Volume 1, No. 4, Nov-Dec 2010
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Experimenting Oriya Text Chunking with Divide-Conquer Strategy

Manoj kumar Jena, Rakesh Chandra Balabantaray*
Department of IT, DRIEMS, Cuttack
IIIT, Bhubaneswar
manojk_j@yahoo.co.in,
rakesh_b_ray@yahoo.co.in

Abstract: The traditional oriya text chunking approach identifies phrase structure or local word group by using only one model and phrases with the same types of features. 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 chunking quite challenging It has been shown that the limitations of using only one model are that: the use of the same types of features is not suitable for all phrases.. In this paper, the divide-conquer approach is proposed and applied in the identification of phrases or local word group. This strategy divides the task of chunking into several sub-tasks according to sensitive features of each phrase and identifies different phrases in parallel. Then, a two-stage decreasing conflict strategy is used to synthesize each sub-task's answer We argue that we might not need an explicit intermediate POS-tagging step for parsing when a sufficient amount of training material is available and word form information is used for low-frequency words. By applying and testing the approach on the public training and test corpus, the F score for arbitrary phrases identification using divide-conquer strategy achieves 91.3% compared to the previous best F score of 92.18%.

Keywords: feature structure, chunking, local word grouping, parsing(LWG), free word order languages, part-of-speech tagging, morphological analysis sensitive features; divide-conquer strategy

I. INTRODUCTION

Chunking or *local word grouping* refers to the task of recovering limited amount of syntactic information sentences from sentences or identification of syntactically correlated parts of words in a sentence and is usually the first step towards parsing of a natural language sentence [1].Although a lot of work has undergone for developing full syntactic parsers, in recent times emphasis is given to partial parsing or identification of phrase. Tong Zhang, Fred Damerau and David Johnson (2002) introduced the Winnow method that is also a machine learning method for English text chunking [4]. The advantage of this particular method is that it can isolate related features within a large quantity of features. However, this method has a lower searching efficiency and needs large memory because of the large quantity of features employed. At the same time, data sparseness also occurs as it applies the feature of "word" to all phrases. However, the problem of chunking takes a new dimension for free-word order languages, where the internal structure of the chunks or local word groups (LWG) is relatively unrestricted. Often sentence level clues or constraints become necessary for identification of the LWGs, but at the same time, a robust and efficient chunker for a free-word order language can drastically reduce the complexity of the parser. We propose here a computational framework for chunking of free-word order languages and describe the implementation of a Oriya chunker in this framework. Due to the lack of machine-readable linguistic resources in Oriya (tagged and chunked corpus to be more precise), the rules for the system have been manually designed. However, given a sufficiently large tagged corpus, where the LWG boundaries are marked, the rules can be statistically learnt as well.

The approach described here is motivated by the [Liang 2006] and local word grouping [6].This strategy remedies the shortcomings in using only one model to identify multiple types of phrases and also has several advantages:

(1) It applies the theory of divide-conquer into the field of Natural Language Processing and concentrate on the characteristics of each phrase. This focus does not occur when only one model is used to identify multiple types of phrases;

(2) Different models and sensitive features are used to identify different phrase, so this not only avoids data sparseness but also improves the speed and performance of chunking.

There are also sentence level constraints over the local word groups, which can identify ill-formed groups or errors in the POS tags of the words. This paper is organized into six sections. Section 2 defines and analyses the problem and gives a brief introduction to local word grouping and the divide conquer strategy in the context of chunking and parsing of free-word order languages. Section 3 describes the structure of phrases and the model used for developing chunker. Section 4 elucidates the
algorithm for grouping of words followed by a section on the implementation details and the results observed. The concluding section summarizes the paper and describes how this method can be modified for development of a parser for free-word order languages. In this paper, Bengali script has been written in italicized Roman fonts following the ITRANS notation [Chopde, 2002].

II. ANALYSIS OF THE PROBLEM

Text chunking has been defined as the process forming groups of words based on local information

[Bharati *et al*, 1995]. Abney [1991] viewed chunks or local word groups to be connected sub graphs of the parse tree of a sentence. The distinction between function and content words were used to design a chunker based on non-deterministic LR – parsing. Recent techniques in chunking use statistical and machine learning approaches. Indo-Aryan languages being relatively free word ordered are difficult to tackle using a generative grammar approach. Moreover, unavailability of chunked corpora precludes the use of available statistical approaches. The problem of dividing an Oriya sentence into word groups has been explored by Bharati *et al* [1995]. The output of the grouper served as an input to a computational Paninian parser. However, in their work, they made a distinction between local word groups and phrases. They assert "from a computational point, the recognition of noun phrases and verb phrases is neither simple nor efficient". Therefore, the scope of the grouper was limited and much of the
disambiguation task was left to the parser.

Structure of Phrases and LWG in Oriya

The word order of oriya is not as rigid as English. To illustrate this, consider the following sentence.

S*ehi nali mandira samnare mu gotie dhala bagha ku dekhilli*

that red temple[of] in front [first person] one white tiger saw

(I saw a white tiger in front of that red temple.)

Table 1 shows the different ways in which one can permute the words of this sentence without changing the meaning. However, not all permutations are allowed. For example, the sentence

She*i mandira nali samnare mu dhala gotie bagha ku dekhilli*

That temple red in front first person] white one tiger's saw .

(I saw that red horse in front of a white church.)

This permutations of the sentence is ungrammatical and Some of them also conveys a different sense.This can be explained by the fact that certain groups of words have fixed order, which we define as the local word groups. The groups can be permuted without restriction, but the words within a group must occur contiguously. In Table 1, the fragments that occur in a chunk are the LWGs. LWGs illustrated in the table, however allow certain degree of permutation of the words within themselves. For example, the LWG "gotie dhala bagha" can also be stated as "dhala bagha gotie" without changing the sense. However, the sub-group "mandira samnare" is completely fixed. Therefore, we define the concepts of *strong* and *weak* LWGs.

Table 1. Different arrangements of words in a sentence.

S*ehi nali mandira samnare*	Mu	*gotie dhala bagha ku* dekhili
Mu S*ehi nali mandira samnare*	Dekhili	*gotie dhala bagha ku*
gotie dhala bagha ku	Mu dekhili	S*ehi nali mandira samnare*
Mu dekhili	*gotie dhala bagha ku*	S*ehi nali mandira samnare*

Definition 1: A *strong word group* is one that has an internal rigid word order and any permutation of the constituent words either changes the sense of the group or is grammatically incorrect.

Definition 2: A *weak word group* is composed of more than one *strong word groups*, and there is negligible change in the sense of the word group, when the individual strong groups are permuted among themselves, but the constituent strong groups may not be placed beyond the weak group boundary.

It may be mentioned here that in [Bharati *et al*, 1995] only strong groups have been identified, where as in [Ray *et al*, 2003] some of the weak groups were also considered. The identification of weak word groups is a step ahead towards parsing as they capture phrases and sometimes even clauses in a sentence. An example sentence that portrays both *strong* and *weak* word groups can be like:

{ (banare) (jau jau) } {(gotie) (chhota) (nadi) } (mo) (akhire padila) .

forest through go[participle] go[participle] one small brook [1st person]eyes[in] fell .

(While travelling through the forest I noticed a small brook .)

Here, (banare) and (jau jau) form two strong word groups that in turn form a larger weak group, which can be considered to be a adverbial phrase..Another syntactic feature of Oriya is the deletion of be-verbs in present tense. This is also referred to as *hidden copula*. For example, in the sentence *ehi bahita mora*. (This book is mine.) the 'is' has been dropped. Here, the groups are *ehi bahita* and *mora*. However, a sentence like "*ehi bahita mora bhari bhAla*" (this book of mine is very nice), is comprised of two weak groups – *ehi bahita mora* and *bhari bhAla*. This clearly illustrates the need for sentence level constraints during local word grouping of free word order languages.

III. DIVIDE-CONQUER STRATEGY FOR ORIYA TEXT CHUNKING

A. The Model of OriyaText Chunking Based on Divide-conquer Strategy

Divide-conquer strategy solves a problem in a decomposing way. The procedure of getting the answer to a question is divided into three parts:

(1) decomposing the task: The task is to be decomposed into smaller sub-tasks i.e sentence

(2) Getting the answer to a sub-task to get the answer for each sub-task separately.

(3) Synthesizing each sub-task's answer: Integrating each sub-task's answer to get the final answer.

In this paper, Oriya text chunking using divide-conquer strategy is proposed and sensitive features of each phrase are considered. The architecture of Oriya text chunking using divide-conquer strategy is shown in Figure 1.[7]

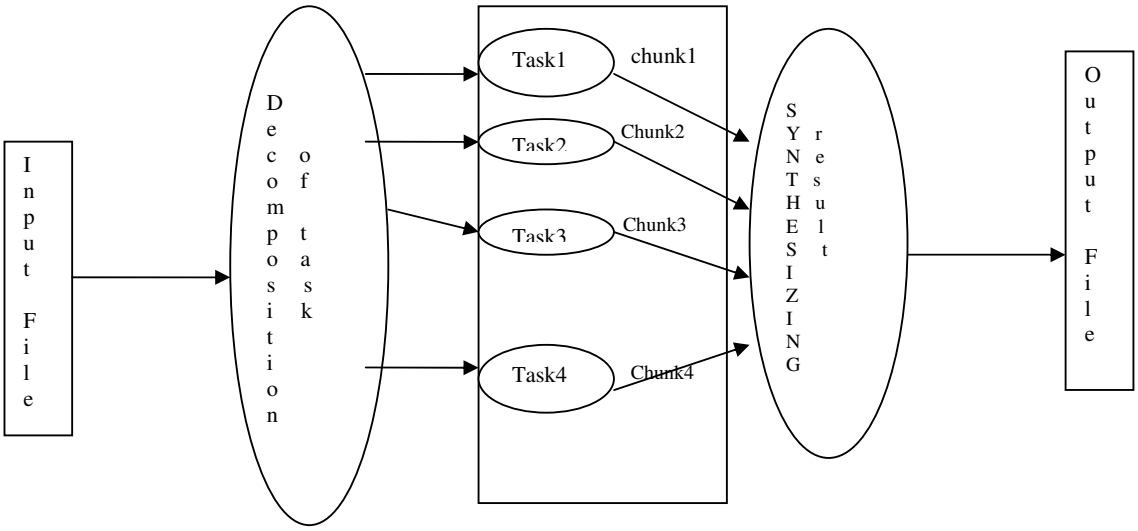


Figure 1. The data flow diagram of Oriya text chunking using divide-conquer strategy

Above partition is not exclusive. From Figure 1, we can find that chunking task is divided into 4 sub-tasks. Then each sub-task's answer is to be achieved separately. The right part in Figure 1 is to synthesize each sub-task's answer.

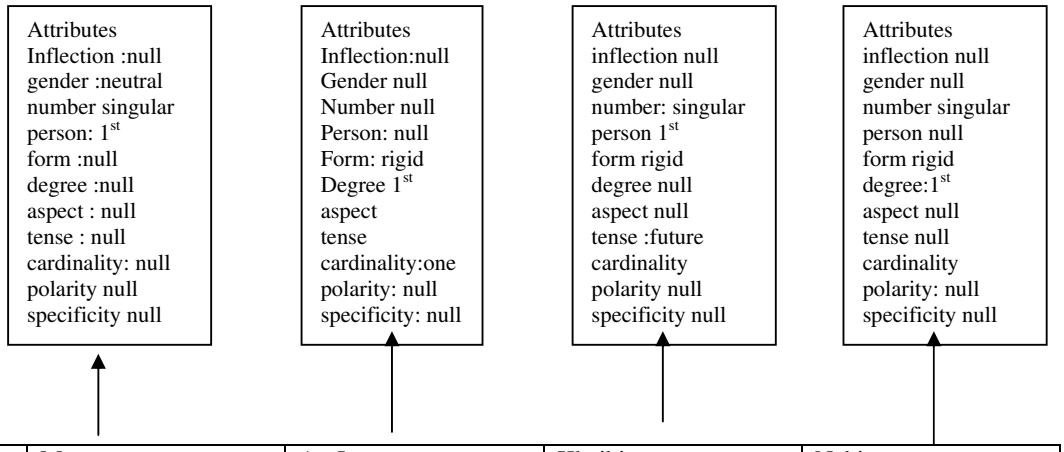
Table 2 provides the brief introduction of the 4 sub tasks.

Sub Task	**Algorithm**	**Function**
Task1	Combination of boundary statistic and rule revise	To identify NP
Task2	Longest string matching	To identify VP
Task3	Binary search and longest string matching	To identify PP,ADVP,ADJP
Task4	Binary search	To identify SBAR,CONJP,NPL

B.The feature addition based on grammatical role of phrases

In our system, 7 types of phrases are identified and the Table 3 give a brief outline of the feature added for each chunk /phrase/pos. The feature added in case of the each chunk is demonstrated thorough the following example.

The example sentence in Oriya "Mu au jor khaibi nahi."Figure 2 gives the outline of the task.



Sentence	Mu	Au Jor	Khaibi	Nahi
Chunks	NP	ADVP	VP	NPL(negative partciples)
Transliteration	I	Further More	Eat	No
Translation	Further, I will not eat more.			

Figure 2: The feature structure and it's corresponding attributes

In the course of synthesizing each sub-task's answer, the above features are added the individual phrases based on grammatical categories and relationships. From the synthesis of the result one can resolve any conflict arising out of giving the feature to each syntactic category. The identification of weak and strong word group is done through the feature selection from each pos.

IV. IMPLEMENTATION AND RESULTS

A chunker for oriya based on the model described here has been implemented. The inherent object oriented nature of the feature structures has prompted us to implement the chunker in Java. The chunker also depends on a morphological analyzer (MA) and a part-of-speech (POS) tagger, which have not been described in this paper. However for the sake of completeness, a few words must be mentioned about them. Both of these tools have been developed in house. The POS tagger has been designed with divide and conquer technique. Its accuracy is around 80%. The MA can provide information about the inflections, gender, number, person, tense, aspect, polarity, specificity etc. for a word. It has been implemented using Directed Acyclic Word Graph methods. The accuracy of the MA is around 95%.

The salient features of the chunker for Oriya have been summarized below:

• The POS tags a sentence using 8 tags, which are considered as the basic POS categories of the feature structures.

• The rules for instantiation of the feature structures are specified separately in a file. For a given POS category, the file contains information about its left and right hands side probable data. There are around 25 such rules.

• The implemented chunker can form strong word groups as well as weak ones. But it fails to recognize multi-word expressions.

• Multiple word groupings for the same sentence are also recognized by the chunker.

• Errors in POS tags by the POS tagger are identified to certain extent using local and global constraint checks, but rectification of those errors though active interaction with the POS tagger has not yet been implemented.

• The chunker cannot handle cases where semantic issues are involved.

The chunker has been tested on a set of 50 randomly selected sentences. The strong word groups are identified with 90 % accuracy provided the POS tags are correct. Weak groups are identified with around 80% accuracy for correct POS tags. For a few cases, it could also identify the errors in the POS tags. A detailed testing of the system is underway.

V. CONCLUSION

In this paper, we have described a new model for chunking of free word order languages based on devide and conquer strategy implemented . We defined the concepts of strong and weak local word groups. Weak local word groups may correspond to phrases or clauses in a sentence and therefore identification of such groups can be considered to be a significant step towards parsing and hence the system can be referred to as a shallow parser. This formalism can be extended to complete parsing of free word order languages. The idea is similar to that of karaka or traditional valencies of verbs. Future research can also focus on extending the concept of affinity from strong, weak or null to a probability value between 0 and 1, where 0 denotes no affinity and 1 denotes a strong affinity. Weak affinities can be something around 0.5. These affinities can be learnt from a chunked corpus.

VI.REFERENCES

[1] Abney. S., Parsing by chunks, Principle Based Parsing, Berwick, Abney and Tenny (eds.), Kluwer A. Publishers, 1991.

[2] Abney. S., Partial parsing via finite-state cascades, Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information, conference, Prague, Czech Republic, pp. 8-15,1996.

[3] Tjong Kim Sang. E.F. Introduction to the CoNLL-2000 Shared Task: Chunking, Proceedings of CoNLL-2000 and LLL-2000, conference, Lisbon, Portugal, pp. 127-132, 2000.

[4] Zhang T., Damerau F and Johnson D., Text Chunking based on a Generalization of Winnow, Machine Learning Research, Vol. 2,No.2, pp. 615-637, March 2002.

[5] A. Bharati, V.Chaitanya, and R.Sangal. *Natural Language Processing A PaninianPerspective*; Prentice Hall India, 1995.

[6] P.R. Ray, V. Harish, S. Sarkar, and A. Basu. *Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi*, Proceedings of International Conference on Natural Language Processing (ICON 2003), Mysore 2003.

[7] Liang .Y,Wang .N,Su .J,Ren.H, Devide- Conquer Strategy for English Text chunking, Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006

[8] Mohapatra, Pandit and Das “Sarbasara Byakarana” Student Book Store, Cuttack.

[9] Sarangi Nrusingha,”Bruhat Oriya Byakarana” Satyanarayana Book Store Binod Bihari, Cuttack.

[10] A. Chopde. *ITRANS*, version 5.30, July 2002, http://www.aczone.com/itrans/