Computational Linguistics II: Parsing

Formal Languages: Overview & Regular Languages

Frank Richter & Jan-Philipp Söhn

fr@sfs.uni-tuebingen.de, jp.soehn@uni-tuebingen.de

Origins of Formal Language Theory

- Biology (neuron nets)
- Electrical Engineering (switching circuits, hardware design)
- Mathematics (foundations of logic)
- Linguistics (grammars for natural languages)

The Big Picture

hierarchy	grammar	machine	other
type 3	reg. grammar	DFA	reg. expressions
det. cf.	LR(k) grammar	DPDA	
type 2	CFG	PDA	
type 1	CSG	LBA	
type 0	unrestricted	Turing	
	grammar	machine	

The Big Picture

hierarchy	grammar	machine	other
type 3	reg. grammar	DFA	reg. expressions
det. cf.	LR(k) grammar	DPDA	
type 2	CFG	PDA	
type 1	CSG	LBA	
type 0	unrestricted grammar	Turing machine	

DFA: Deterministic finite state automaton

(D)PDA: (Deterministic) Pushdown automaton

CFG: Context-free grammar

CSG: Context-sensitive grammar

LBA: Linear bounded automaton

Form of Grammars of Type 0–3

For $i \in \{0, 1, 2, 3\}$, a grammar $\langle N, T, P, S\rangle$ of Type i, with N the set of non-terminal symbols, T the set of terminal symbols (N and T disjoint, $\Sigma = N \cup T$), P the set of productions, and S the start symbol ($S \in N$), obeys the following restrictions:

T3: Every production in P is of the form $A \longrightarrow aB$ or $A \longrightarrow \epsilon$, with $B, A \in N$, $a \in T$.

T2: Every production in P is of the form $A \longrightarrow x$, with $A \in N$ and $x \in \Sigma^*$.

T1: Every production in P is of the form $x_1 A x_2 \longrightarrow x_1 y x_2$, with $x_1, x_2 \in \Sigma^*$, $y \in \Sigma^+$, $A \in N$ and the possible exception of $C \longrightarrow \epsilon$ in case C does not occur on the righthand side of a rule in P.

T0: No restrictions.

Deterministic Finite-State Automata

Definition 1 (DFA) A deterministic FSA (DFA) is a quintuple $(\Sigma, Q, i, F, \delta)$ where

Σ is a finite set called *the alphabet*,

Q is a finite set of *states*,

$i \in Q$ is the *initial state*,

$F \subseteq Q$ the set of *final states*, and

δ is the transition function from $Q \times \Sigma$ to Q.

Transition Closure

Definition 2 For each DFA $(\Sigma, Q, i, F, \delta)$, for each $q \in Q$, for each $a \in \Sigma$, for each $x \in \Sigma^*$,

$\hat{\delta}(q, \epsilon) = q$, and

$\hat{\delta}(q, ax) = \hat{\delta}(\delta(q, a), x)$

Acceptance

Definition 3 (Acceptance)

Given a DFA $M = (\Sigma, Q, i, F, \delta)$, the language $L(M)$ accepted by M is

$$L(M) = \{x \in \Sigma^* | \hat{\delta}(i, x) \in F\}.$$