge of semantics with argument structure knowledge.

6 Conclusion

In this paper, we have presented a study of Bengali Complex Predicates (CPs) with a special focus on compound verbs, proposed automatic methods for their extraction from a corpus and diagnostic tests for their evaluation. The problem arises in case of distinguishing Complex Predicates (CPs) from Non-MonoClausal verbs, as only the lexical patterns are insufficient to identify the verbs. In future task, the subcategorization frames or argument structures of the sentences are to be identified for solving the issues related to the errors of the present system.

References


