Chapter 44
Parsing Bangla Grammar Using Context Free Grammar

Al-Mahmud
Khulna University of Engineering and Technology, Bangladesh

Bishnu Sarker
Khulna University of Engineering and Technology, Bangladesh

K. M. Azharul Hasan
Khulna University of Engineering and Technology, Bangladesh

ABSTRACT

Parsing plays a very prominent role in computational linguistics. Parsing a Bangla sentence is a primary need in Bangla language processing. This chapter describes the Context Free Grammar (CFG) for parsing Bangla language, and hence, a Bangla parser is proposed based on the Bangla grammar. This approach is very simple to apply in Bangla sentences, and the method is well accepted for parsing grammar. This chapter introduces a parser for Bangla language, which is, by nature, a predictive parser, and the parse table is constructed for recognizing Bangla grammar. Parse table is an important tool to recognize syntactical mistakes of Bangla sentences when there is no entry for a terminal in the parse table. If a natural language can be successfully parsed then grammar checking of this language becomes possible. The parsing scheme in this chapter works based on a top-down parsing method. CFG suffers from a major problem called left recursion. The technique of left factoring is applied to avoid the problem.

INTRODUCTION

Language plays the most important role in human communication. Human communication is based on exchange of feelings, sharing of knowledge, etc. Feelings can be exchanged through voice, symbols and signs. Language provides the most convenient way of expressing the expressions of feelings through providing those necessary phonetics that could be spoken and the symbols that could be written to be preserved. Language is the driving force to human communication. Language is the most important medium to represent and express human knowledge and human communication.

Bangla (or Bengali) is one of the more important Indo-Iranian languages, is the sixth most

DOI: 10.4018/978-1-4666-6042-7.ch044

popular in the world and spoken by a population that now exceeds 250 million. Geographically, Bangla-speaking population percentages are as follows: Bangladesh (over 95%), and the Indian States of Andaman & Nicobar Islands (26%), Assam (28%), Tripura (67%), and West Bengal (85%). The global total includes those who are now in diaspora in Canada, Malawi, Nepal, Pakistan, Saudi Arabia, Singapore, United Arab Emirates, United Kingdom, and United States.

Bangla is still in degraded stage at least as far as work in the area of computational linguistics is concerned. Natural languages like English and even Hindi is rapidly progressing as far as work done in processing by computers is concerned. Unfortunately, Bangla lags more or less behind in some crucial areas of research like parts of speech tagging, text summarization and categorization, information retrieval and most importantly in the area of grammar checking. The grammar checking for a language has a wide variety of applications.

The activity of breaking down a sentence into its constituent parts is known as parsing. Parsing is an earlier term for the diagramming of sentences of natural languages, and is still used for the diagramming. Parsing a sentence involves the use of linguistic knowledge of a language to discover the way in which a sentence is structured. Exactly how this linguistic knowledge is represented and can be used to understand sentences is one of the questions that has engaged the interest of psycholinguists, linguists, computational linguists, and computer scientists. Bangla parsing is a challenging task. This chapter has a detail discussion over Bangla parsing using Context Free Grammar.

BACKGROUND

In computing, a parser is one of the components in an interpreter or compiler that checks for correct syntax and builds a data structure (often some kind of parse tree, abstract syntax tree or other hierarchical structure) implicitly in the input tokens. Parsing can be defined as a method where a parser algorithm is used to determine whether a given input string is grammatically correct or not for a given grammar. Parsing is a fundamental problem in language processing for both machines and humans. In general, the parsing problem includes the definition of an algorithm to map any input sentence to its associated syntactic tree structure (Saha, 2006). The parser often uses a separate lexical analyzer to create tokens from the sequence of input characters. Parsers may be programmed by hand or may be automatically or semi-automatically generated (in some programming languages) by a tool.

A parse tree for a grammar is a tree where the root of the tree is the start symbol for the grammar, the interior nodes are the non-terminals of the grammar, the leaf nodes are the terminals of the grammar and the children of a node starting from the left to the right correspond to the symbols on the right hand side of some production for the node in the grammar. Every valid parse tree represents a string generated by the grammar (Yarowsky, 1995).

A parser analyzes the sequence of symbols presented to it based on the grammar (Yarowsky, 1995). Natural language applications namely Information Extraction, Machine Translation, and Speech Recognition, need to have an accurate parser (Haque & Khan, 2005). Parsing natural language text is more difficult than the computer languages such as compiler and word processor because the grammars for natural languages are complex, ambiguous, and infinite in number of vocabulary. It is difficult to prepare formal rules to describe informal behavior even though it is clear that some rules are being followed. For a syntax based grammar checking the sentence is completely parsed to check the correctness of it. If the syntactic parsing fails, the text is considered incorrect. On the other hand, for statistics based approach, Parts of Speech (POS) tag sequences are prepared from an annotated corpus, and hence the frequency and the probability (Sengupta &