

11. Gramatiky typu 0 a 1

Formální jazyky a automaty

Jiří Balun

Obsah

1 Kontextové gramatiky (CSG)

- Popis CSG
- Vztah CSG a ostatních gramatik

2 Gramatiky bez omezení

- Popis gramatik typu 0
- Nezkracující gramatiky

3 Existence dalších jazyků

- Nespočetnost 2^{Σ^*}
- Existence jazyka, který není REL

Kontextové gramatiky (CSG)

Kontextová gramatika (typ 1)

Intuice: kontextové gramatiky (oproti CFG) umožňují podmínit aplikaci pravidla podřetězci, které jsou v okolí přepisovaného neterminálu.

Definice

Kontextová gramatika (CSG z *context-sensitive grammar*) má všechna pravidla v jednom z těchto tvarů:

- 1 $\alpha A \beta \rightarrow \alpha \gamma \beta$, kde $A \in N$, $\alpha, \beta, \gamma \in (\Sigma \cup N)^*$ a zároveň $\gamma \neq \varepsilon$,
- 2 $S \rightarrow \varepsilon$, pokud se S nevyskytuje na pravé straně jakéhokoliv pravidla.

- levá strana pravidel může obsahovat kromě přepisovaného neterminálu i tzv. **kontext**:
 - v definici je kontext pravidla $\alpha A \beta \rightarrow \alpha \gamma \beta$ dán řetězcí α a β
 - celý řetězec na levé straně pravidla podmiňuje aplikování daného pravidla
 - kontext nemůže být daným pravidlem přepsán (je zachován i na pravé straně pravidla)
- CSG generují **kontextové jazyky** (CSL z *context-sensitive language*):
 - jsou nadtřídou bezkontextových jazyků
 - CSL jsou uzavřené na sjednocení, průnik, zřetězení, Kleeneho uzávěr i doplněk
- problém náležení řetězce do jazyka CSG je rozhodnutelný

Příklad: kontextová gramatika

Příklad

CSG $\mathcal{G} = (\{S, A, B, C, D\}, \{a, b, c\}, P, S)$ s jazykem $L(\mathcal{G}) = \{a^m b^m c^n \mid 1 < n < m\}$, který je kontextový (a zároveň není bezkontextový):

$S \rightarrow aaABBcc$	$S \Rightarrow aa$ A B B cc	$A \rightarrow ab$
$A \rightarrow aAB \mid ab$	$\Rightarrow aaa$ b B B cc	$bB \rightarrow bC$
$bB \rightarrow bC$	$\Rightarrow aaa$ b C B cc	$bC \rightarrow bD$
$CB \rightarrow CC$	$\Rightarrow aaa$ b D B cc	$bD \rightarrow bb$
$bC \rightarrow bD$	$\Rightarrow aaab$ b B cc	$bB \rightarrow bC$
$bD \rightarrow bb$	$\Rightarrow aaab$ b C cc	$bC \rightarrow bD$
$DC \rightarrow DB$	$\Rightarrow aaab$ b D cc	$bD \rightarrow bb$
$BC \rightarrow BB \mid BBc$	$\Rightarrow aaabbbcc$	

Příklady: kontextové jazyky

Příklad

Jazyk $L_p = \{a^p \mid p \text{ je prvočíslo}\}$ řetězců, které mají délku prvočísel.

Příklad

Jazyk $L_{exp} = \{a^{2^n} \mid n \in \mathbb{N}\}$ řetězců, které mají délku mocnin čísla 2.

Příklad

Jazyk $L_{mul} = \{a^m b^n c^{mn} \mid m, n \in \mathbb{N}\}$, kde počet c je násobkem počtu znaků a a b .

Příklad

Jazyk $L_{rep} = \{w^{|w|} \mid w \in \Sigma^*\}$, jehož řetězce jsou n -té mocniny všech možných řetězců w nad Σ , kde n je délka daného řetězce w .

Vztah CSG a ostatních gramatik

Věta

Každá regulární gramatika je zároveň kontextová.

Důkaz

Z definice regulární gramatiky triviálně platí, že její pravidla zachovávají kontext. ☐

Věta

Ke každé CFG $\mathcal{G}$ existuje ekvivalentní gramatika $\mathcal{G}'$, která je zároveň kontextová

Důkaz

Všechna pravidla CFG $\mathcal{G} = (N, \Sigma, P, S)$ jsou ve tvaru $A \rightarrow \gamma$, kde $\gamma \in (\Sigma \cup N)^*$:

- pokud $\gamma \neq \varepsilon$, pak pravidlo zachovává kontext triviálně,
- je to případ kontextového pravidla $\alpha A \beta \rightarrow \alpha \gamma \beta$, kde $\alpha = \varepsilon$ a $\beta = \varepsilon$,
- jediným problémem jsou proto ε -pravidla, tj. pravidla ve tvaru $A \rightarrow \varepsilon$ (kde $A \neq S$),
- ε -pravidla odstraníme algoritmem z převodu CFG do Chomského normální formy. ☐

Gramatiky bez omezení

Gramatika bez omezení (typ 0)

Intuice: naše definice gramatiky vyžaduje, aby každé pravidlo bylo **generativní**, tj. pravidlo má na levé straně alespoň jeden neterminál.

Poznámka: některé definice gramatik povolují i negenerativní pravidla, ale síla gramatiky se tím nezmění, proto je uvažovat nebudeme.

Definice

Každá gramatika, jejíž pravidla jsou generativní, je **gramatika bez omezení** (z našeho pohledu tuto definici splňuje každá korektně definovaná gramatika).

- generují **rekurzivně spočetné jazyky** (REL z *recursively enumerable language*):
 - třída REL je nadtřídou kontextových jazyků
 - jsou uzavřené na sjednocení, průnik, zřetězení a Kleeneho uzávěr
 - nejsou uzavřené na doplněk a množinový rozdíl
- problém náležení řetězce do jazyka gramatiky typu 0 obecně není rozhodnutelný
 - pro řetězec, který gramatika negeneruje, se může ověřování zacyklit
 - problém je rozhodnutelný pouze pro tzv. **rekurzivní jazyky** (jsou podmnožinou REL)

Příklad: gramatika bez omezení

Příklad

Gramatika $\mathcal{G} = (\{S, A, B\}, \{a, b, c\}, P, S)$ s jazykem $L(\mathcal{G}) = \{a^n b^n c^n \mid n \geq 1\}$:

$S \rightarrow abc \mid aAbc$	$S \Rightarrow aAbc$	$Ab \rightarrow bA$
$Ab \rightarrow bA$	$\Rightarrow abAc$	$Ac \rightarrow Bbcc$
$Ac \rightarrow Bbcc$	$\Rightarrow abBbcc$	$bB \rightarrow Bb$
$bB \rightarrow Bb$	$\Rightarrow aBbbcc$	$aB \rightarrow aa$
$aB \rightarrow aa \mid aaA$	$\Rightarrow aabbcc$	

- *Poznámka:* $L(\mathcal{G})$ je rekurzivně spočetný jazyk, který je zároveň kontextový ($\mathcal{G}$ je tzv. **nezkracující**, proto lze ji převést na CSG).

Příklady: rekurzivně spočetné jazyky

Příklad

Jazyk vhodně zakódovaných (například binárně) ekvivalentních deterministických zásobníkových automatů je REL, jehož problém náležení **je rozhodnutelný**:

$$DPDA_{eq} = \{ \langle A_1, A_2 \rangle \mid A_1 \text{ a } A_2 \text{ jsou DPDA takové, že } L(A_1) = L(A_2) \}.$$

Příklad

Jazyk neekvivalentních bezkontextových gramatik, jehož náležení **není rozhodnutelné**:

$$CFG_{neq} = \{ \langle \mathcal{G}_1, \mathcal{G}_2 \rangle \mid \mathcal{G}_1 \text{ a } \mathcal{G}_2 \text{ jsou CFG takové, že } L(\mathcal{G}_1) \neq L(\mathcal{G}_2) \}.$$

Příklad

Problém zastavení je obecně definovaný pro Turingův stroj (více v kurzu o *vyčíslitelnosti*). Pro jednoduchost proto uvažujme jazyk programů jazyka C nad abecedou ASCII, které vždy ukončí svůj výpočet. Náležení do tohoto jazyka **není rozhodnutelné**:

$$HALT_C = \{ \langle c, w \rangle \mid c \text{ je zdrojový kód jazyka C, který pro vstup } w \text{ zastaví} \}.$$

Nezkracující gramatiky

Definice

Pravidlo $\alpha \rightarrow \beta$ je **nezkracující**, pokud platí $|\alpha| \leq |\beta|$. Gramatika $\mathcal{G} = (N, \Sigma, P, S)$ je nezkracující, pokud jsou nezkracující všechny její pravidla kromě $S \rightarrow \varepsilon$.

Lemma

Ke každé gramatice $\mathcal{G}$ existuje gramatika $\mathcal{G}'$, ve které jsou všechna pravidla obsahující terminální symboly ve tvaru $A \rightarrow a$, kde $A \in N$ a $a \in \Sigma$.

Důkaz

Pravidla z gramatiky $\mathcal{G}$ transformujeme v $\mathcal{G}'$ tímto způsobem:

- 1 pro každý terminál $a \in \Sigma$ zavedeme nový neterminál N_a a pomocné pravidlo $N_a \rightarrow a$,
- 2 každý výskyt terminálu a v pravidlech $\mathcal{G}$ nahradíme za N_a .

Například pravidlo $Cb \rightarrow aBb$ se změní na $CN_b \rightarrow N_aBN_b$, kde $N_a \rightarrow a$ a $N_b \rightarrow b$ jsou nová pomocná pravidla $\mathcal{G}'$. □

Kontextové a nezkracující gramatiky

Věta

Ke každé nezkracující gramatice $\mathcal{G}$ existuje CSG $\mathcal{G}'$ taková, že $L(\mathcal{G}) = L(\mathcal{G}')$.

Důkaz

Předpokládejme, že pravidla $\mathcal{G}$ jsou buď ve tvaru $A \rightarrow a$ (z předchozího lemma), nebo $A_1 A_2 \dots A_m \rightarrow B_1 B_2 \dots B_n$, kde $A_1, \dots, A_m, B_1, \dots, B_n \in N$ a navíc $m \leq n$.

Tyto pravidla nahradíme s pomocí nových neterminálů $X_1, \dots, X_m$ za kontextová pravidla:

$$A_1 A_2 \dots A_m \rightarrow X_1 A_2 \dots A_m$$

$$X_1 A_2 \dots A_m \rightarrow X_1 X_2 \dots A_m$$

$$\vdots$$

$$X_1 X_2 \dots X_{m-1} A_m \rightarrow X_1 X_2 \dots X_{m-1} X_m B_{m+1} \dots B_n$$

$$X_1 X_2 \dots X_m B_{m+1} \dots B_n \rightarrow B_1 X_2 \dots X_m B_{m+1} \dots B_n$$

$$\vdots$$

$$B_1 B_2 \dots B_{m-1} X_m B_{m+1} \dots B_n \rightarrow B_1 B_2 \dots B_{m-1} B_m B_{m+1} \dots B_n$$



Příklad: nezkracující gramatika

Příklad

CSG $\mathcal{G}' = (\{S, A, B, X_1, X_2, X_3, X_4, N_a, N_b, N_c\}, \{a, b, c\}, P, S)$ s jazykem $L(\mathcal{G}') = \{a^n b^n c^n \mid n \geq 1\}$, kterou získáme z předchozího příkladu gramatiky bez omezení:

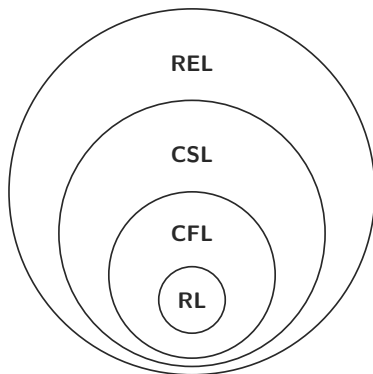
$$\begin{array}{ll} S \rightarrow N_a N_b N_c \mid N_a A N_b N_c & N_b B \rightarrow X_3 B \\ A N_b \rightarrow X_1 N_b & X_3 B \rightarrow X_3 X_4 \\ X_1 N_b \rightarrow X_1 X_2 & X_3 X_4 \rightarrow B X_4 \\ X_1 X_2 \rightarrow N_b X_2 & B X_4 \rightarrow B N_b \\ N_b X_2 \rightarrow N_b A & N_a \rightarrow a \\ A N_c \rightarrow B N_b N_c N_c & N_b \rightarrow b \\ N_a B \rightarrow N_a N_a \mid N_a N_a A & N_c \rightarrow c \end{array}$$

Průběh transformace na CSG:

- nejprve nahradíme každý výskyt terminálu $\sigma \in \Sigma$ za pomocný neterminál N_σ ,
- poté stačí nahradit dle důkazu předchozí věty pravidla, které nezachovávají kontext.

Existence dalších jazyků

Chomského hierarchie



- Je tu něco?

Existuje jazyk, který není rekurzivně spočetný?

- otázku lze formulovat jako: existuje jazyk, který nelze generovat žádnou gramatikou?
- odpověď je **ano**, například jazyk dvojic ekvivalentních CFG (doplňek CFG_{neq}):

$$CFG_{eq} = \{ \langle \mathcal{G}_1, \mathcal{G}_2 \rangle \mid \mathcal{G}_1 \text{ a } \mathcal{G}_2 \text{ jsou takové CFG, že } L(\mathcal{G}_1) = L(\mathcal{G}_2) \}$$

Nespočetnost 2^{Σ^*}

Věta

Třída všech jazyků nad abecedou Σ , kterou značíme 2^{Σ^*} , je nespočetná.

Důkaz

Z definice abecedy je Σ konečná množina, a proto je Kleeneho uzávěr Σ^* spočetně nekonečná množina a lze jej seřadit pomocí shortlex uspořádat, tj. $\Sigma^* = \{w_1, w_2, \dots\}$.

Důkaz vedeme sporem, proto předpokládejme, že 2^{Σ^*} je spočetná množina:

- potom by prvky (jazyky) z 2^{Σ^*} měli jít nějak uspořádat, tj. $2^{\Sigma^*} = \{L_1, L_2, L_3, \dots\}$,
- 2^{Σ^*} obsahuje i jazyk definovaný jako $L_{diag} = \{w_i \mid w_i \notin L_i\}$,
- řekněme, že existuje index j takový, že platí $L_{diag} = L_j$,
- pak ale $w_j \in L_{diag}$ právě tehdy, když $w_j \notin L_j$, což je spor, protože $L_{diag} = L_j$.

Všimněte si, že tvrzení platí i pro případ, kdy je Σ jednoprvková.



Nespočetnost 2^{Σ^*}

	w_1	w_2	w_3	$\dots$	w_j	$\dots$
L_1	✓	✗	✗	$\dots$	✓	$\dots$
L_2	✓	✗	✓	$\dots$	✗	$\dots$
L_3	✗	✓	✓	$\dots$	✓	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
$L_j = L_{diag}$	✗	✓	✗	$\dots$	??	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	

Reprezentace důkazu tabulkou:

- řádek pro $L_j = L_{diag} = \{w_i \mid w_i \notin L_i\}$ je definován jako negace prvků na diagonále
- ve sloupci pro slovo w_j dostáváme spor, protože $w_j \in L_j$ právě tehdy, když $w_j \notin L_j$

Existence jazyka, který není REL

Věta

Nad jednoprvkovou abecedou $\Sigma = \{a\}$ existuje jazyk, který není generován gramatikou.

Důkaz

Mějme abecedu $\Gamma = \{a, N, \rightarrow, \{, \}, (,), , \}$ (i symbol čárky je součástí Γ), pomocí které jsme schopni zakódovat libovolnou gramatiku nad abecedou Σ :

- neterminály kódujeme unárně jako $N_1 = N$, $N_2 = NN$, $N_3 = NNN$, ...
- příklad řetězce nad Γ je gramatika $(\{N\}, \{a\}, \{N \rightarrow a\}, N)$ popisující jazyk $\{a\}$,
- Γ^* je spočetná (stejně jako Σ^*), proto existuje spočetně mnoho gramatik nad Σ ,
- každá gramatika zakódovaná pomocí Γ popisuje právě jeden jazyk nad Σ ,
- proto lze pomocí Γ reprezentovat jen spočetně mnoho jazyků nad abecedou Σ .

Třída jazyků 2^{Σ^*} je ale nespočetná, proto musí existovat jazyk $L \subseteq \Sigma^*$, který nelze generovat gramatikou z Γ^* . □