Abstract

Identification of phrases and idiomsis an indispensable part of computational linguistics work. Phrase may refer to any group of words functioning as a single constituent within a sentence. The syntax of a phrase may be different for different languages. Within a language too, phrases may take different form. Idioms are group of words having a special meaning, which is not the compositional meaning of the words present in the group (or phrase). If the idioms are not identified and handled properly, it may mislead the process of language translation and language understanding. This is a challenging topic mainly because of the cases and affixes used in Assamese language. Though, Assamese is an Eastern Indo-Aryan language spoken by around 30 million people, this topic has not been studied much in this language, 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 report, different production rules are defined using modified context free grammar. In this modified context free grammar, the right hand side of the production rules is treated as a free string. So that free word order phenomenon can be dealt with. These production rules are used to parse and identify the phrases. And finally, we have developed a parser to parse Assamese phrases. 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 solution techniques are also provided to overcome those difficulties.

Keywords: Phrase, Idiom, Assamese, Context free grammar, Computational linguistics.