Parsing

Lecture 5

Robb T. Koether

Hampden-Sydney College

Fri, Jan 30, 2009

Outline

Syntax Analysis

Definition (Context-Free Grammar)

A context-free grammar is a grammar in which each grammar rule has the form

$$A \to \alpha$$

where A is a nonterminal and α is a string of terminals and nonterminals, i.e.,

$$\alpha \in (\Sigma \cup N)^*.$$

- The syntax of a language is described by a context-free grammar.

Grammar Convention

Grammar Convention

- Terminals are represented by lowercase letters near the beginning of the alphabet (a, b, c, ...) or by non-alphabetic characters (2, +, {, ...).
- Nonterminals are represented by uppercase letters near the beginning of the alphabet (A, B, C, ...).

Grammar Convention

Grammar Convention

- Grammar symbols that may be either terminals or nonterminals are represented by uppercase letters near the end of the alphabet (Z, Y, X, ...).
- Strings of terminals are represented by lowercase letters near the end of the alphabet (z, y, x, ...).

Grammar Convention

Grammar Convention

- Strings of grammar symbols (both terminals and nonterminals) are represented by lowercase Greek letters (α, β, γ, ...).

Parsers

Definition (Parser)

A parser is a program that matches a sequence of tokens to the grammar rules of a context-free grammar, thereby building the parse tree for the sequence.

- The parser will receive, as its input, tokens from a lexer,
- The parser will output an abstract syntax tree (parse tree) representing the grammatical structure of the input.

Parsing Algorithms

Definition

A top-down parser begins with the start symbol and applies productions until only terminals remain.

Definition

A bottom-up parser begins by matching terminals to productions and continues until the string of input symbols is reduced to the start symbol.

Parsing Algorithms

- Top-down parsers traverse the parse tree from the top down.
- They tend to be simpler, but less powerful.
- Bottom-up parsers traverse the parse tree from the bottom up.
- They tend to be complex, but more powerful.

Parsing Algorithms

- There are two basic methods of implementing top-down parsers.
  - Recursive descent parsers (Lab 4).
  - Table-driven parsers, also called LL parsers (Lab 5).
- Bottom-up parsers are table-driven (Lab 6).
- These are called LR parsers.

LL Parsers

Definition (LL Parser)

An LL parser parses the input from left to right (L) and uses a leftmost derivation (L).

Definition (LR Parser)

An LR parser parses the input from left to right (L) and uses a rightmost derivation (R).

LL Grammars

- The LL grammars form a proper subset of the LR grammars.
- Most modern compilers use LR parsers.
  - LR(0) (lookahead 0)
  - LR(1) (lookahead 1)
  - LALR (lookahead LR)
- CUP uses LALR.

Parsing Algorithms

- Although programming languages are described by context-free grammars, none of the known parsing algorithms is capable of parsing an arbitrary context-free grammar.
- Therefore, we must use caution when writing a grammar for a language; we must be able to parse it.