Compiler Construction

Lecture 6: Syntax Analysis II ($LL(k)$ Grammars)

Thomas Noll

Lehrstuhl für Informatik 2
(Software Modeling and Verification)

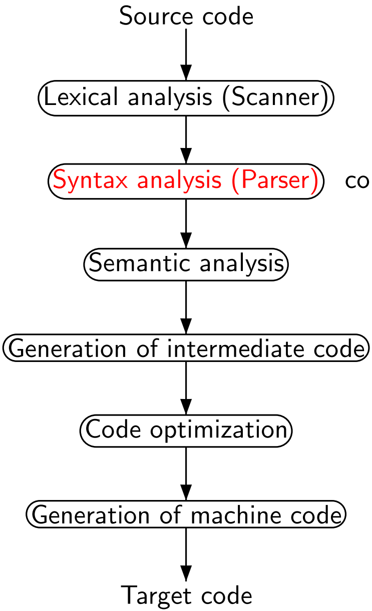
RWTH AACHEN UNIVERSITY

noll@cs.rwth-aachen.de

http://moves.rwth-aachen.de/teaching/ss-14/cc14/

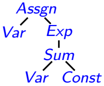
Summer Semester 2014

Outline



(id, $\texttt{x1}$)(gets,)(id, $\texttt{y2}$)(plus,)(int, 1)

context-free grammars/pushdown automata



Approach:

1. Given $G \in CFG_\Sigma$, construct a nondeterministic pushdown automaton (PDA) which accepts $L(G)$ and which additionally computes corresponding leftmost derivations (similar to the proof of "$L(CFG_\Sigma) \subseteq L(PDA_\Sigma)$")
   - input alphabet: Σ
   - pushdown alphabet: X
   - output alphabet: $[p]$
   - state set: not required
2. Remove nondeterminism by allowing lookahead on the input: $G \in LL(k)$ iff $L(G)$ recognizable by deterministic PDA with lookahead of k symbols

The Nondeterministic Top-Down Automaton

Definition (Nondeterministic top-down parsing automaton)

Let $G = \langle N, \Sigma, P, S\rangle \in CFG_\Sigma$. The nondeterministic top-down parsing automaton of G, $\mathrm{NTA}(G)$, 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 left)
- Transitions for $w \in \Sigma^*$, $\alpha \in X^*$, and $z \in [p]^*$:

  expansion steps: if $\pi_i = A \to \beta$, then $(w, A\alpha, z) \vdash (w, \beta\alpha, zi)$

  matching steps: for every $a \in \Sigma$, $(aw, a\alpha, z) \vdash (w, \alpha, z)$
- Initial configuration for $w \in \Sigma^*$: (w, S, ε)
- Final configurations: $\{\varepsilon\} \times \{\varepsilon\} \times [p]^*$

Remark: $\mathrm{NTA}(G)$ is nondeterministic iff G contains $A \to \beta \mid \gamma$

Outline

Correctness of $\mathrm{NTA}(G)$

Theorem 6.1 (Correctness of $\mathrm{NTA}(G)$)

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

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

Proof.

$\Longrightarrow$ (soundness): see exercises

$\Longleftarrow$ (completeness): on the board

□

Outline

Adding Lookahead

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

$\Longrightarrow$ determination of expanding A-production by next k symbols

Definition 6.2 (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, $\mathrm{first}_k(\alpha) \subseteq \Sigma^*$, is given by

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

Remark: $\mathrm{first}_k(\alpha)$ is effectively computable. If $\alpha \in \Sigma^*$, then $|\mathrm{first}_k(\alpha)| = 1$.

Example 6.3 (first_k set)

Let $G : S \to aSb \mid \varepsilon$.

1. $\mathrm{first}_1(ab) = \{a\} = \mathrm{first}_2(a)$
2. $\mathrm{first}_3(S) = \{\varepsilon, ab, aab, aaa\}$
3. $\mathrm{first}_3(Sa) = \{a, aba, aab, aaa\}$

Outline

$LL(k)$ Grammars I

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

Definition 6.4 ($LL(k)$ 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 $\beta \neq \gamma$, it follows that $\mathrm{first}_k(x) \neq \mathrm{first}_k(y)$
(i.e., different productions must not yield the same lookahead).

Remarks:

- If $G \in LL(k)$, then the leftmost derivation step for $wA\alpha$ in
$$S \Rightarrow_l^* wA\alpha \begin{cases} \Rightarrow_l w\beta\alpha \Rightarrow_l^* wx \\ \Rightarrow_l w\gamma\alpha \Rightarrow_l^* wy \end{cases}$$
is determined by the next k symbols following w.
- Corresponding computations of $\mathrm{NTA}(G)$:
$$\begin{array}{ccccccc} (wx, S, \varepsilon) & \vdash^* & (x, A\alpha, z) & \overset{(*)}{\vdash} & (x, \beta\alpha, zi) & \vdash^* & (\varepsilon, \varepsilon, ziz') \\ (wy, S, \varepsilon) & \vdash^* & (y, A\alpha, z) & \overset{(*)}{\vdash} & (y, \gamma\alpha, zj) & \vdash^* & (\varepsilon, \varepsilon, zjz'') \end{array}$$
where $\pi_i = A \rightarrow \beta$ and $\pi_j = A \rightarrow \gamma$
- Deterministic decision in $(*)$ possible if $\mathrm{first}_k(x) \neq \mathrm{first}_k(y)$
- **Problem:** how to determine the A-production from the lookahead (potentially infinitely many derivations $\beta\alpha \Rightarrow_l^* x$ / $\gamma\alpha \Rightarrow_l^* y$)?

Lemma 6.5 (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 $\mathrm{first}_k(\beta\alpha) \cap \mathrm{first}_k(\gamma\alpha) = \emptyset$.

Proof.

omitted □

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 right contexts α to be considered (if β [or γ] "too short", i.e., $\mathrm{first}_k(\beta\alpha) \neq \mathrm{first}_k(\beta)$).
- **Idea:** α derives to "everything that follows A"

Outline

The follow_k Sets

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

Definition 6.6 (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, $\mathrm{follow}_k(A) \subseteq \Sigma^*$, is given by

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

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 6.7

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 6.8 (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 6.9

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* $\varepsilon \in \mathrm{fo}(A)$