Compiler Construction

Lecture 7: Syntactic Analysis III ($LL(1)$ 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

Adding Lookahead

Goal: resolve nondeterminism of NTA(G) by supporting lookahead of $k \in \mathbb{N}$ symbols on the input
$\Longrightarrow$ determination of expanding A-production by next k symbols

Definition (first$_k$ set)

Let $G = \langle N, \Sigma, P, S\rangle \in CFG_\Sigma$, $\alpha \in X^*$, and $k \in \mathbb{N}$. Then the first$_k$ set of α, $\text{first}_k(\alpha) \subseteq \Sigma^*$, is given by

$$\begin{aligned}\text{first}_k(\alpha) := {} & \{v \in \Sigma^k \mid \text{ex. } w \in \Sigma^* \text{ such that } \alpha \Rightarrow^* vw\} \cup \\ & \{v \in \Sigma^{<k} \mid \alpha \Rightarrow^* v\}\end{aligned}$$

$LL(k)$ Grammars I

$LL(k)$: reading of input from left to right with k-lookahead, computing a leftmost analysis

Definition ($LL(k)$ grammar)

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

$$S \Rightarrow_l^* wA\alpha \begin{cases} \Rightarrow_l w\beta\alpha \Rightarrow_l^* wx \\ \Rightarrow_l w\gamma\alpha \Rightarrow_l^* wy \end{cases}$$

such that $\text{first}_k(x) = \text{first}_k(y)$, it follows that $\beta = \gamma$ (i.e., the same production is applied to A).

Remarks:

- If $G \in LL(k)$, then the leftmost derivation step for $wA\alpha$ in the above diagram is determined by the next k symbols following w.
- Problem: how to determine the A-production from the lookahead (potentially infinitely many derivations to wx/wy)?

Lemma (Characterization of $LL(k)$)

$G \in LL(k)$ iff for all leftmost derivations of the form

$$S \Rightarrow_l^* wA\alpha \begin{cases} \Rightarrow_l w\beta\alpha \\ \Rightarrow_l w\gamma\alpha \end{cases}$$

such that $\beta \neq \gamma$, it follows that $\mathrm{first}_k(\beta\alpha) \cap \mathrm{first}_k(\gamma\alpha) = \emptyset$.

Remarks:

- If $G \in LL(k)$, then the A-production is determined by the lookahead sets $\mathrm{first}_k(\beta\alpha)$ (for every $A \to \beta \in P$).
- Problem: still infinitely many rightmost contexts α to be considered (if β "too short", i.e., $\mathrm{first}_k(\beta\alpha) \neq \mathrm{first}_k(\beta)$).

Outline

Goal: determine all possible lookaheads from production alone (by combining all possible right contexts)

Definition 7.1 (follow_k set)

Let $G = \langle N, \Sigma, P, S \rangle \in CFG_\Sigma$, $A \in N$, and $k \in \mathbb{N}$. Then the follow_k set of A, $\text{follow}_k(A) \subseteq \Sigma^*$, is given by

$$\text{follow}_k(A) := \{v \in \text{first}_k(\alpha) \mid \text{ex. } w \in \Sigma^*, \alpha \in X^* \text{ such that } S \Rightarrow_l^* wA\alpha\}.$$

Outline

The Case $k = 1$

Motivation:

- $k = 1$ sufficient to resolve nondeterminism in “most” practical applications
- Implementation of $LL(k)$ parsers for $k > 1$ rather involved (cf. ANTLR [ANother Tool for Language Recognition; formerly PCCTS] at `http://www.antlr.org/`)

Abbreviations: $\mathrm{fi} := \mathrm{first}_1$, $\mathrm{fo} := \mathrm{follow}_1$, $\Sigma_\varepsilon := \Sigma \cup \{\varepsilon\}$

Corollary 7.2

1. *For every* $\alpha \in X^*$,
$$\mathrm{fi}(\alpha) = \{a \in \Sigma \mid \textit{ex. } w \in \Sigma^* : \alpha \Rightarrow^* aw\} \cup \{\varepsilon \mid \alpha \Rightarrow^* \varepsilon\} \subseteq \Sigma_\varepsilon$$
2. *For every* $A \in N$,
$$\mathrm{fo}(A) = \{x \in \mathrm{fi}(\alpha) \mid \textit{ex. } w \in \Sigma^*, \alpha \in X^* : S \Rightarrow_l^* wA\alpha\} \subseteq \Sigma_\varepsilon.$$

Lookahead Sets

Definition 7.3 (Lookahead set)

Given $\pi = A \to \beta \in P$,

$$\mathrm{la}(\pi) := \mathrm{fi}(\beta \cdot \mathrm{fo}(A)) \subseteq \Sigma_\varepsilon$$

is called the lookahead set of π (where $\mathrm{fi}(\Gamma) := \bigcup_{\gamma \in \Gamma} \mathrm{fi}(\gamma)$).

Corollary 7.4

1. *For all* $a \in \Sigma$,
$$a \in \mathrm{la}(A \to \beta) \textit{ iff } a \in \mathrm{fi}(\beta) \textit{ or } (\beta \Rightarrow^* \varepsilon \textit{ and } a \in \mathrm{fo}(A))$$
2. $\varepsilon \in \mathrm{la}(A \to \beta)$ *iff* $\beta \Rightarrow^* \varepsilon$ *and* $\varepsilon \in \mathrm{fo}(A)$

Characterization of $LL(1)$

Theorem 7.5 (Characterization of $LL(1)$)

$G \in LL(1)$ iff for all pairs of rules $A \to \beta \mid \gamma \in P$ (where $\beta \neq \gamma$):

$$\mathrm{la}(A \to \beta) \cap \mathrm{la}(A \to \gamma) = \emptyset.$$

Proof.

on the board □

Remark: the above theorem generally does not hold if $k > 1$ (cf. exercises)

Outline

Computing Lookahead Sets I

(see Waite/Goos: *Compiler Construction*, p. 164f)

Lemma 7.6 (Computation of fi/fo)

The sets $\mathrm{fi}(\alpha) \subseteq \Sigma_\varepsilon$ *(for* $\alpha \in X^*$*) and* $\mathrm{fo}(A) \subseteq \Sigma_\varepsilon$ *(for* $A \in N$*) are the least sets such that:*

1. $\mathrm{fi}(Y)$ *for* $Y \in X$*:*
   - $Y \in \Sigma \implies \mathrm{fi}(Y) = \{Y\}$
   - $Y \to A_1 \dots A_k Z\alpha \in P, k \in \mathbb{N}, Z \in X, \varepsilon \in \mathrm{fi}(A_1) \cap \dots \cap \mathrm{fi}(A_k)$, $a \in \mathrm{fi}(Z) \implies a \in \mathrm{fi}(Y)$
   - $Y \to A_1 \dots A_k \in P, k \in \mathbb{N}, \varepsilon \in \mathrm{fi}(A_1) \cap \dots \cap \mathrm{fi}(A_k) \implies \varepsilon \in \mathrm{fi}(Y)$
2. $\mathrm{fi}(Y_1 \dots Y_n)$ *for* $n \in \mathbb{N}, Y_i \in X$*:*
   - $\varepsilon \in \mathrm{fi}(Y_1 \dots Y_{k-1}), a \in \mathrm{fi}(Y_k), k \in [n] \implies a \in \mathrm{fi}(Y_1 \dots Y_n)$
   - $\varepsilon \in \mathrm{fi}(Y_1) \cap \dots \cap \mathrm{fi}(Y_n) \implies \varepsilon \in \mathrm{fi}(Y_1 \dots Y_n)$
3. $\mathrm{fo}(A)$ *for* $A \in N$*:*
   - $\varepsilon \in \mathrm{fo}(S)$
   - $A \to \alpha B\beta \in P, a \in \mathrm{fi}(\beta) \implies a \in \mathrm{fo}(B)$
   - $A \to \alpha B\beta \in P, \varepsilon \in \mathrm{fi}(\beta), x \in \mathrm{fo}(A) \implies x \in \mathrm{fo}(B)$

Computing Lookahead Sets II

Corollary 7.7

1. $A \to a\alpha \in P \implies a \in \mathrm{fi}(A)$
2. $A \to B\alpha \in P, a \in \mathrm{fi}(B) \implies a \in \mathrm{fi}(A)$
3. $A \to \varepsilon \in P \implies \varepsilon \in \mathrm{fi}(A)$
4. $\mathrm{fi}(\varepsilon) = \{\varepsilon\}$
5. $a \in \mathrm{fi}(A) \implies a \in \mathrm{fi}(A\alpha)$
6. $A \to \alpha B \in P, x \in \mathrm{fo}(A) \implies x \in \mathrm{fo}(B)$

Example 7.8

Grammar for arithmetic expressions (cf. Example 5.11):

$$
\begin{array}{lrcl}
G_{AE}: & E & \to & E\texttt{+}T \mid T \\
& T & \to & T\texttt{*}F \mid F \\
& F & \to & \texttt{(}E\texttt{)} \mid \texttt{a} \mid \texttt{b}
\end{array}
$$

- $F \to a \in P \implies a \in \mathrm{fi}(F)$
- $T \to F \in P, a \in \mathrm{fi}(F) \implies a \in \mathrm{fi}(T)$
- $a \in \mathrm{fi}(T)$
  $\implies \mathrm{la}(T \to T\texttt{*}F) = \mathrm{fi}(T\texttt{*}F \cdot \mathrm{fo}(T)) \ni a$
- $a \in \mathrm{fi}(F)$
  $\implies \mathrm{la}(T \to F) = \mathrm{fi}(F \cdot \mathrm{fo}(T)) \ni a$
- $\implies \mathrm{la}(T \to T\texttt{*}F) \cap \mathrm{la}(T \to F) \neq \emptyset$
- $\implies G_{AE} \notin LL(1)$

Fixing the Problem

(general methods later)

Example 7.8 (continued)

Restructuring (such that $L(G'_{AE}) = L(G_{AE})$):

$$
\begin{array}{llll}
G'_{AE}: & E & \rightarrow & TE' \\
 & E' & \rightarrow & \mathtt{+}TE' \mid \varepsilon \\
 & T & \rightarrow & FT' \\
 & T' & \rightarrow & \mathtt{*}FT' \mid \varepsilon \\
 & F & \rightarrow & \mathtt{(}E\mathtt{)} \mid \mathtt{a} \mid \mathtt{b}
\end{array}
$$

$A \in N$	$\mathrm{fi}(A)$	$\mathrm{fo}(A)$
E	$\{\mathtt{(}, \mathtt{a}, \mathtt{b}\}$	$\{\varepsilon, \mathtt{)}\}$
E'	$\{\mathtt{+}, \varepsilon\}$	$\{\varepsilon, \mathtt{)}\}$
T	$\{\mathtt{(}, \mathtt{a}, \mathtt{b}\}$	$\{\mathtt{+}, \varepsilon, \mathtt{)}\}$
T'	$\{\mathtt{*}, \varepsilon\}$	$\{\mathtt{+}, \varepsilon, \mathtt{)}\}$
F	$\{\mathtt{(}, \mathtt{a}, \mathtt{b}\}$	$\{\mathtt{*}, \mathtt{+}, \varepsilon, \mathtt{)}\}$

$A \rightarrow \beta \in P$	$\mathrm{la}(A \rightarrow \beta) = \mathrm{fi}(\beta \cdot \mathrm{fo}(A))$
$E \rightarrow TE'$	$\{\mathtt{(}, \mathtt{a}, \mathtt{b}\}$
$E' \rightarrow \mathtt{+}TE'$	$\{\mathtt{+}\}$
$E' \rightarrow \varepsilon$	$\{\varepsilon, \mathtt{)}\}$
$T \rightarrow FT'$	$\{\mathtt{(}, \mathtt{a}, \mathtt{b}\}$
$T' \rightarrow \mathtt{*}FT'$	$\{\mathtt{*}\}$
$T' \rightarrow \varepsilon$	$\{\mathtt{+}, \varepsilon, \mathtt{)}\}$
$F \rightarrow \mathtt{(}E\mathtt{)}$	$\{\mathtt{(}\}$
$F \rightarrow \mathtt{a}$	$\{\mathtt{a}\}$
$F \rightarrow \mathtt{b}$	$\{\mathtt{b}\}$

$\Longrightarrow$ $G'_{AE} \in LL(1)$

Outline

Deterministic Top-Down Parsing

Approach: given $G \in CFG_\Sigma$,

1. Verify that $G \in LL(1)$ by computing the lookahead sets and checking alternatives for disjointness
2. Start with nondeterministic top-down parsing automaton $\mathrm{NTA}(G)$
3. Use 1-symbol lookahead to control the choice of expanding productions:
   - $(aw, A\alpha, z) \vdash (aw, \beta\alpha, zi)$
     if $\pi(i) = A \to \beta$ and $a \in \mathrm{la}(\pi(i))$
   - $(\varepsilon, A\alpha, z) \vdash (\varepsilon, \beta\alpha, zi)$
     if $\pi(i) = A \to \beta$ and $\varepsilon \in \mathrm{la}(\pi(i))$
   - [as before: $(aw, a\alpha, z) \vdash (w, \alpha, z)$]

$\Longrightarrow$ deterministic top-down parsing automaton $\mathrm{DTA}(G)$

Remarks:

- $\mathrm{DTA}(G)$ is actually not a pushdown automaton (a is read but not consumed). But: can be simulated using the finite control.
- Advantage of using lookahead is twofold:
  - Removal of nondeterminism
  - Earlier detection of syntax errors
    (in configurations $(aw, A\alpha, z)$ where $a \notin \bigcup_{A \to \beta \in P} \mathrm{la}(A \to \beta)$)

The Deterministic Top-Down Automaton I

Definition 7.9 (Deterministic top-down parsing automaton)

Let $G = \langle N, \Sigma, P, S \rangle \in LL(1)$. The deterministic top-down parsing automaton of G, $\mathrm{DTA}(G)$, is defined by the following components.

- Input alphabet Σ, pushdown alphabet X, output alphabet $[p]$
- Configurations $\Sigma^* \times X^* \times [p]^*$, initial configuration (w, S, ε), final configurations $\{\varepsilon\} \times \{\varepsilon\} \times [p]^*$ (as $\mathrm{NTA}(G)$)
- Action function

$$\mathrm{act} : \Sigma_\varepsilon \times X_\varepsilon \to \{(\alpha, i) \mid \pi(i) = A \to \alpha\} \cup \{\mathsf{pop}, \mathsf{accept}, \mathsf{error}\}$$

$$\begin{aligned} \text{with } \mathrm{act}(x, A) &:= (\alpha, i) \text{ if } \pi(i) = A \to \alpha \text{ and } x \in \mathrm{la}(\pi(i)) \\ \mathrm{act}(a, a) &:= \mathsf{pop} \\ \mathrm{act}(\varepsilon, \varepsilon) &:= \mathsf{accept} \\ \mathrm{act}(x, y) &:= \mathsf{error} \qquad \text{otherwise} \end{aligned}$$

- Transitions for $x \in \Sigma_\varepsilon$, $w \in \Sigma^*$, $Y \in X$, $\beta \in X^*$, and $z \in [p]^*$:

$$(xw, Y\beta, z) \vdash \begin{cases} (xw, \alpha\beta, zi) & \text{if } \mathrm{act}(x, Y) = (\alpha, i) \\ (w, \beta, z) & \text{if } \mathrm{act}(x, Y) = \mathsf{pop} \end{cases}$$

Example 7.10

$$
\begin{array}{llll}
G'_{AE}: & E & \rightarrow TE' & (1) \\
& E' & \rightarrow \texttt{+}TE' \mid \varepsilon & (2,3) \\
& T & \rightarrow FT' & (4) \\
& T' & \rightarrow \texttt{*}FT' \mid \varepsilon & (5,6) \\
& F & \rightarrow \texttt{(}E\texttt{)} \mid \texttt{a} \mid \texttt{b} & (7,8,9)
\end{array}
$$

$A \rightarrow \beta \in P$	$\mathrm{la}(A \rightarrow \beta)$
$E \rightarrow TE'$	$\{\texttt{(}, \texttt{a}, \texttt{b}\}$
$E' \rightarrow \texttt{+}TE'$	$\{\texttt{+}\}$
$E' \rightarrow \varepsilon$	$\{\varepsilon, \texttt{)}\}$
$T \rightarrow FT'$	$\{\texttt{(}, \texttt{a}, \texttt{b}\}$
$T' \rightarrow \texttt{*}FT'$	$\{\texttt{*}\}$
$T' \rightarrow \varepsilon$	$\{\texttt{+}, \varepsilon, \texttt{)}\}$
$F \rightarrow \texttt{(}E\texttt{)}$	$\{\texttt{(}\}$
$F \rightarrow \texttt{a}$	$\{\texttt{a}\}$
$F \rightarrow \texttt{b}$	$\{\texttt{b}\}$

$\mathrm{act} : \Sigma_\varepsilon \times X_\varepsilon \rightarrow \{(\alpha, i) \mid \pi(i) = A \rightarrow \alpha\} \cup \{\mathsf{pop}, \mathsf{accept}, \mathsf{error}\}$ (empty $=$ error)

act	E	E'	T	T'	F	a	b	(	)	*	+	ε
a	$(TE',1)$		$(FT',4)$		$(\texttt{a},8)$	pop						
b	$(TE',1)$		$(FT',4)$		$(\texttt{b},9)$		pop					
(	$(TE',1)$		$(FT',4)$		$(\texttt{(}E\texttt{)},7)$			pop				
)		$(\varepsilon,3)$		$(\varepsilon,6)$					pop			
*				$(\texttt{*}FT',5)$						pop		
+		$(\texttt{+}TE',2)$		$(\varepsilon,6)$							pop	
ε		$(\varepsilon,3)$		$(\varepsilon,6)$								accept

Example 7.10 (continued)

act	E	E'	T	T'	F	a	b	(	)	*	+	ε
a	$(TE',1)$		$(FT',4)$		$(\mathtt{a},8)$	pop						
b	$(TE',1)$		$(FT',4)$		$(\mathtt{b},9)$		pop					
(	$(TE',1)$		$(FT',4)$		$(\mathtt{(}E\mathtt{)},7)$			pop				
)		$(\varepsilon,3)$		$(\varepsilon,6)$					pop			
*				$(\mathtt{*}FT',5)$						pop		
+		$(\mathtt{+}TE',2)$		$(\varepsilon,6)$							pop	
ε		$(\varepsilon,3)$		$(\varepsilon,6)$								accept

Leftmost analysis of `(a)*b`:

$$
\begin{array}{ll}
 & (\mathtt{(a)*b},\ E,\ \varepsilon) \\
\vdash & (\mathtt{(a)*b},\ TE',\ 1) \\
\vdash & (\mathtt{(a)*b},\ FT'E',\ 14) \\
\vdash & (\mathtt{(a)*b},\ \mathtt{(}E\mathtt{)}T'E',\ 147) \\
\vdash & (\mathtt{a)*b},\ E\mathtt{)}T'E',\ 147) \\
\vdash & (\mathtt{a)*b},\ TE'\mathtt{)}T'E',\ 1471) \\
\vdash & (\mathtt{a)*b},\ FT'E'\mathtt{)}T'E',\ 14714) \\
\vdash & (\mathtt{a)*b},\ \mathtt{a}T'E'\mathtt{)}T'E',\ 147148) \\
\vdash & (\mathtt{)*b},\ T'E'\mathtt{)}T'E',\ 147148) \\
\vdash & (\mathtt{)*b},\ E'\mathtt{)}T'E',\ 1471486) \\
\vdash & (\mathtt{)*b},\ \mathtt{)}T'E',\ 14714863) \\
\vdash & (\mathtt{*b},\ T'E',\ 14714863) \\
\vdash & (\mathtt{*b},\ \mathtt{*}FT'E',\ 147148635) \\
\vdash & (\mathtt{b},\ FT'E',\ 147148635) \\
\vdash & (\mathtt{b},\ \mathtt{b}T'E',\ 1471486359) \\
\vdash & (\varepsilon,\ T'E',\ 1471486359) \\
\vdash & (\varepsilon,\ E',\ 14714863596) \\
\vdash & (\varepsilon,\ \varepsilon,\ 147148635963)
\end{array}
$$