Compiler Construction

Lecture 9: Syntactic Analysis V ($LR(k)$ Grammars)

Thomas Noll

Lehrstuhl für Informatik 2
(Software Modeling and Verification)

RWTH Aachen University

noll@cs.rwth-aachen.de

http://www-i2.informatik.rwth-aachen.de/i2/cc08/

Summer semester 2008

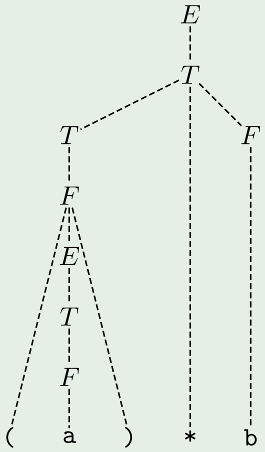
Outline

Example

Grammar for
arithmetic expressions:

$$G_{AE}: \quad \begin{array}{lll} E \to E\texttt{+}T \mid T & (1,2) \\ T \to T\texttt{*}F \mid F & (3,4) \\ F \to \texttt{(}E\texttt{)} \mid \texttt{a} \mid \texttt{b} & (5,6,7) \end{array}$$

Leftmost analysis of `(a)*b`:
2 3 4 5 2 4 6 7



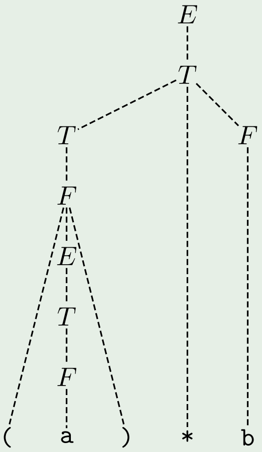
Outline

Example 9.1

Grammar for arithmetic expressions:

$$G_{AE}: \quad \begin{array}{lll} E \rightarrow E\texttt{+}T \mid T & (1,2) \\ T \rightarrow T\texttt{*}F \mid F & (3,4) \\ F \rightarrow \texttt{(}E\texttt{)} \mid \texttt{a} \mid \texttt{b} & (5,6,7) \end{array}$$

Reversed rightmost analysis of `(a)*b`:
6 4 2 5 4 7 3 2



Approach:

1. Given $G \in CFG_\Sigma$, construct a nondeterministic bottom-up parsing automaton (NBA) which accepts $L(G)$ and which additionally computes corresponding (reversed) rightmost analyses
   - input alphabet: Σ
   - pushdown alphabet: X
   - output alphabet: $[p]$ (where $p := |P|$)
   - state set: omitted
   - transitions:

     shift: shifting input symbols onto the pushdown

     reduce: replacing the right-hand side of a production by its left-hand side (= inverse expansion steps)
2. Remove nondeterminism by allowing lookahead on the input: $G \in LR(k)$ iff $L(G)$ recognizable by deterministic bottom-up parsing automaton with lookahead of k symbols

Outline

Nondeterministic Bottom-Up Automaton I

Definition 9.2 (Nondeterministic bottom-up parsing automaton)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$. The nondeterministic bottom-up parsing automaton of G, NBA(G), is defined by the following components.

- Input alphabet: Σ
- Pushdown alphabet: X
- Output alphabet: $[p]$
- Configurations: $\Sigma^* \times X^* \times [p]^*$ (top of pushdown to the right)
- Transitions for $w \in \Sigma^*$, $\alpha \in X^*$, and $z \in [p]^*$:
  shifting steps: $(aw, \alpha, z) \vdash (w, \alpha a, z)$ if $a \in \Sigma$
  reduction steps: $(w, \alpha\beta, z) \vdash (w, \alpha A, zi)$ if $\pi(i) = A \to \beta$
- Initial configuration for $w \in \Sigma^*$: $(w, \varepsilon, \varepsilon)$
- Final configurations: $\{\varepsilon\} \times \{S\} \times [p]^*$

Nondeterministic Bottom-Up Automaton II

Example 9.3

Grammar for arithmetic expressions (cf. Example 5.11):

$$
\begin{array}{llll}
G_{AE}: & E \to E\texttt{+}T \mid T & & (1,2) \\
& T \to T\texttt{*}F \mid F & & (3,4) \\
& F \to \texttt{(}E\texttt{)} \mid \texttt{a} \mid \texttt{b} & & (5,6,7)
\end{array}
$$

Bottom-up parsing of `(a)*b`:

$$
\begin{array}{ll}
 & (\texttt{(a)*b}, \varepsilon, \varepsilon) \\
\vdash & (\texttt{a)*b}, \texttt{(}, \varepsilon) \\
\vdash & (\texttt{)*b}, \texttt{(a}, \varepsilon) \\
\vdash & (\texttt{)*b}, \texttt{(}F, 6) \\
\vdash & (\texttt{)*b}, \texttt{(}T, 64) \\
\vdash & (\texttt{)*b}, \texttt{(}E, 642) \\
\vdash & (\texttt{*b}, \texttt{(}E\texttt{)}, 642) \\
\vdash & (\texttt{*b}, F, 6425) \\
\vdash & (\texttt{*b}, T, 64254) \\
\vdash & (\texttt{b}, T\texttt{*}, 64254) \\
\vdash & (\varepsilon, T\texttt{*b}, 64254) \\
\vdash & (\varepsilon, T\texttt{*}F, 642547) \\
\vdash & (\varepsilon, T, 6425473) \\
\vdash & (\varepsilon, E, 64254732)
\end{array}
$$

Correctness of NBA(G)

Theorem 9.4 (Correctness of NBA(G))

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ *and* NBA(G) *as before. Then, for every* $w \in \Sigma^*$ *and* $z \in [p]^*$,

$$(w, \varepsilon, \varepsilon) \vdash^* (\varepsilon, S, z) \quad \textit{iff} \quad \overleftarrow{z} \textit{ is a rightmost analysis of } w$$

Proof.

similar to the top-down case (Theorem 5.15) □

Nondeterminisn in NTA(G)

Remark: NTA(G) is generally nondeterministic

- Shift or reduce? Example:

$$(bw, \alpha a, z) \vdash \begin{cases} (w, \alpha ab, z) \\ (bw, \alpha A, zi) \quad \text{if } \pi(i) = A \to a \end{cases}$$

- If reduce: which "handle" β? Example:

$$(w, \alpha ab, z) \vdash \begin{cases} (w, \alpha A, zi) \\ (w, \alpha aB, zj) \end{cases} \text{if } \pi(i) = A \to ab \text{ and } \pi(j) = B \to b$$

- If reduce β: which left-hand side A? Example:

$$(w, \alpha a, z) \vdash \begin{cases} (w, \alpha A, zi) \\ (w, \alpha B, zj) \end{cases} \text{if } \pi(i) = A \to a \text{ and } \pi(j) = B \to a$$

- When to terminate parsing? Example:

$$\underbrace{(\varepsilon, S, z)}_{\text{final}} \vdash (\varepsilon, A, zi) \text{ if } \pi(i) = A \to S$$

Resolving Termination Nondeterminism I

General assumption in the following: every grammar is start separated

Definition 9.5 (Start separation)

A grammar $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ is called start separated if S only occurs in productions of the form $S \to A$ where $A \neq S$.

Remarks:

- Start separation always possible by adding $S' \to S$ with new start symbol S'
- From now on consider only reduced grammars of this form ($\pi(0) = S' \to S$)

Resolving Termination Nondeterminism II

Start separation removes last form of nondeterminism (when to terminate parsing):

Corollary 9.6

If $G \in CFG_\Sigma$ is start separated, then no successor of a final configuration (ε, S', z) in $\mathrm{NBA}(G)$ is again a final configuration. (Thus parsing should be stopped in the first final configuration.)

Corollary 9.7

- *To (ε, S', z), only reductions by ε-productions can be applied:*
  $$(\varepsilon, S', z) \vdash (\varepsilon, S'A, zi) \quad \text{if } \pi(i) = A \to \varepsilon$$
- *Thereafter, only reductions by productions of the form $A_0 \to A_1 \ldots A_n$ $(n \geq 0)$ can be applied*
- *Every resulting configuration is of the (non-final) form*
  $$(\varepsilon, S'B_1 \ldots B_k, z) \quad \text{where } k \geq 1$$

Outline

$LR(k)$ Grammars

Goal: resolve remaining nondeterminism of NBA(G) by supporting lookahead of $k \in \mathbb{N}$ symbols on the input

$\Longrightarrow$ $LR(k)$: reading of input from left to right with k-lookahead, computing a rightmost analysis

Definition 9.8 ($LR(k)$ grammar)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ be start separated and $k \in \mathbb{N}$. Then G has the $LR(k)$ property (notation: $G \in LR(k)$) if for all rightmost derivations of the form

$$S \begin{cases} \Rightarrow_r^* & \alpha A w & \Rightarrow_r \alpha\beta w \\ \Rightarrow_r^* & \alpha' A' w' & \Rightarrow_r \alpha\beta v \end{cases}$$

such that $\text{first}_k(w) = \text{first}_k(v)$, it follows that $\alpha' = \alpha$, $A' = A$, and $w' = v$.

Remarks:

- If $G \in LR(k)$, then the reduction of $\alpha\beta w$ to $\alpha A w$ is already determined by $\alpha\beta\,\text{first}_k(w)$.
- Therefore NBA(G) in configuration $(w, \alpha\beta, z)$ can decide whether to shift or to reduce and, in the second case, how to reduce.

Outline

$LR(0)$ Grammars

The case $k = 0$ is relevant (in contrast to $LL(0)$): here the decision is just based on the contents of the pushdown, without any lookahead.

Corollary 9.9 ($LR(0)$ grammar)

$G \in CFG_\Sigma$ has the $LR(0)$ property if for all rightmost derivations of the form

$$S \begin{cases} \Rightarrow_r^* & \alpha A w & \Rightarrow_r \alpha\beta w \\ \Rightarrow_r^* & \alpha' A' w' & \Rightarrow_r \alpha\beta v \end{cases}$$

it follows that $\alpha' = \alpha$, $A' = A$, and $w' = v$.

Goal: derive a finite information from the pushdown which suffices to resolve the nondeterminism (similar to abstraction of right context in LL parsing by fo-sets)

$LR(0)$ Items and Sets

Definition 9.10 ($LR(0)$ items and sets)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ be start separated by $S' \to S$ and $S' \Rightarrow_r^* \alpha A w \Rightarrow_r \alpha\beta_1\beta_2 w$ (i.e., $A \to \beta_1\beta_2 \in P$).

- $[A \to \beta_1 \cdot \beta_2]$ is called an $LR(0)$ item for $\alpha\beta_1$.
- Given $\gamma \in X^*$, $LR(0)(\gamma)$ denotes the set of all $LR(0)$ items for γ, called the $LR(0)$ set (or: $LR(0)$ information) of γ.
- $LR(0)(G) := \{LR(0)(\gamma) \mid \gamma \in X^*\}$.

Corollary 9.11

1. *For every* $\gamma \in X^*$, $LR(0)(\gamma)$ *is finite.*
2. $LR(0)(G)$ *is finite.*
3. *The item* $[A \to \beta\cdot] \in LR(0)(\gamma)$ *indicates the possible reduction* $(w, \alpha\beta, z) \vdash (w, \alpha A, zi)$ *where* $\pi(i) = A \to \beta$ *and* $\gamma = \alpha\beta$.
4. *The item* $[A \to \beta_1 \cdot Y\beta_2] \in LR(0)(\gamma)$ *indicates a possible shift step (with incomplete handle* β_1*).*

$LR(0)$ Conflicts

Definition 9.12 ($LR(0)$ conflicts)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ and $I \in LR(0)(G)$.

- I has a shift/reduce conflict if there exist $A \to \alpha_1 a \alpha_2, B \to \beta \in P$ such that
$$[A \to \alpha_1 \cdot a\alpha_2], [B \to \beta \cdot] \in I.$$

- I has a reduce/reduce conflict if there exist $A \to \alpha, B \to \beta \in P$ with $A \neq B$ or $\alpha \neq \beta$ such that
$$[A \to \alpha \cdot], [B \to \beta \cdot] \in I.$$

Lemma 9.13

$G \in LR(0)$ iff no $I \in LR(0)(G)$ contains conflicting items.

Computing $LR(0)$ Sets I

Theorem 9.14 (Computing $LR(0)$ sets)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$ be start separated by $S' \to S$ and reduced.

1. *$LR(0)(\varepsilon)$ is the least set such that*
   - *$[S' \to \cdot S] \in LR(0)(\varepsilon)$ and*
   - *if $[A \to \cdot B\gamma] \in LR(0)(\varepsilon)$ and $B \to \beta \in P$, then $[B \to \cdot\beta] \in LR(0)(\varepsilon)$.*
2. *$LR(0)(\alpha Y)$ ($\alpha \in X^*, Y \in X$) is the least set such that*
   - *if $[A \to \gamma_1 \cdot Y\gamma_2] \in LR(0)(\alpha)$, then $[A \to \gamma_1 Y \cdot \gamma_2] \in LR(0)(\alpha Y)$ and*
   - *if $[A \to \gamma_1 \cdot B\gamma_2] \in LR(0)(\alpha Y)$ and $B \to \beta \in P$, then $[B \to \cdot\beta] \in LR(0)(\alpha Y)$.*

Computing $LR(0)$ Sets II

Example 9.15

$$G: \quad \begin{aligned} S' &\to S \\ S &\to B \mid C \\ B &\to aB \mid b \\ C &\to aC \mid c \end{aligned}$$

$[S' \to \cdot S] \in$

$LR(0)(\varepsilon)$ $\quad [A \to \cdot B\gamma] \in LR(0)(\varepsilon), B \to \beta \in P \implies [B \to \cdot\beta] \in LR(0)(\varepsilon)$ $\quad [A \to \gamma_1 \cdot Y\gamma_2] \in LR(0)(\alpha) \implies [A \to \gamma_1 Y \cdot \gamma_2] \in LR(0)(\alpha Y)$

$I_0 := LR(0)(\varepsilon)$: $[S' \to \cdot S]$ $[S \to \cdot B]$ $[S \to \cdot C]$ $[B \to \cdot aB]$ $[B \to \cdot b]$ $[C \to \cdot aC]$ $[C \to \cdot c]$

$I_1 := LR(0)(S)$: $[S' \to S\cdot]$

$I_2 := LR(0)(B)$: $[S \to B\cdot]$

$I_3 := LR(0)(C)$: $[S \to C\cdot]$

$I_4 := LR(0)(a)$: $[B \to a \cdot B]$ $[C \to a \cdot C]$ $[B \to \cdot aB]$ $[B \to \cdot b]$ $[C \to \cdot aC]$ $[C \to \cdot c]$

$I_5 := LR(0)(b)$: $[B \to b\cdot]$

$I_6 := LR(0)(c)$: $[C \to c\cdot]$

$I_7 := LR(0)(aB)$: $[B \to aB\cdot]$

$I_8 := LR(0)(aC)$: $[C \to aC\cdot]$

$(LR(0)(aa) = LR(0)(a) = I_4,\ LR(0)(ab) = LR(0)(b) = I_5,$
$LR(0)(ac) = LR(0)(c) = I_6,\ I_9 := LR(0)(\gamma) = \emptyset$ in all remaining cases)