Formální jazyky a automaty

Bezkontextové gramatiky, derivační stromy, redukované gramatiky

Jan Křetínský

Fakulta informatiky, MU Brno

Jaro 2024

Bezkontextové jazyky

Bezkontextová gramatika (Context-free grammar, CFG)

Bezkontextová gramatika $\mathcal{G}$ je čtveřice (N, Σ, P, S), kde

- N je neprázdná konečná množina **neterminálních symbolů**,
- Σ je konečná množina **terminálních symbolů** taková, že $N \cap \Sigma = \emptyset$ (značení: $V = N \cup \Sigma$),
- $S \in N$ je **počáteční neterminál**,
- $P \subseteq N \times V^*$ je konečná množina **pravidel**.

Jazyk je **bezkontextový**, pokud je generovaný nějakou bezkontextovou gramatikou.

Příklad

$\mathcal{G} = (\{E, T, F\}, \{+, *, (,), i\}, P, E)$, kde P obsahuje pravidla

$$
\begin{aligned}
E &\rightarrow E + T \mid T \\
T &\rightarrow T * F \mid F \\
F &\rightarrow (E) \mid i
\end{aligned}
$$

Derivační stromy pro bezkontextové gramatiky

Definice 3.1.

Nechť $\mathcal{G} = (N, \Sigma, P, S)$ je CFG.
Strom T nazveme derivačním stromem v $\mathcal{G}$ právě když

1. kořen má návěští S, vnitřní uzly mají návěští z N, listy mají navěští z $N \cup \Sigma \cup \{\varepsilon\}$,
2. má-li vnitřní uzel návěští A a jeho všichni synové $n_1, \ldots, n_k$ mají v uspořádání zleva doprava návěští $X_1, \ldots, X_k \in N \cup \Sigma \cup \{\varepsilon\}$, pak $A \to X_1 \ldots X_k \in P$,
3. každý list s návěštím ε je jediným synem svého otce.

Výsledkem derivačního stromu T nazveme slovo vzniklé zřetězením návěští listů v uspořádání zleva doprava.

Derivační stromy pro bezkontextové gramatiky

Definice 3.1.

Nechť $\mathcal{G} = (N, \Sigma, P, S)$ je CFG.
Strom T nazveme derivačním stromem v $\mathcal{G}$ právě když

1. kořen má návěští S, vnitřní uzly mají návěští z N, listy mají navěští z $N \cup \Sigma \cup \{\varepsilon\}$,
2. má-li vnitřní uzel návěští A a jeho všichni synové $n_1, \ldots, n_k$ mají v uspořádání zleva doprava návěští $X_1, \ldots, X_k \in N \cup \Sigma \cup \{\varepsilon\}$, pak $A \to X_1 \ldots X_k \in P$,
3. každý list s návěštím ε je jediným synem svého otce.

Výsledkem derivačního stromu T nazveme slovo vzniklé zřetězením návěští listů v uspořádání zleva doprava.

Věta 3.3. Vztah mezi derivačními stromy a relací $\Rightarrow^*$

Nechť $\mathcal{G} = (N, \Sigma, P, S)$ je CFG. Pak pro libovolné $\alpha \in (N \cup \Sigma)^*$ platí $S \Rightarrow^* \alpha$ právě když v $\mathcal{G}$ existuje derivační strom s výsledkem α.

Důkaz. Rozšířením na $\forall A \in N$ + indukcí

Jednoznačnost derivačních stromů

Derivace

Derivace je sekvence $S \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \ldots \Rightarrow \alpha_n$.
Levá derivace je taková derivace, kde každé α_{i+1} vznikne z α_i přepsáním nejlevějšího neterminálu.

Každému derivačnímu stromu odpovídá jediná levá derivace.
Každé levé derivaci odpovídá jediný derivační strom.
Existuje pro každé $w \in L(\mathcal{G})$ právě jeden derivační strom?

Jednoznačnost derivačních stromů

Derivace

Derivace je sekvence $S \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \ldots \Rightarrow \alpha_n$.
Levá derivace je taková derivace, kde každé α_{i+1} vznikne z α_i přepsáním nejlevějšího neterminálu.

Každému derivačnímu stromu odpovídá jediná levá derivace.
Každé levé derivaci odpovídá jediný derivační strom.
Existuje pro každé $w \in L(\mathcal{G})$ právě jeden derivační strom?

Definice 3.5.

CFG $\mathcal{G}$ se nazývá víceznačná (nejednoznačná) právě když existuje $w \in L(\mathcal{G})$ mající alespoň dva různé derivační stromy.
V opačném případě říkáme, že $\mathcal{G}$ je jednoznačná.

Bezkontextový jazyk L se nazývá vnitřně (inherentně) víceznačný, právě když každá bezkontextová gramatika, která jej generuje, je víceznačná.

Příklad

Jednoznačnost derivačních stromů

Derivace

Derivace je sekvence $S \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \ldots \Rightarrow \alpha_n$.
Levá derivace je taková derivace, kde každé α_{i+1} vznikne z α_i přepsáním nejlevějšího neterminálu.

Každému derivačnímu stromu odpovídá jediná levá derivace.
Každé levé derivaci odpovídá jediný derivační strom.
Existuje pro každé $w \in L(\mathcal{G})$ právě jeden derivační strom?

Definice 3.5.

CFG $\mathcal{G}$ se nazývá víceznačná (nejednoznačná) právě když existuje $w \in L(\mathcal{G})$ mající alespoň dva různé derivační stromy.
V opačném případě říkáme, že $\mathcal{G}$ je jednoznačná.

Bezkontextový jazyk L se nazývá vnitřně (inherentně) víceznačný, právě když každá bezkontextová gramatika, která jej generuje, je víceznačná.

Příklad

$\{w\#vw^R v' \mid v, v, w \in \{0,1\}^*\}$

Kanonické tvary bezkontextových gramatik

- redukované bezkontextové gramatiky
- gramatiky bez ε-pravidel
- gramatiky bez jednoduchých pravidel
- vlastní gramatiky
- Chomského normální forma
- gramatiky bez levé rekurze
- Greibachové normální forma

Redukované bezkontextové gramatiky

Definice 3.7.

Symbol $X \in N \cup \Sigma$ je nepoužitelný v CFG $\mathcal{G} = (N, \Sigma, P, S)$ právě když v $\mathcal{G}$ neexistuje derivace tvaru

$$S \Rightarrow^* wXy \Rightarrow^* wxy$$

pro žádné $w, x, y \in \Sigma^*$. Řekneme, že $\mathcal{G}$ je redukovaná, jestliže neobsahuje žádné nepoužitelné symboly.

X je nepoužitelný typu I (tj. nenormovaný) $\iff$ neexistuje $w \in \Sigma^*$ splňující $X \Rightarrow^* w$

X je nepoužitelný typu II (tj. nedosažitelný) $\iff$ neexistují $\alpha, \beta \in (N \cup \Sigma)^*$ splňující $S \Rightarrow^* \alpha X \beta$

Nalezení nepoužitelných symbolů typu I

Nenormované neterminály: neexistuje $w \in \Sigma^*$: $A \Rightarrow^* w$

Vstup: CFG $\mathcal{G} = (N, \Sigma, P, S)$
Výstup: $N_e = \{A \mid \exists w \in \Sigma^*.\ A \Rightarrow^* w\}$ (normované neterminály)

1: $i := 0;\ N_0 := \emptyset$
2: **repeat**
3: $\quad i := i + 1$
4: $\quad N_i := N_{i-1} \cup \{A \mid A \to \alpha \in P,\ \alpha \in (N_{i-1} \cup \Sigma)^*\}$
5: **until** $N_i = N_{i-1}$
6: $N_e := N_i$

Normované terminály: Korektnost algoritmu 1/2

Konečnost

Normované terminály: Korektnost algoritmu 1/2

Konečnost.

Správnost výsledku: Dokážeme $A \in N_e \iff \exists w \in \Sigma^*.\ A \Rightarrow^* w$.

($\Longrightarrow$) Indukcí k i dokážeme $A \in N_i \implies \exists w \in \Sigma^*.\ A \Rightarrow^* w$.

Základní krok $i = 0$: Platí triviálně, protože $N_0 = \emptyset$.

Indukční krok: **(IP)** Tvrzení platí pro i. Dokážeme pro $i + 1$.

- $A \in N_i$. Tvrzení plyne z (IP).
- $A \in N_{i+1} \setminus N_i$. Pak existuje $A \to X_1 \dots X_k \in P$,
  kde každé X_j je terminál nebo neterminál patřící do N_i.
  Podle (IP) existuje w_j tak, že $X_j \Rightarrow^* w_j$.
  Tedy $A \Rightarrow X_1 \dots X_k \Rightarrow^* w_1 X_2 \dots X_k \Rightarrow^* \dots \Rightarrow^* w_1 \dots w_k$,
  kde $w_1 \dots w_k \in \Sigma^*$.

Normované terminály: Korektnost algoritmu 2/2

($\Longleftarrow$) Indukcí k n dokážeme

$$A \overset{n}{\Rightarrow} w, w \in \Sigma^* \implies A \in N_i \text{ pro nějaké } i.$$

Základní krok $n = 1$: $A \to w \in P$ okamžitě dává $i = 1$.

Indukční krok: **(IP)** Předpokládejme, že tvrzení platí pro všechna $n' \leq n$.

Nechť $A \overset{n+1}{\Rightarrow} w$. Pak $A \Rightarrow X_1 \ldots X_k \overset{n}{\Rightarrow} w$, kde $X_j \overset{n_j}{\Rightarrow} w_j$ a $n_j \leq n$.
Pokud $X_j \in N$, pak podle (IP) $X_j \in N_{i_j}$ pro nějaké i_j.
Pokud $X_j \in \Sigma$, klademe $i_j = 0$.
Položme $i = 1 + \max\{i_1, \ldots, i_k\}$. Pak zřejmě $A \in N_i$.

□

Normovaná gramatika

Důsledek 3.10.

Existuje algoritmus, který pro libovolnou danou CFG $\mathcal{G}$ rozhoduje, zda $L(\mathcal{G}) = \emptyset$.

Důkaz.

Stačí ověřit, zda $S \notin N_e$. □

Normovaná gramatika

Důsledek 3.10.

Existuje algoritmus, který pro libovolnou danou CFG $\mathcal{G}$ rozhoduje, zda $L(\mathcal{G}) = \emptyset$.

Důkaz.

Stačí ověřit, zda $S \notin N_e$. □

Věta.

Nechť $\mathcal{G} = (N, \Sigma, P, S)$ je CFG taková, že $L(\mathcal{G}) \neq \emptyset$. Pak existuje ekvivalentní CFG $\mathcal{G}'$ bez nepoužitelných neterminálů typu I.

Důkaz.

Stačí spočítat množinu N_e a položit $\mathcal{G}' = (N_e, \Sigma, P', S)$, kde $P' = P \cap N_e \times (N_e \cup \Sigma)^*$. □

Nalezení nepoužitelných symbolů typu II

Nedosažitelné symboly: neexistují $\alpha, \beta \in (N \cup \Sigma)^* : S \Rightarrow^* \alpha X \beta$

Vstup: CFG $\mathcal{G} = (N, \Sigma, P, S)$
Výstup: CFG $\mathcal{G}' = (N', \Sigma', P', S)$ bez nedosažitelných symbolů splňující $L(\mathcal{G}) = L(\mathcal{G}')$

Nalezení nepoužitelných symbolů typu II

Nedosažitelné symboly: neexistují $\alpha, \beta \in (N \cup \Sigma)^* : S \Rightarrow^* \alpha X \beta$

Vstup: CFG $\mathcal{G} = (N, \Sigma, P, S)$
Výstup: CFG $\mathcal{G}' = (N', \Sigma', P', S)$ bez nedosažitelných symbolů splňující $L(\mathcal{G}) = L(\mathcal{G}')$

1: $i := 0;\ V_0 := \{S\}$
2: **repeat**
3: $\quad i := i + 1$
4: $\quad V_i := V_{i-1} \cup \{X \in N \cup \Sigma \mid \exists A \in V_{i-1} \,.\, A \to \alpha' X \beta' \in P\}$
5: **until** $V_i = V_{i-1}$
6: $N' := N \cap V_i;\ \Sigma' := \Sigma \cap V_i;\ P' := P \cap (V_i \times V_i^*)$

Nalezení nepoužitelných symbolů typu II

Nedosažitelné symboly: neexistují $\alpha, \beta \in (N \cup \Sigma)^* : S \Rightarrow^* \alpha X \beta$

Vstup: CFG $\mathcal{G} = (N, \Sigma, P, S)$
Výstup: CFG $\mathcal{G}' = (N', \Sigma', P', S)$ bez nedosažitelných symbolů splňující $L(\mathcal{G}) = L(\mathcal{G}')$

1: $i := 0;\ V_0 := \{S\}$
2: **repeat**
3: $\quad i := i + 1$
4: $\quad V_i := V_{i-1} \cup \{X \in N \cup \Sigma \mid \exists A \in V_{i-1} \,.\, A \to \alpha' X \beta' \in P\}$
5: **until** $V_i = V_{i-1}$
6: $N' := N \cap V_i;\ \Sigma' := \Sigma \cap V_i;\ P' := P \cap (V_i \times V_i^*)$

Idea korektnosti: $X \in N' \cup \Sigma' \iff \exists \alpha, \beta \in (N' \cup \Sigma')^* \,.\, S \Rightarrow^* \alpha X \beta$

Příklad

$\mathcal{G} = (\{S, A, B\}, \{a, b, c, d, e\}, P, S)$, kde P obsahuje pravidla

$$
\begin{aligned}
S &\rightarrow aSb \mid c \mid aB \\
A &\rightarrow dA \mid d \\
B &\rightarrow eB
\end{aligned}
$$

Eliminace nepoužitelných symbolů

Věta 3.11.

Každý neprázdný bezkontextový jazyk L je generován nějakou redukovanou CFG.

Důkaz.

Nechť L je generován nějakou CFG $\mathcal{G}$.
Krok 1. Z $\mathcal{G}$ odstraníme symboly typu I *(výsledek označme $\mathcal{G}_1$)*.
Krok 2. Z $\mathcal{G}_1$ odstraníme symboly typu II *(výsledek označme $\mathcal{G}_2$)*.
Platí $L(\mathcal{G}) = L(\mathcal{G}_1) = L(\mathcal{G}_2)$. □

Korektnost: Dokážeme, že $\mathcal{G}_2$ je redukovaná CFG.
Nechť X je libovolný symbol z $\mathcal{G}_2$.

- v $\mathcal{G}_2$ existuje derivace $S \Rightarrow^*_{\mathcal{G}_2} \alpha X \beta$
- všechny symboly z $\mathcal{G}_2$ jsou též v $\mathcal{G}_1$
- pro nějaký terminální řetěz w platí $S \Rightarrow^*_{\mathcal{G}_2} \alpha X \beta \Rightarrow^*_{\mathcal{G}_1} w$
- žádný symbol z derivace $\alpha X \beta \Rightarrow^*_{\mathcal{G}_1} w$ není krokem 2 eliminován a proto $\alpha X \beta \Rightarrow^*_{\mathcal{G}_2} w$

Víme tedy, že existuje derivace $S \Rightarrow^*_{\mathcal{G}_2} \alpha X \beta \Rightarrow^*_{\mathcal{G}_2} w$, kde w je terminální řetěz. Tudíž X není nepoužitelný v $\mathcal{G}_2$. □

Korektnost: Dokážeme, že $\mathcal{G}_2$ je redukovaná CFG.
Nechť X je libovolný symbol z $\mathcal{G}_2$.

- v $\mathcal{G}_2$ existuje derivace $S \Rightarrow^*_{\mathcal{G}_2} \alpha X \beta$
- všechny symboly z $\mathcal{G}_2$ jsou též v $\mathcal{G}_1$
- pro nějaký terminální řetěz w platí $S \Rightarrow^*_{\mathcal{G}_2} \alpha X \beta \Rightarrow^*_{\mathcal{G}_1} w$
- žádný symbol z derivace $\alpha X \beta \Rightarrow^*_{\mathcal{G}_1} w$ není krokem 2 eliminován a proto $\alpha X \beta \Rightarrow^*_{\mathcal{G}_2} w$

Víme tedy, že existuje derivace $S \Rightarrow^*_{\mathcal{G}_2} \alpha X \beta \Rightarrow^*_{\mathcal{G}_2} w$, kde w je terminální řetěz. Tudíž X není nepoužitelný v $\mathcal{G}_2$. □

Příklad

$S \to AB \mid a$
$A \to a$