Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 84 (2016) 65 – 69

Procedia Computer Science

7th International conference on Intelligent Human Computer Interaction, IHCI 2015

Phrase and Idiom Identification in Assamese

Shinjit Kamal Borah[a*], Utpal Sharma[a]

[a]*Department of CSE,TezpurUniversity, Napaam-784 028, India*

Abstract

Identification of phrases and idioms is an indispensable part of computational linguistics work. In case of Assamese, this is a challenging topic mainly because of the cases and affixes used in the language. Though, this language is an Eastern Indo-Aryan language spoken by around 30 million people, this topic has not been studied much, as very little computational linguistics work has been done for this language. Assamese language is a relatively free word order language. Context Free Grammar (CFG) can be applied in phrase level by taking extra care in defining the production rules. In this paper, we explain about a method which can be considered as modified context free grammar. Different production rules for phrases can be defined using this modified context free grammar. In this method, the right hand side of the production rules is treated as a free string. So that free word order phenomenon can be dealt with. Different idioms are also analyzed in terms of their syntax and use, to find out the similarities among them to build a dictionary of idioms. Difficulties in parsing phrases and idioms are also discussed and some of the techniques are also provided to overcome those difficulties.

Peer-review under responsibility of the Organizing Committee of IHCI 2015

*Keywords:*Phrase; Idiom; Assamese; Context free grammar; Computational linguistics.

1. Introduction

Phrase is defined as a group of words that does not contain a verb and its subject together, and is used as a single part of speech. A phrase consists of two parts head and dependents. The head gives the name of the phrase category[1]. Thus, the head of a noun phrase is a noun. Dependents may act as a modifiers or complements. For example ‘The house at the end of the street is red.’ The words ‘The house at the end of the street’ form a phrase; together they act like a noun. Similarly in Assamese, ‘চাকৰি বিছাৰি হাবাথুৰি খোৱা ল’ৰাজন আজি আমাৰ ঘৰলৈ আহিছিল। (cAkari bichAri HAbAthuri khowA l’rAjan Aji AmAr gharalE Ahichil)’ In this sentence ‘চাকৰি বিছাৰি হাবাথুৰি খোৱা ল’ৰাজন (cAkari

* Corresponding author. Tel.: +919854446484.
E-mail address:borahshinjit@gmail.com

Peer-review under responsibility of the Organizing Committee of IHCI 2015
doi:10.1016/j.procs.2016.04.067

bichAri HAbAthuri khowA l'rAjan)' is a noun phrase due to the head noun 'ল'ৰা (l'rA)'. There are five major phrase types[2] as follows

- Adverb Phrase (AdvP)
- Preposition Phrase/ Postposition Phrase (PP)
- Adjective Phrase (AP)
- Noun Phrase (NP)
- Verb Phrase (VP)

An idiom is a group of words having a special meaning due to its common usage. An idiom is an expression whose meaning cannot be determined simply from the meaning of its component words and their syntax[1]. For example 'Fred kicked the bucket'. In this sentence 'kicked the bucket' is an idiom. Actual meaning of this sentence is 'Fred is understood to have died'. In Assamese, we can consider the sentence 'ৰামৰ কপাল ফুলিল। (rAmar kapAl phulil)', in this sentence 'কপাল ফুলিল (kapAl phulil)' is an idiom and meaning is 'blessed with fortune'. Though 'কপাল (kapAl)' means forehead and 'ফুলিল (phulil)' means blossom. Idioms can be categorized into six different types[2] as follows

- Nominal Compound Roots
- Idiomatic uses of Nouns
- Idiomatic uses of Verbs
- Idiomatic uses of Verbs with particular Nouns
- Idiomatic uses of Adjectives
- Idiomatic uses of miscellaneous Words or Phrases

Identification of phrases is very important because phrases act as a single constituent and their syntax may be different for different languages. For example, modifiers of nouns appear before the noun in English but in case of Assamese, they appear after the noun. We can consider the phrase 'on the table'. Here the head is table, which is a noun and appears after the dependents. If we translate the phrase into Assamese, it will be 'মেজ খনৰ ওপৰত (mez khanar oparat)'. In this phrase, head noun is 'মেজ (mez)' and dependents follow the head. Similarly, if the idioms are not identified and handled properly, it may mislead the process of language analysis and language understanding due to their non-compositional meaning.

Assamese is an Eastern Indo-Aryan language spoken by around 30 million people[3]. Assamese evolved at least before 7th century A.D. from the Magadhi Prakrit, which is developed from a dialect or a group of dialects that are close to, but different from Vedic and Classical Sanskrit. Not much computational linguistic work has been done for Assamese so far. We have not come across any work like the one discussed here.

2. Related work

In English, phrases are generally parsed using context free grammar (CFG) [4]. A context free grammar consists of a set of rules or productions. Each of the productions of CFG expresses the ways that symbols of the language can be grouped and ordered together[5]. In the paper[6], the authors showed how idioms can be parsed in lexicalized Tree Adjoining Grammars (TAGs). They considered idioms of different syntactic categories in both English and French. Lin (1999) presented a method for automatic identification of non-compositional expressions using their statistical properties in a text corpus[7]. There are many techniques for parsing phrases and idioms in languages, particularly in English. But, most of the techniques are not effective for Assamese. A few of these techniques can be applied to Assamese only with some modification.