Einführung in die Computerlinguistik

Kontextfreie Grammatiken

Laura Kallmeyer
Heinrich-Heine-Universität Düsseldorf

Sommersemester 2013

Overview

1. Kontextfreie Grammatiken (CFG)
2. Kellerautomaten (PDA)
3. Äquivalenz von CFG und PDA

[Hopcroft and Ullman, 1979]

CFG (1)

Example: Grammar $G_{telescope}$:

Productions:

S → NP VP	NP → D N	N → N PP
VP → VP PP	VP → V NP	PP → P NP
N → man	N → girl	N → telescope
D → the	NP → John	NP → Mary
P → with	V → saw	

CFG (2)

Sentences one can generate with this grammar:

(1) John saw Mary

(2) John saw the girl

(3) the man with the telescope saw John

(4) John saw the girl with the telescope

...

CFG (3)

A context-free grammar (CFG) is a tuple $G = \langle N, T, P, S \rangle$ such that

- N and T are disjoint alphabets, the nonterminals and terminals,
- $S \in N$ is the start symbol, and
- P is a set of productions of the form $A \to \beta$ with $A \in N, \beta \in (N \cup T)^*$.

Any $\beta \in (N \cup T)^*$ with $S \stackrel{*}{\Rightarrow} \beta$ is called a sentential form.

CFG (4)

Bsp.: CFG $G_{a,b}$ mit Produktionen

S → aB	S → bA	
A → a	A → aS	A → bAA
B → b	B → bS	B → aBB

erzeugt die Sprache $\{w \mid w \in \{a,b\}^+ . |w|_a = |w|_b\}$

CFG (5)

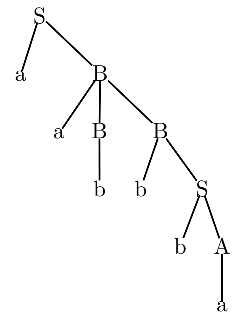
A tree t is a parse tree for a CFG $G = \langle N, T, P, S \rangle$ iff

- each node in t is labeled with a $x \in N \cup T \cup \{\epsilon\}$;
- the root label is S;
- if there is a node with label A that has n daughters labeled (from left to right) $x_1, \dots, x_n$, then $A \to x_1 \dots x_n \in P$.
- if a node has label ϵ, it is a leaf and the unique daughter of its mother node.

$S \stackrel{*}{\Rightarrow} \alpha$ in G iff there is a parse tree for G with yield α.

CFG (6)

	S → aB	S → bA	
$G_{a,b}$:	A → a	A → aS	A → bAA
	B → b	B → bS	B → aBB



CFG (7)

- Let $G = \langle N, T, P, S\rangle$ be a CFG. The string language $L(G)$ of G is the set $\{w \in T^* \mid S \overset{*}{\Rightarrow} w\}$ where
  - for $w, w' \in (N \cup T)^*$: $w \Rightarrow w'$ iff there is a $A \to \beta \in P$ and there are $v, u \in (N \cup T)^*$ such that $w = vAu$ and $w' = v\beta u$.
  - $\overset{*}{\Rightarrow}$ is the reflexive transitive closure of $\Rightarrow$.
- A derivation of a word $w \in T^*$ is a sequence $S \Rightarrow \alpha_1 \cdots \Rightarrow w$ of derivation steps leading to w.
- The tree language is the set of all parse trees with root label S and all leaves labelled with $a \in T \cup \{\varepsilon\}$.

CFG (8)

For a single parse tree, there might be more than one corresponding derivation.

A derivation is called a

- leftmost derivation iff, in each derivation step, a production is applied to the leftmost non-terminal of the already derived sentential form.
- rightmost derivation iff, in each derivation step, a production is applied to the rightmost non-terminal of the already derived sentential form.

CFG (9)

For a single word w, there might be more than one parse tree:

- A CFG giving more than one parse tree for some word w is called ambiguous.

  Example: $G_{telescope}$ with
  $w =$ John saw the man with the telescope

  $G_{a,b}$ with $w = aabbab$
- A CFL L is called inherently ambiguous if each CFG G with $L = L(G)$ is ambiguous.

  Example:

  $\{a^n b^n c^m d^m \mid n \geq 1, m \geq 1\} \cup \{a^n b^m c^m d^n \mid n \geq 1, m \geq 1\}$

PDA (1)

A push-down automaton is a FSA with an additional stack. The moves of the automaton depend on

- the current state,
- the next input symbol, and
- the topmost stack symbol.

Each move consists of

- changing state,
- popping the topmost symbol from the stack, and
- pushing a new sequence of symbols on the stack.

PDA (2)

Example: Automaton that

- starts with q_1 and stack $\#$,
- in q_1: pushes A on the stack for an input symbol a,
- in q_1: leaves stack unchanged and goes to q_2 for an input symbol c,
- in q_2: pops an A from the stack for input symbol b,
- in q_2: moves to q_3 if the top of the stack is $\#$

The automaton accepts all words that allow to end up in q_3.

Language $\{a^n c b^n \mid n \geq 0\}$

PDA (3)

In general, PDAs are non-deterministic, since a given state, input symbol and topmost stack symbol can allow for more than one move.

In contrast to FSA, the deterministic version of the automaton is not equivalent to the non-deterministic one: There are languages that are accepted by a non-deterministic PDA but not by any deterministic PDA.

CFLs are the languages accepted by (non-deterministic) PDAs.

PDA (4)

A push-down automaton (PDA) M is a tuple $\langle Q, \Sigma, \Gamma, \delta, q_0, Z_0, F\rangle$ with

- Q is a finite set of states.
- Σ is a finite set, the input alphabet.
- Γ is a finite set, the stack alphabet.
- $q_0 \in Q$ is the initial state.
- $Z_0 \in \Gamma$ is the initial stack symbol.
- $F \subseteq Q$ is the set of final states.
- $\delta : Q \times (\Sigma \cup \{\epsilon\}) \times \Gamma \rightarrow \mathcal{P}_{fin}(Q \times \Gamma^*)$ is the transition function. ($\mathcal{P}_{fin}(X)$ is the set of finite subsets of X).

PDA (5)

An instantaneous description of a PDA is a triple (q, w, γ) with

- $q \in Q$ is the current state of the automaton,
- $w \in \Sigma^*$ is the remaining part of the input string, and
- $\gamma \in \Gamma^*$ is the current stack.

$(q, aw, Z\alpha) \vdash (q', w, \beta\alpha)$ iff $\langle q', \beta\rangle \in \delta(q, a, Z)$ for all $q, q' \in Q, a \in \Sigma \cup \{\epsilon\}, w \in \Sigma^*, Z \in \Gamma, \alpha, \beta \in \Gamma^*$.

$\overset{*}{\vdash}$ is the reflexive transitive closure of $\vdash$.

PDA (6)

There are two alternatives for the definition of the language accepted by a PDA $M = \langle Q, \Sigma, \Gamma, \delta, q_0, Z_0, F\rangle$:

- The language accepted by M with a final state is

  $L(M) := \{w \mid (q_0, w, Z_0) \overset{*}{\vdash} (q_f, \epsilon, \gamma) \text{ for a } q_f \in F \text{ and a } \gamma \in \Gamma^*\}$
- The language accepted by M with an empty stack is

  $N(M) := \{w \mid (q_0, w, Z_0) \overset{*}{\vdash} (q, \epsilon, \epsilon) \text{ for a } q \in Q\}$

The two modes of acceptance are equivalent, i.e., for each language L: there is a PDA M_1 with $L = L(M_1)$ iff there is a PDA M_2 with $L = N(M_2)$.

PDA (7)

Example: PDA M for $L(M) = \{wcw^R \mid w \in \{a, b\}^*\}$.

- $Q = \{q_1, q_2, q_3\}$, $\Sigma = \{a, b, c\}$, $\Gamma = \{\#, A, B\}$.
  $q_0 = q_1$, $Z_0 = \#$, $F = \{q_3\}$.
- $\delta(q_1, a, \#) = \{\langle q_1, A\#\rangle\}$ $\delta(q_1, b, \#) = \{\langle q_1, B\#\rangle\}$
  $\delta(q_1, a, A) = \{\langle q_1, AA\rangle\}$ $\delta(q_1, b, A) = \{\langle q_1, BA\rangle\}$
  $\delta(q_1, a, B) = \{\langle q_1, AB\rangle\}$ $\delta(q_1, b, B) = \{\langle q_1, BB\rangle\}$
- $\delta(q_1, c, \#) = \{\langle q_2, \#\rangle\}$
  $\delta(q_1, c, A) = \{\langle q_2, A\rangle\}$
  $\delta(q_1, c, B) = \{\langle q_2, B\rangle\}$
- $\delta(q_2, a, A) = \{\langle q_2, \epsilon\rangle\}$ $\delta(q_2, b, B) = \{\langle q_2, \epsilon\rangle\}$
- $\delta(q_2, \epsilon, \#) = \{\langle q_3, \#\rangle\}$

PDA (8)

Example: PDA M for $N(M) = \{wcw^R \mid w \in \{a, b\}^*\}$.

- $Q = \{q_1, q_2\}$, $\Sigma = \{a, b, c\}$, $\Gamma = \{\#, A, B\}$.
  $q_0 = q_1$, $Z_0 = \#$, $F = \emptyset$.
- $\delta(q_1, a, \#) = \{\langle q_1, A\#\rangle\}$ $\delta(q_1, b, \#) = \{\langle q_1, B\#\rangle\}$
  $\delta(q_1, a, A) = \{\langle q_1, AA\rangle\}$ $\delta(q_1, b, A) = \{\langle q_1, BA\rangle\}$
  $\delta(q_1, a, B) = \{\langle q_1, AB\rangle\}$ $\delta(q_1, b, B) = \{\langle q_1, BB\rangle\}$
- $\delta(q_1, c, \#) = \{\langle q_2, \#\rangle\}$
  $\delta(q_1, c, A) = \{\langle q_2, A\rangle\}$
  $\delta(q_1, c, B) = \{\langle q_2, B\rangle\}$
- $\delta(q_2, a, A) = \{\langle q_2, \epsilon\rangle\}$ $\delta(q_2, b, B) = \{\langle q_2, \epsilon\rangle\}$
- $\delta(q_2, \epsilon, \#) = \{\langle q_2, \epsilon\rangle\}$

PDA (9)

A PDA $M = \langle Q, \Sigma, \Gamma, \delta, q_0, Z_0, F\rangle$ is a deterministic PDA (DPDA) iff

- for all $q \in Q, Z \in \Gamma, a \in \Sigma \cup \{\epsilon\}$: $|\delta(q, a, Z)| \leq 1$, and
- for all $q \in Q, Z \in \Gamma$: if $\delta(q, \epsilon, Z) \neq \emptyset$, then $\delta(q, a, Z) = \emptyset$ for all $a \in \Sigma$.

The preceding example was a DPDA.

The class of languages accepted by DPDAs is smaller than the class accepted by (non-deterministic) PDAs.

Example of a language that requires a non-determinstic PDA: $\{ww^R \mid w \in \{a, b\}^*\}$.

PDA and CFG (1)

For each CFL L, there is a PDA M with $L = N(M)$.

Construction: Assume that $\epsilon \notin L$. $L = L(G)$ for a CFG $G = \langle N, T, P, S \rangle$ in Greibach-Normal Form (GNF).

This means that all productions have the form $A \to a\gamma$ with $A \in N, a \in T, \gamma \in N^*$.

Every CFG can be transformed into an equivalent CFG in GNF.

Equivalent PDA:
$M = \langle \{q\}, T, N, \delta, q, S, \emptyset \rangle$ with $\langle q, \gamma \rangle \in \delta(q, a, A)$ iff $A \to a\gamma \in P$.

The automaton simulates leftmost derivations in G.

PDA and CFG (2)

For each PDA M with $L = N(M)$: L is a context-free language.

Construction of equivalent CFG for given PDA
$M = \langle Q, \Sigma, \Gamma, \delta, q_0, Z_0, F \rangle$:

- nonterminals: S and all $[q_1, Z, q_2]$ with $q_1, q_2 \in Q$, $Z \in \Gamma$.
- productions: $S \to [q_0, Z_0, q]$ for every $q \in Q$, and
  $[q, A, q_{m+1}] \to a[q_1, B_1, q_2], \ldots, [q_m, B_m, q_{m+1}]$ for
  $q, q_1, \ldots, q_{m+1} \in Q, a \in \Sigma \cup \{\epsilon\}, A, B_1, \ldots, B_m \in \Gamma$ such that
  $\langle q_1, B_1 \ldots B_m \rangle \in \delta(q, a, A)$
  $[q, A, q_1] \to a$ if $\langle q_1, \epsilon \rangle \in \delta(q, a, A)$.

$[q_1, A, q_2] \stackrel{*}{\Rightarrow} w$ iff $(q_1, w, A) \stackrel{*}{\vdash} (q_2, \epsilon, \epsilon)$.

References

[Hopcroft and Ullman, 1979] Hopcroft, J. E. and Ullman, J. D. (1979). *Introduction to Automata Theory, Languages and Computation.* Addison Wesley.