Shallow morphology based complex predicates extraction in Oriya

R.C. Balabantaray
IIIT BHUBANESWAR
Gothapatana, Malipada
Bhubaneswar, India

M.K. Jena
DRIEMS
Cuttack, Orissa
India

S. Mohanty
Utkal University
Bhubaneswar
Orissa, India

ABSTRACT

This paper presents the extraction of Complex Predicates (*CPs*) in Oriya based on shallow morphology and available seed lists of verbs. Generally Oriya language is a free word order language. Free word order languages have relatively unrestricted local word group or phrase structures that make the problem of complex predicates extraction quite challenging. The complex predicates are generally the special multi word expression which is extracted with a special emphasis on compound verbs (*Verb + Verb*) and conjunct verbs (*Noun /Adjective +Verb*)/ (*Verb + Noun /Adjective*). The lexicalization of compound and conjunct verbs is done based on the information of shallow morphology. Lexical scopes of compound and conjunct verbs in consecutive sequence of Complex Predicates (*CPs*) have been identified. Aim of the current work is, to investigate the possibility of improving the accuracy of complex predicates extraction making it sensitive to verb sub categorization and to evaluate the recall, precision and F-score on different operational environment.

Keywords

Complex predicates (cps), shallow morphology, free word order, multi word expression(MWE) , *Control Construction* (*CC*), *Modal Control Construction* (*MCC*), *Passives* (*Pass*), *Auxiliary Construction* (*AC*)

1. INTRODUCTION

Automatic multi word extraction (MWE) extraction techniques especially complex predicates extraction operates on by using statistical methods of evaluation. The complex predicates (*CPs*) contain [*verb*] +*verb* (*compound verbs*) or [*noun/ adjective/adverb*], verb (*conjunct verbs*) combinations in *South Asian languages* (Hook, 1974). Informally, complex predicates extraction is a multi word chunking task (MWC). .MWC is a chunk with two or more words, where each word links to a semantic head through different dependency relations. Here we are emphasizing the cp extraction for Oriya. But the peculiarities of Oriya language is that the construction of sentences is relatively free. Especially the cp extraction in case of poem is relatively confusing. So the thrust is given for cp extraction in Oriya prose.

Identifying Complex Predicates (*CPs*) facilitates in adding values for building lexical resources (e.g. WordNet (Miller *et al*.,1990; VerbNet (Kipper-Schuler, 2005)) and machine translation systems. The complex predicates extraction can be considered as a type of clause identification.(Ghosh et al ,2010)

Oriya is less computerized in comparison 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 Oriya consists of two types, compound verbs (*CompVs*) and conjunct verbs (*ConjVs*).The compound verbs (*CompVs*) (e.g. *maridela,* 'kill', chalibaku lagila, "started walking ") consist of two verbs. The first verb is termed as *Full Verb* (*FV*) that is present at surface level either as conjunctive participle form. 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). The Conjunct verbs (*ConjVs*) are like (e.g. bharasha kara (adj+verb) 'to depend', samparka rakha (verb+noun) 'keep relationship') 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.*

The sentence is first shallow parsed and the lexical pattern is identified based morphological analysis of the word basing upon information available in Oriya corpora. The information is tagged to the word by analyzing its root word, present category like type (for example noun, verb, adjective etc.) tense, aspect and inflections etc.

2. ANALYSIS OF THE PROBLEM

The problem of complex predicate extraction is analyzed based upon the linguistic information available to us. Generally the complex predicates we mean a group of words having close relationship with each other in terms of ordering of words according to its syntactic categories. Here we are using shallow morphology for identifying the lexical pattern for identifying cps

i.e. Compound verbs (*Verb + Verb*) and conjunct verbs (*Noun /Adjective +Verb*)/ (*Verb + Noun /Adjective*).

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. *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 khai gala 'eat-go' (took and went), dekhi *gala* 'see-go' (see 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 Oriya.

The following cases are the invalid criteria of complex predicates (*CPs*).

1. *Control Construction* (*CC*)*: jiba ku kahila* 'said to go',

Khaibaku badhya kala 'forced to eat'

2. *Modal Control Construction* (*MCC*)*: lekhibar achi'* have to write' padhibar achi 'have to read'

3. *Passives* (*Pass*) *: khia haba* 'will be eaten', *hana haba* 'was slaughtered'

4. *Auxiliary Construction* (*AC*)*:* jau achi 'is going', dia hela 'had given'.

Sometimes, the successive sequence of the Complex Predicates (*CPs*) shows a problem of deciding the scopes of individual Complex Predicates (*CPs*) present in that sequence. For example the sequence, *dain padi bcnhila 'jump-fall-live''* (*jumped and saved*) seems to contain two Complex Predicates (*CPs*) (dain padi 'jump' and *padi bnchila* 'saved'). But there is actually one Complex Predicate (*CP*). The first one dain padi 'jump' is a compound verb (*CompV*) as well as a Complex Predicate (*CP*). Another one is banchila 'saved' that is a simple verb. As the sequence is not monoclausal, the Complex Predicate (*CP*) dain padi 'jump' associated with *banchila* 'save' is to be separated by a lexical boundary. Thus the determination of lexical scopes of Complex Predicates (*CPs*) from a long consecutive sequence is indeed a crucial task. 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 Oriya 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, ten predefined verbs 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*). Another feature for consideration is complex predicates extraction in Oriya prose and poem where ordering differs in case of conjunct verb and compound verb.

For example:

Conjunct verb: vala dela.(adj+verb, good give) ' sufficiently gave' ,this cp can be written especially in poem dela vala (verb+adj, give good) gave sufficiently.

Compound verb : kahi dela (Full verb+ Light verb, tell give) 'told' in case of prose.

This may happen in poem the same compound verb can be written as dela kahi (light verb+ full verb, 'give tell') 'told'

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 a Oriya corpus.

3. EXTRACTING COMPLEX PREDICATES (CPs)

Experimenting manually for extracting the Complex Predicates (*CPs*) for the sentences containing the lexical pattern {MMM] (**n**/**adj**) [NNN] (**v**)} in the shallow parsed sentences where MMM and NNN represent any word. But, the lexical category of the root word of MMM is either noun (**n)** or adjective (**adj**) and the lexical category of the root word of is verb (**v**).

3.1 About Oriya Corpora

The Oriya corpora used in this experiment is developed by CIL, Mysore, from which we have taken 1000 test sentences. Those sentences are passed to our approach.

3.2 The model of cp extraction procedure

The shallow parsed sentences are pre-processed to generate the simplified patterns. The procedure of extracting the cps is as such

1) The input file is fed into the system and the sentences are taken for shallow parsing one by one
2) The root of each lexical category is determined.
3) Cps are determined from the probable candidates
4) Check the candidates based upon the invalid criteria like whether the word group is coming under CC, MCC, PASS and AC.
5) Cp extraction

The following figure demonstrates the procedure of cp extraction from preprocessed shallow parsed text and an example of similar lexical pattern of the shallow parsed result and its simplified output is shown below.

(NP Vata **NO** (vata ,n,…O))
(VP Khaa **V** (khaa ,v, 'p' ,**r. khaiba**))

Table 1. Preprocessed shallow parsed result

The following example is of conjunct verb (*ConjV*). The extraction of Oriya compound verbs (*CompVs*) is straightforward rather than conjunct verbs (*ConjVs*). The lexical pattern of compound verb is {[*MMM*](**v**) [NNN] (**v**)} where the lexical or basic POS categories of the root words of "MMM" and "NNN" are only verb. If the basic POS tags of the root forms of "MMM" and "NNN*"* 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 is a compound verb (*CompV*) but Example 2 is not.

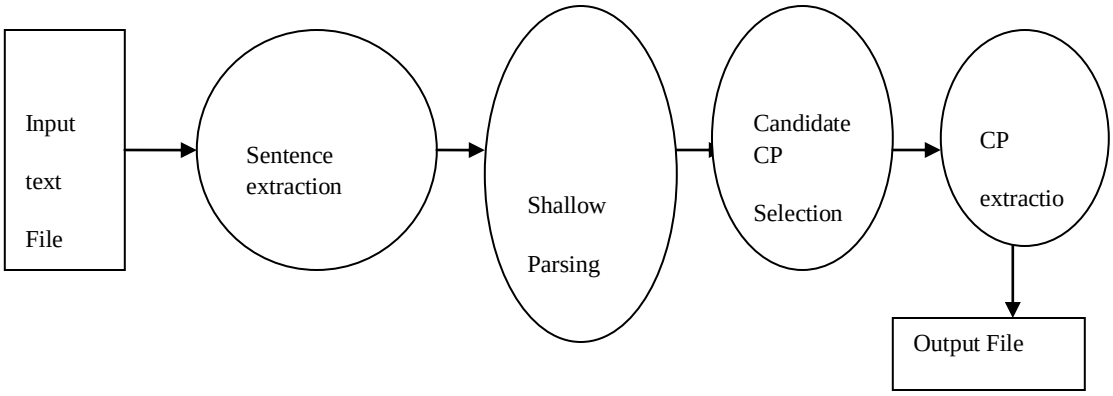

Figure 1. CP Extraction Procedure

The corresponding lexical categories of the root words vata 'rice' (e.g. *noun* for '**n**') and khaa, 'eat (e.g. *verb* for '**v**') are shown in bold face in Table 2.

Example 1 "Mu 'I' (**pn**) gapa 'story' (**NN)** (kahibaku 'telling' (**V**) lagili 'saying' (**V**)CompVs)"

In Example 2, the lexical category or the basic POS of the *Full Verb* (*FV*) is noun (*n*) and hence the pattern is discarded as non-compound verb (*non-CompV*).

*Example 2: "mote 'me' (**pn**) se 'he'(**pn**) badhya 'forced' (**V**) kari (**V**) dekheila 'shown'(**v**) "*

Oriya, like any other Indian languages, is morphologically very rich. Different suffixes may be attached to a *Light Verb* (*LVs*) (in this case [NNN]) 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 common *Light Verbs* is specified in Table 2

Oriya light verb	English Counterpart
Kariba	do
Deba	give
patHa	send
Basiba	Sit
tHiaheba	stand
utHa	awake
kata	cut

Table 2 : List of common light verbs

Irrespective of this, sometimes negative markers are added with light verb (na, nai maens NO). For example uthilana nai " awakened or not", karana 'don't do"

So, the suffixes to be checked for the identification of light verb.

4. IDENTIFYING COMPLEX PREDICATES (CPs) LEXICAL SCOPE

Lexical scopes of the Complex Predicates (*CPs*) from their successive sequences shows that 19 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 or the infinitive form 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*).

For example:

sethare (dekHi jiba(CP, infinitive form)), kahinkina gote pagili katuri (dHari kari(CP infinitive form)) basithiba "beware up ,because one mad woman is sitting by holding a knife"

5. EVALUATION

The system is tested on 800 development sentences and finally applied on a collection of 500 sentences from each of the Oriya corpora. As there is no annotated corpus available for evaluating Complex Predicates (*CPs*), the manual evaluation of total 1000 sentences from the Oriya 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 result for the corpora are shown in Table 3 . The results show that the system identifies the Complex Predicates (*CPs*) satisfactorily from the corpora. The error analysis is also conducted for the 4 invalid criteria.

Oriya Corpora	Recall	precision	F-score
CompVs	61.2	76.3	71.6
ConjVs	87.4	79.2	89.90

Table 3: The Recall, precision, F-score of the experiment on oriya corpora for 'Cp' extraction

6. CONCLUSION

In this paper, we have presented a study of Oriya Complex Predicates (*CPs*) with a special focus on compound verbs and the 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-Mono-Clausal verbs, as only the lexical patterns are insufficient to identify the verbs. In future task, the sub categorization frames or argument structures of the sentences are to be identified for solving the issues related to the errors of the present system.

7. REFERENCES

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

[2] Timothy, Baldwin, Su Nam Kim. 2010. Multiword Expressions. *In Nitin Indurkhya and Fred J. Damerau (eds.) Handbook of Natural Language Processing, Second Edition, Chapman & Hall/CRC*, London, UK, pp. 267-292

[3] Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine Miller. 1990. Five Papers on WordNet. *CSL Report 43*, Cognitive Science Laboratory, Princeton University, Princeton.

[4] Das, Pradeep Kumar. 2009. The form and function of Conjunct verb construction in Hindi. *Global Association of Indo-ASEAN Studies*, Daejeon, South Korea.

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

[6] Bharati,Chaitanya and Singhal "Natural language processing a paninian perspective",PHI.

[7] Mohapatra, Pandit and Das "Sarbasara Byakarana" Student Book Store, Cuttack.

[8] Sarangi Nrusingha,"Bruhat Oriya Byakarana" Satyanarayana Book Store Binod Bihari, Cuttack.

[9] Mohanty et.al,"Oriya wordNet" proceedings of 1st global Wordnet conference,2002, CIL, Moysore.