



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Języki i operacje na językach

Teoria automatów i języków formalnych

Dr inż. Janusz Majewski
Katedra Informatyki

Definicja języka

Definicja języka

Niech Σ będzie alfabetem, Σ^* - zbiorem wszystkich łańcuchów nad alfabetem Σ .

Dowolny podzbiór L zbioru Σ^* nazywamy językiem L nad alfabetem Σ .

$$L \subseteq \Sigma^*$$

Przykłady:

$$L_0 = \emptyset$$

- język pusty

$$L_1 = \{\varepsilon\}$$

- język zawierający tylko słowo puste

$$L_2 = \Sigma^*$$

- język zawierający wszystkie słowa nad alfabetem Σ

$$L_3 = \{\varepsilon, 0, 01, 001\}$$

- język zawierający skończoną liczbę słów

$$L_4 = \{0, 01, 011, 0111, \dots\} = \{01^n \mid n \geq 0\}$$

- język nieskończony

Operacje na językach (1)

Niech L , L_1 i L_2 będą językami odpowiednio nad alfabetami Σ , Σ_1 i Σ_2 .

$$L \subseteq \Sigma^* \quad L_1 \subseteq \Sigma_1^* \quad L_2 \subseteq \Sigma_2^*$$

Najczęściej wykorzystuje się następujące operacje na językach:

Suma teoriomnogościowa

$$L_1 \cup L_2 = \{x \mid x \in L_1 \vee x \in L_2\}$$

Złożenie języków

$$L_1 L_2 = \{x_1 x_2 \mid x_1 \in L_1 \wedge x_2 \in L_2\}$$

Domknięcie Kleene'a (gwiazdka Kleene'a) L^*

$$L^0 = \{\varepsilon\}$$

$$L^1 = L$$

$$L^2 = L^1 L$$

.....

$$L^n = L^{n-1} L$$

$$L^* = L^0 \cup L^1 \cup L^2 \cup L^3 \cup \dots$$

$$L^+ = L^1 \cup L^2 \cup L^3 \cup \dots$$

Operacje na językach (2)

Rozpatruje się także operacje przecięcia (iloczynu teoriomnogościowego), dopełnienia, podstawienia, homomorfizmu i ilorazu.

Przecięcie (iloczyn teoriomnogościowy)

$$L_1 \cap L_2 = \{ x \mid x \in L_1 \wedge x \in L_2 \}$$

Dopełnienie języka L względem Σ^*

$$\bar{L} = \Sigma^* - L$$

Iloraz języków

Niech będą dane dwa języki: $L_1 \subseteq \Sigma^*$, $L_2 \subseteq \Sigma^*$. Definiujemy iloraz L_1/L_2 tych języków jako:

$$L_1/L_2 = \{ x \mid (\exists y \in L_2) (xy \in L_1) \}$$

Operacje na językach (3)

Przypomnienie definicji ilorazu

$$L_1/L_2 = \{ x \mid (\exists y \in L_2) (xy \in L_1) \}$$

Rozważamy języki:

- $L_1 = \{0^n10^m \mid m \geq 0, n \geq 0\} =$
 $= \{1, 01, 10, 001, 010, 100, 0001, 0010, 0100, 1000, \dots\}$
- $L_2 = \{10^n1 \mid n \geq 0\} = \{11, 101, 1001, 10001, \dots\}$

$L_1/L_2 = \emptyset$ gdyż każdy łańcuch $y \in L_2$ zawiera dwie jedynki, a każdy łańcuch $xy \in L_1$ może zawierać tylko jedną jedynkę, więc nie istnieje łańcuch x , taki że $xy \in L_1$ i $y \in L_2$.

Operacje na językach (4)

Przypomnienie definicji ilorazu

$$L_1/L_2 = \{ x \mid (\exists y \in L_2) (xy \in L_1) \}$$

Rozważamy języki:

- $L_1 = \{0^n10^m \mid m \geq 0, n \geq 0\} =$
 $= \{1, 01, 10, 001, 010, 100, 0001, 0010, 0100, 1000, \dots\}$
- $L_3 = \{0^n1 \mid n \geq 0\} = \{1, 01, 001, 0001, \dots\}$

$L_1/L_3 = \{0^n \mid n \geq 0\} = \{\varepsilon, 0, 00, 000, \dots\}$ gdyż w rachubę wchodzi tylko słowa $1, 01, 001, 0001, \dots$ z L_1 i wszystkie słowa z L_3 (choć wystarczyłoby tylko słowo 1).

Operacje na językach (5)

Przypomnienie definicji ilorazu

$$L_1/L_2 = \{ x \mid (\exists y \in L_2) (xy \in L_1) \}$$

Rozważamy języki:

- $L_2 = \{10^n1 \mid n \geq 0\} = \{11, 101, 1001, 10001, \dots\}$
 - $L_3 = \{0^n1 \mid n \geq 0\} = \{1, 01, 001, 0001, \dots\}$
- $L_2/L_3 = \{10^n \mid n \geq 0\} = \{1, 10, 100, 1000, \dots\}$.**

Podstawienia (1)

Podstawienie f jest odwzorowaniem alfabetu Σ na podzbiory zbioru Γ^* dla pewnego alfabetu Γ . Zatem f przyporządkowuje każdemu symbolowi z Σ pewien język.

$$f: \Sigma \mapsto 2^{\Gamma^*}$$

Odwzorowanie f rozszerzamy na łańcuchy

$$f: \Sigma^* \mapsto 2^{\Gamma^*}$$

w następujący sposób:

$$(1) \quad f(\varepsilon) = \{\varepsilon\}$$

$$(2) \quad f(xa) = f(x)f(a)$$

Wreszcie odwzorowanie f rozszerzamy na zbiory łańcuchów, czyli na języki

$$f: 2^{\Sigma^*} \mapsto 2^{\Gamma^*}$$

definiując:

$$f(L) = \bigcup_{x \in L} f(x)$$

Podstawienia (2)

Przykład:

Niech

$$\Sigma = \{0, 1\}$$

$$\Gamma = \{a, b\}$$

$$f(0) = \{a\}$$

$$f(1) = \{b^n \mid n \geq 0\} = \{\varepsilon, b, bb, bbb, \dots\}$$

Wtedy dla łańcucha 010 mamy:

$$f(010) = \{a\} \{b^n \mid n \geq 0\} \{a\} = \{aa, aba, abba, abbba, \dots\} = \{ab^n a \mid n \geq 0\}$$

Niech

$$L = \{0^m 1 \mid m \geq 0\} = \{1, 01, 001, 0001, \dots\}$$

Wtedy

$$f(L) = \{a^m b^n \mid m \geq 0, n \geq 0\} =$$

$$= \{\varepsilon, b, bb, bbb, \dots, a, ab, abb, abbb, \dots, aa, aab, aabb, aabbb, \dots, aaa, aaab, aaabb, \dots\}$$

Homomorfizmy (1)

Homomorfizmem h nazywany takie podstawienie, które każdemu symbolowi alfabetu Σ przypisuje dokładnie jeden łańcuch ze zbioru Γ^* , czyli homomorfizm to odwzorowanie:

$$h: \Sigma \mapsto \Gamma^*$$

Rozszerzamy odwzorowanie h na łańcuchy

$$h: \Sigma^* \mapsto \Gamma^*$$

w taki sam sposób, jak to miało miejsce z podstawieniem:

$$(1) \quad h(\varepsilon) = \varepsilon$$

$$(2) \quad h(xa) = h(x)h(a)$$

Dalej rozszerzamy homomorfizm h na języki

$$h: 2^{\Sigma^*} \mapsto 2^{\Gamma^*}$$

w taki sam sposób, jak podstawienie

$$h(L) = \bigcup_{x \in L} h(x)$$



Homomorfizmy (2)

Definiujemy przeciwobraz homomorficzny $h^{-1}(x)$ łańcucha x jako:

$$h^{-1}(x) = \{y \mid h(y) = x\}$$

oraz przeciwobraz homomorficzny $h^{-1}(L)$ języka L jako:

$$h^{-1}(L) = \{x \mid h(x) \in L\}$$

Zachodzi przy tym:

$$h^{-1}(h(L)) \supseteq L$$

oraz:

$$h(h^{-1}(L)) \subseteq L$$

Homomorfizmy (3)

Przykład: Niech

$$\Sigma = \{0, 1, 2\} \quad \text{oraz} \quad \Gamma = \{a, b\}$$

$$h(0) = a$$

$$h(1) = aab$$

$$h(2) = ab$$

Wtedy dla łańcucha 012 mamy:

$$h(012) = aaabab$$

Niech: $L = \{01, 02\}$

Wtedy: $h(L) = \{aaab, aab\}$

Wyznaczmy $h^{-1}(h(L))$

$$h^{-1}(h(L)) = \{002, 01, 02, 1\} \neq L$$

Widzimy, że:

$$h^{-1}(h(L)) \supseteq L$$



Homomorfizmy (4)

Przykład:

Niech

$$\Sigma = \{0, 1\} \quad \text{oraz} \quad \Gamma = \{a, b\}$$

$$h(0) = aa$$

$$h(1) = aba$$

Niech

$$L = \{(ab)^n a \mid n \geq 0\} = \{a, aba, ababa, abababa, \dots\}$$

Wtedy

$$h^{-1}(L) = \{1\}$$

Wyznaczmy $h(h^{-1}(L))$

$$h(h^{-1}(L)) = \{aba\} \neq L$$

Widzimy, że:

$$h(h^{-1}(L)) \subseteq L$$

Przedrostki, przyrostki (1)

Niech $z \in L \subseteq \Sigma^*$ będzie słowem z języka L .

Przedstawimy z w postaci:

$$z = xy \quad x, y \in \Sigma^*$$

x nazywamy przedrostkiem (prefiksem) słowa z , zaś y nazywamy przyrostkiem (sufiksem) słowa z .

x nazywamy przedrostkiem właściwym słowa $z \Leftrightarrow y \neq \varepsilon$.

y nazywamy przyrostkiem właściwym słowa $z \Leftrightarrow x \neq \varepsilon$.

Przykład:

Rozważamy słowo $abbb$

Przedrostki tego słowa to: $\varepsilon, a, ab, abb, abbb$

Przedrostki właściwe tego słowa to: ε, a, ab, abb



Przedrostki, przyrostki (2)

Język L ma własność przedrostkową gdy:

$(\forall z \in L) (\forall s - \text{będącego przedrostkiem właściwym słowa } z \in L) (s \notin L)$

czyli język ma własność przedrostkową, jeśli żaden przedrostek właściwy słowa tego języka nie jest identyczny z żadnym słowem tego języka.

Język L ma własność przyrostkową gdy:

$(\forall z \in L) (\forall s - \text{będącego przyrostkiem właściwym słowa } z \in L) (s \notin L)$

czyli język ma własność przyrostkową, jeśli żaden przyrostek właściwy słowa tego języka nie jest identyczny z żadnym słowem tego języka.

Przedrostki, przyrostki (3)

Przykład :

$$L = \{10^n \mid n \geq 0\} = \{1, 10, 100, 1000, \dots\}$$

L nie posiada własności przedrostkowej, gdyż np. słowo *1000* ma przedrostek właściwy *10* będący słowem tego języka.

L posiada własność przyrostkową, gdyż wszystkie przyrostki właściwe słów tego języka mają postać $\{0^n \mid n \geq 0\}$, i żaden z nich nie jest identyczny z żadnym słowem tego języka.

Uporządkowanie słów należących do języka

- Zbiór słownikowy można uważać za zbiór liniowo lub dobrze uporządkowany, np. poprzez porządek leksykograficzny (Σ^*, \leq_L) lub standardowy (Σ^*, \leq_S) .
- W taki sam sposób można uporządkować słowa dowolnego języka $L \subseteq \Sigma^*$ (określając relację \leq na alfabecie Σ oraz redukując relację \leq_S lub \leq_L określoną na Σ^* do L). Mówimy wówczas o leksykograficznym lub standardowym porządku słów danego języka.
- Przykładem porządku leksykograficznego \leq_L dla skończonych zbiorów (języków) może być uporządkowanie słów w encyklopediach, słownikach, leksykonach – wówczas \leq jest powszechnie przyjętym uporządkowaniem liter w alfabecie pewnego języka naturalnego.

Moc zbioru wszystkich języków (1)

Lemat: Zbiór \mathcal{B} wszystkich nieskończonych ciągów zerojedynkowych jest nieprzeliczalny.

Założmy dla dowodu nie wprost, że zbiór wszystkich nieskończonych łańcuchów zerojedynkowych jest przeliczalny. Można więc te łańcuchy wypisać i ponumerować, na przykład tak:

numer	łańcuch
1	0 1100010010100...
2	1 0 001001001110...
3	01 1 10101010010...
4	111 0 1100111011...
...	...

Skonstruujemy łańcuch x różny od wszystkich wypisanych łańcuchów. Jeśli n -ty łańcuch ma na n -tej pozycji zero, to x będzie miał na n -tej pozycji jedynekę i na odwrót, jeśli n -ty łańcuch ma na n -tej pozycji jedynekę, to x będzie miał na n -tej pozycji zero. U nas $x = \mathbf{1101}...$

Łańcuch x jest różny co najmniej na jednej pozycji od każdego z wypisanych łańcuchów, wobec tego jest różny od każdego z wszystkich łańcuchów.

Doszliśmy do sprzeczności. Zbioru wszystkich nieskończonych łańcuchów zerojedynkowych nie da się ponumerować, jest to więc zbiór nieprzeliczalny. *(Jest to metoda diagonalizacji Cantora).*

Moc zbioru wszystkich języków (2)

Twierdzenie: Zbiór $\mathcal{L} = 2^{\Sigma^*}$ wszystkich języków (wszystkich podzbiorów zbioru Σ^* - zbioru słownikowego nad danym alfabetem Σ) jest nieprzeliczalny.

Pokażemy, że \mathcal{L} jest nieprzeliczalny konstruując bijekcję między \mathcal{B} (zbiorem wszystkich nieskończonych łańcuchów zerojedynkowych) i \mathcal{L} dowodzącą, że oba te zbiory są tej samej mocy. Ponumerujmy słowa z Σ^* (wiadomo, że można, np. stosując porządek standardowy): $\Sigma^* = \{s_1, s_2, s_3, \dots\}$. Każdy język $L \in \mathcal{L} = 2^{\Sigma^*}$ odpowiada unikalnemu ciągowi z \mathcal{B} - i-ty bit tego ciągu jest równy jeden wtedy i tylko wtedy, gdy $s_i \in L$, w przeciwnym przypadku bit ten jest równy zero. Taki ciąg nazywamy ciągiem charakterystycznym χ_L języka L .

Przykład: $\Sigma = \{a, b\}$

$\Sigma^* = \{ \varepsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, baa, \dots \}$

$L = \{ \quad a, \quad aa, ab, \quad bb, aaa, aab, \quad abb, \quad \dots \}$

$\chi_L = \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad \dots$

Odwzorowanie $f: \mathcal{L} \mapsto \mathcal{B}$, gdzie $f(L) = \chi_L$ jest bijekcją, zatem, ponieważ \mathcal{B} jest nieprzeliczalny, to \mathcal{L} także jest nieprzeliczalny.



Relacja indukowana przez język (1)

Relacja prawostronnie niezmiennicza

Relację $R \subseteq \Sigma^* \times \Sigma^*$ (gdzie Σ jest skończonym alfabetem symboli) nazywamy prawostronnie niezmienniczą wtedy i tylko wtedy, gdy

$$(\forall u, v \in \Sigma^*) (u R v \Rightarrow (\forall z \in \Sigma^*) uz R vz)$$

Przykładem relacji prawostronnie niezmienniczej jest relacja R_L indukowana przez język L

Relacja indukowana przez język

Relacją indukowaną przez język $L \subseteq \Sigma^*$ nazywamy relację $R_L \subseteq \Sigma^* \times \Sigma^*$ (gdzie Σ jest skończonym alfabetem symboli) taką, że

$$(\forall u, v \in \Sigma^*) (u R_L v \equiv ((\forall z \in \Sigma^*) uz \in L \Leftrightarrow vz \in L))$$

Uzasadnienie, że relacja R_L jest relacją prawostronnie niezmienniczą można znaleźć na stronie <http://kompilatory.agh.edu.pl>

Relacja R_L indukowana przez język L jest relacją równoważności.

Relacja równoważności o indeksie skończonym

Mówimy, że relacja równoważności jest relacją o indeksie skończonym, jeżeli ta relacja równoważności posiada skończoną liczbę klas abstrakcji.

Relacja indukowana przez język (2)

Przykład:

Dany jest język:

$$\{a^m b^n c^k \mid m+n > 0; n+k > 0\}$$

Znaleźć liczbę klas abstrakcji relacji R_L .

Rozważmy następujące zbiory:

- $K_0 = \{\varepsilon\}$
- $K_1 = \{a^p \mid p \geq 1\}$
- $K_2 = \{a^p b^r \mid p \geq 0; r \geq 1\}$
- $K_3 = \{a^p b^q c^r \mid p+q \geq 1, r \geq 1\}$
- K_4 – pozostałe słowa nad alfabetem $\Sigma = \{a, b, c\}$

Zbiory K_0, K_1, K_2, K_3, K_4 stanowią podział zbioru wszystkich słów nad alfabetem $\Sigma = \{a, b, c\}$.

Można uzasadnić, że każde dwa słowa z dowolnego ze zbiorów K_0, K_1, K_2, K_3, K_4 pozostają ze sobą w relacji R_L oraz że żadne dwa słowa z różnych zbiorów nie będą ze sobą w relacji R_L (por. <http://kompilatory.agh.edu.pl>). Tak więc liczba klas abstrakcji relacji R_L wynosi 5.

Relacja indukowana przez język (3)

Przykład

Znaleźć liczbę klas abstrakcji relacji R_L indukowanej przez język:

$$L = \{a^n b^n \mid n \geq 1\}$$

Rozważmy jednoelementowe zbiory $K_{i,\bullet} = \{a^i \mid i \geq 0\}$, wieloelementowe zbiory $K_{\bullet,j} = \{a^i b^k \mid i \geq 1, k \geq 1, 0 \leq j \leq i, j = i - k\}$ oraz zbiór K_x zawierający wszystkie pozostałe słowa nad alfabetem $\Sigma = \{a, b\}$. Zbiory jednoelementowe $K_{i,\bullet}$, wieloelementowe zbiory $K_{\bullet,j}$ oraz zbiór K_x stanowią podział zbioru wszystkich słów nad alfabetem $\Sigma = \{a, b\}$. Elementy każdego ze zbiorów $K_{i,\bullet}$ oraz $K_{\bullet,j}$ są oczywiście w relacji z samymi sobą. Element któregośkolwiek ze zbiorów $K_{i,\bullet}$ oraz $K_{\bullet,j}$ nie jest w relacji z elementem żadnego innego zbioru $K_{i,\bullet}$ czy $K_{\bullet,j}$. Żaden element zbioru K_x nie jest w relacji R_L z elementem któregośkolwiek ze zbiorów $K_{i,\bullet}$ oraz $K_{\bullet,j}$, gdyż prawostronne uzupełnienie każdego z łańcuchów z K_x o dowolny łańcuch daje słowo nie będące elementem języka L , zaś dla każdego ze zbiorów $K_{i,\bullet}$ oraz $K_{\bullet,j}$ istnieje jakiś łańcuch, po dopisaniu którego z prawej strony do elementu tego zbioru otrzymamy słowo języka. Tak więc zbiory $K_{i,\bullet}$ oraz $K_{\bullet,j}$ oraz zbiór K_x zawierający wszystkie pozostałe słowa nad alfabetem $\Sigma = \{a, b\}$ są klasami abstrakcji relacji R_L . Liczba klas abstrakcji relacji R_L jest w tym przypadku nieskończona. Wszystkie słowa badanego języka należą do $K_{\bullet,0}$.