AGH

AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Gramatyki, wyprowadzenia, hierarchia Chomsky'ego

Teoria automatów i języków formalnych

Dr inż. Janusz Majewski
Katedra Informatyki

AGH

Gramatyka

Gramatyką *G* nazywamy czwórkę uporządkowaną

$G = \langle V, \Sigma, P, S \rangle$

gdzie:

- V – skończony zbiór symboli nieterminalnych (zmiennych językowych),
- Σ – zbiór (alfabet) symboli terminalnych,
- P – skończony zbiór produkcji, z których każda ma postać $\alpha \rightarrow \beta$
- $S \in V$ – wyróżniony symbol początkowy (nieterminal)

przy czym:

$P \subseteq (V \cup \Sigma)^+ \times (V \cup \Sigma)^*$

$P = \{\alpha \rightarrow \beta \mid \alpha \in (V \cup \Sigma)^+, \beta \in (V \cup \Sigma)^* \}$

Czasami produkcje definiuje się inaczej:

$P \subseteq (V \cup \Sigma)^* V (V \cup \Sigma)^* \times (V \cup \Sigma)^*$

$P = \{\alpha \rightarrow \beta \mid \alpha \in (V \cup \Sigma)^* V (V \cup \Sigma)^*, \beta \in (V \cup \Sigma)^* \}$

Wyprowadzalność (1)

Słowo ψ jest wyprowadzalne bezpośrednio ze słowa ω w gramatyce G, co zapisujemy

$\omega \Rightarrow_G \psi$

jeżeli:

$\omega = \gamma\alpha\delta$

$\psi = \gamma\beta\delta$

$(\alpha \rightarrow \beta) \in P$

$\alpha, \beta, \gamma, \delta, \psi, \omega \in (V \cup \Sigma)^*$

Wyprowadzalność (2)

Słowo ψ jest wyprowadzalne ze słowa ω w gramatyce G, co zapisujemy

$\omega \Rightarrow_G^+ \psi$

jeżeli istnieją $\varphi_0, \varphi_1, \ldots, \varphi_n \in (V \cup \Sigma)^*$ takie, że:

$\varphi_0 = \omega$

$\varphi_n = \psi$

$\varphi_{i-1} \Rightarrow_G \varphi_i$ dla i = 1, 2, ..., n

Sekwencję $\varphi_0, \varphi_1, \ldots, \varphi_n$ nazywamy wyprowadzeniem o długości n.

Definiujemy ponadto:

$(\omega \Rightarrow_G^* \psi) \Leftrightarrow (\omega \Rightarrow_G^+ \psi) \vee (\omega = \psi)$

Relacje $\Rightarrow_G^+$ oraz $\Rightarrow_G^*$ są odpowiednio przechodnim oraz przechodnim i zwrotnym domknięciem relacji bezpośredniej wyprowadzalności $\Rightarrow_G$. Jeżeli wiadomo, o jaką gramatykę chodzi, pomijamy dolny indeks „G" w oznaczeniu tych relacji pisząc po prostu: $\Rightarrow^+$, $\Rightarrow^*$ oraz $\Rightarrow$.

Język generowany przez gramatykę, forma zdaniowa

Język generowany przez gramatykę

Gramatyka jest jednym ze sposobów definiowania języka formalnego. Mając daną gramatykę G oznaczamy przez $L(G)$ zbiór wszystkich słów, które mogą być w tej gramatyce wyprowadzone z symbolu początkowego S. Zbiór ten nazywamy językiem generowanym przez daną gramatykę.

$L(G) = \{ x \in \Sigma^* \mid S \Rightarrow_G^* x \}$

Forma zdaniowa

Łańcuch $x \in (V \cup \Sigma)^*$ nazywamy *formą zdaniową* gramatyki G, jeśli można go wyprowadzić z symbolu początkowego S.

$x \in (V \cup \Sigma)^*$ jest *formą zdaniową* $\Leftrightarrow$ $S \Rightarrow_G^* x$

Uwaga: terminu *słowo* używamy w rozumieniu łańcucha zbudowane wyłącznie z symboli terminalnych

$x \in \Sigma^*$ jest *słowem* $\Leftrightarrow$ $S \Rightarrow_G^* x$

Hierarchia Chomsky'ego, gramatyki bez ograniczeń

Hierarchia Chomsky'ego

Noam Chomsky zdefiniował cztery klasy gramatyk oraz cztery klasy języków formalnych. Klasy te numerowane są od *0* do *3*.

Klasa 0

Gramatykę $G = \langle V, \Sigma, P, S \rangle$, w której produkcje mają postać $\alpha \rightarrow \beta$, gdzie α i β są dowolnymi łańcuchami symboli tej gramatyki, przy czym $\alpha \neq \varepsilon$ nazywamy *semi-gramatykami Thuego, gramatykami bez ograniczeń, gramatykami struktur frazowych, gramatykami kombinatorycznymi* lub *gramatykami klasy „0"*.

Definicja gramatyk klasy „0", jak widać, nie nakłada żadnych ograniczeń na postać produkcji gramatyki w stosunku do ogólnej definicji gramatyki.

Języki generowane przez gramatyki tego typu noszą nazwę *języków rekurencyjnie przeliczalnych*.

Przez G_{KOMB} oznaczymy klasę gramatyk kombinatorycznych, a przez L_{RP} klasę języków rekurencyjnie przeliczalnych.

Problem

Termin „problem” w uproszczeniu oznacza pytanie związane z jakimś wystąpieniem pewnych obiektów z pewnych klas (u nas tymi obiektami są np. dowolne gramatyki pewnego typu oraz dowolne słowa nad alfabetem definiowanym przez te gramatyki, zaś wystąpieniem obiektu będzie konkretne słowo i konkretna gramatyka), na które to pytania można udzielić odpowiedzi: TAK lub NIE. Termin „nierozstrzygalny” w uproszczeniu znaczy tyle: „nie istnieje jednoznaczny deterministyczny algorytm, który dla każdego wystąpienia danego problemu w skończonej liczbie kroków dałby odpowiedź TAK, jeżeli poprawna odpowiedź na pytanie związane z wystąpieniem rozważanego problemu brzmi TAK, oraz NIE, gdy poprawna odpowiedź brzmi NIE”. Termin „rozstrzygalny” w uproszczeniu znaczy tyle: „istnieje jednoznaczny deterministyczny algorytm, który dla każdego wystąpienia danego problemu w skończonej liczbie kroków dałby odpowiedź TAK, jeżeli poprawna odpowiedź na pytanie związane z wystąpieniem rozważanego problemu brzmi TAK, oraz NIE, gdy poprawna odpowiedź brzmi NIE”.

Problem przynależności słowa do języka generowanego przez daną gramatykę

Fundamentalny problem, który będzie później naszym głównym przedmiotem zainteresowania, mianowicie: „czy słowo x należy do języka generowanego przez daną gramatykę”, jest nierozstrzygalny dla języków generowanych przez gramatyki kombinatoryczne.

Problem: czy $x \in L(G)$ jest nierozstrzygalny dla $G \in \mathcal{G}_{KOMB}$.

Języki i gramatyki kontekstowe (1)

Klasa 1

Gramatykę $G = <V, \Sigma, P, S>$, w której produkcje mają postać $\alpha \rightarrow \beta$, gdzie α i β są takimi łańcuchami symboli tej gramatyki, że łańcuch β jest przynajmniej tak długi jak łańcuch α $(|\alpha| \leq |\beta|)$ oraz dodatkowo dopuszczona jest produkcja $S \rightarrow \varepsilon$, jeśli język zawiera słowo puste, nazywamy *gramatykami kontekstowymi, gramatykami monotonicznym, gramatykami nieskracającymi* lub *gramatykami klasy „1”*.

Termin „kontekstowy” pochodzi od tego, że dla każdej gramatyki monotonicznej można znaleźć równoważną jej (tzn. generującą ten sam język) gramatykę, której produkcje (z wyjątkiem ewentualnej produkcji $S \rightarrow \varepsilon$) mają postać $\alpha_1 A \alpha_2 \rightarrow \alpha_1 \beta \alpha_2$ gdzie A jest nieterminalem $(A \in V)$, zaś α_1, α_2, β są dowolnymi łańcuchami symboli gramatyki, przy czym $\beta \neq \varepsilon$. Produkcje o tej postaci pozwalają na zastąpienie nieterminala A łańcuchem β tylko w „lewostronnym kontekście” α_1 i „prawostronnym kontekście” α_2.

Języki i gramatyki kontekstowe (2)

Języki generowane przez gramatyki tego typu noszą nazwę *języków kontekstowych*.

Przez G_K oznaczymy klasę gramatyk kontekstowych, a przez L_K klasę języków kontekstowych.

Problem: czy $x \in L(G)$ jest rozstrzygalny dla $G \in G_K$.

Ponadto:

$G_K \subset G_{KOMB}$

$L_K \subset L_{RP}$

Przykład

Język $\{a^m b^m c^m d^m \mid m > 0\}$ jest generowany przez przykładowe gramatyki: klasy 0 i klasy 1

$S \rightarrow abcDd$
$cD \rightarrow Dc$
$bD \rightarrow Db$
$aD \rightarrow aaA$
$Ab \rightarrow bA$
$bAc \rightarrow bbcA$
$cAc \rightarrow ccA$
$cAd \rightarrow ccDdd$
$D \rightarrow \varepsilon$

$S \rightarrow abF$
$cD \rightarrow Dc$
$bD \rightarrow Db$
$aD \rightarrow aaA$
$Ab \rightarrow bA$
$bAc \rightarrow bbcA$
$cAc \rightarrow ccA$
$cAd \rightarrow cFd$
$F \rightarrow cDd$
$F \rightarrow cd$

Języki i gramatyki bezkontekstowe

Klasa 2 - klasa gramatyk bezkontekstowych jest chyba najważniejszą (z naszego punktu widzenia) klasą gramatyk, gdyż za pomocą gramatyk tej klasy opisuje się składnię większości języków programowania.

Gramatykę $G = \langle V, \Sigma, P, S \rangle$, w której produkcje mają postać $A \rightarrow \beta$, gdzie A jest nieterminalem $(A \in V)$, zaś łańcuch β jest dowolnym łańcuchem symboli tej gramatyki nazywamy *gramatykami bezkontekstowymi* lub *gramatykami klasy „2"*. Termin „bezkontekstowy" pochodzi od tego, że produkcje takiej gramatyki pozwalają na bezwarunkowe (bez uwzględniania kontekstu) zastąpienie nieterminala A łańcuchem β.

Języki generowane przez gramatyki tego typu noszą nazwę *języków bezkontekstowych*.

Przez G_{BK} oznaczymy klasę gramatyk kontekstowych, a przez L_{BK} klasę języków kontekstowych.

Problem: czy $x \in L(G)$ jest rozstrzygalny dla $G \in G_{BK}$.

Ponadto:

$G_{BK} \subset G_K \subset G_{KOMB}$
$L_{BK} \subset L_K \subset L_{RP}$

Języki i gramatyki regularne (1)

Klasa 3

Gramatykę $G = <V, \Sigma, P, S>$, w której każda produkcja ma postać $A \rightarrow xB$ lub $A \rightarrow x$ gdzie A i B są nieterminalami $(A,B \in V)$, zaś łańcuch x jest dowolnym łańcuchem symboli terminalnych tej gramatyki $(x \in \Sigma^*)$ nazywamy *gramatyką prawostronnie liniową*. Gramatykę $G = <V, \Sigma, P, S>$, w której każda produkcja ma postać $A \rightarrow Bx$ lub $A \rightarrow x$ gdzie A i B są nieterminalami $(A,B \in V)$, zaś łańcuch x jest dowolnym łańcuchem symboli terminalnych tej gramatyki $(x \in \Sigma^*)$ nazywamy *gramatyką lewostronnie liniową*. Gramatyki prawostronnie liniowe i lewostronnie liniowe nazywamy *gramatykami liniowymi*, *gramatykami regularnymi lub gramatykami klasy „3”*.

Języki generowane przez gramatyki tego typu noszą nazwę *języków regularnych*.

Przez $\mathcal{G}_{RG}$ oznaczymy klasę gramatyk regularnych, a przez $\mathcal{L}_{RG}$ klasę języków regularnych.

Języki i gramatyki regularne (2)

Problem: czy $x \in L(G)$ jest rozstrzygalny dla $G \in \mathcal{G}_{RG}$.

Ponadto:

$$\mathcal{G}_{RG} \subset \mathcal{G}_{BK} \subset \mathcal{G}_{K} \subset \mathcal{G}_{KOMB}$$

$$\mathcal{L}_{RG} \subset \mathcal{L}_{BK} \subset \mathcal{L}_{K} \subset \mathcal{L}_{RP}$$

Klasa gramatyk regularnych jest także bardzo ważną (z naszego punktu widzenia) klasą gramatyk, gdyż za pomocą gramatyk tej klasy opisuje się składnię większości podstawowych elementów leksykalnych (słownikowych) języków programowania (takich jak identyfikatory, stałe numeryczne, stałe tekstowe, komentarze, operatory, itd.

Przykład (1)

Przykład:

Rozważymy gramatykę $G = <V, \Sigma, P, S>$, w której:

$V = \{S, A, B, C, D, E, F, G\}$

$\Sigma = \{a, b, c\}$

$P = \{$ $S \rightarrow AbC \mid aD \mid AE \mid aBc \mid abc$
$A \rightarrow a$
$B \rightarrow b$
$bC \rightarrow bc$
$D \rightarrow bc$
$aE \rightarrow abFcG$
$F \rightarrow \varepsilon$
$bcG \rightarrow bc$ $\}$

$S = S$

Ta gramatyka jest gramatyką kombinatoryczną (klasy „0" – w lewych stronach produkcji występują dowolne łańcuchy symboli, są dwie produkcje skracające) i równocześnie nie jest gramatyką żadnej węższej klasy. Język przez nią generowany jest oczywiście językiem rekurencyjnie przeliczalnym.

Przykład (2)

Zbadajmy, jakie słowa są generowane przez tę gramatykę.

$S \Rightarrow AbC \Rightarrow abC \Rightarrow abc$

$S \Rightarrow aD \Rightarrow abc$

$S \Rightarrow AE \Rightarrow aE \Rightarrow abFcG \Rightarrow abcG \Rightarrow abc$

$S \Rightarrow aBc \Rightarrow abc$

$S \Rightarrow abc$

Widać, że jedynym słowem generowanym przez tę gramatykę jest *abc*. Dla języka

$L = \{ abc \}$

można zbudować znacznie prostszą gramatykę $G_1 = <V_1, \Sigma, P_1, S>$, w której:

$V_1 = \{S\}$

$\Sigma = \{a, b, c\}$

$P_1 = \{ S \rightarrow abc \}$

$S = S$

Gramatyka G_1 należy do klasy „3" gramatyk regularnych. Ponieważ

$L = L(G) = L(G_1)$

więc język L należy do klasy języków regularnych, mimo że może być wygenerowany przez gramatykę znacznie szerszej klasy (gramatykę bez ograniczeń).

Uwaga: zawsze interesować nas będzie najwęższa klasa, do której należy badany język.