JAO - Wprowadzenie do Gramatyk bezkontekstowych

Podstawowymi narzędziami abstrakcyjnymi do opisu języków formalnych są gramatyki i automaty. Gramatyka bezkontekstowa (ang. cfg) jest formalnie czwórką $G = (N, T, P, S)$, gdzie

- N - zbiór zmiennych syntaktycznych, inaczej zwanych symbolami nieterminalnymi (w skrócie nieterminalami)
- T - zbiór stałych, inaczej zwanych symbolami terminalnymi (w skrócie terminalami)
- P - zbiór reguł postaci $X \to \alpha$, $S \in N$, inaczej zwanych produkcjami, jest to program generujący słowa języka
- $S \in N$ - symbol początkowy gramatyki

Język $L(G)$ *generowany* przez G to zbiór słow nad alfabetem T które można otrzymać starując z S i stosując produkcje (reguły) gramatyki.

Przykład gramatyki bezkontekstowej dla języka:

$$L(G) = \{ w \in \{a, b\}^+ : \#_a(w) = \#_b(w) \}$$

Zbiorem nieterminali jest $N = \{A, B\}$, $T = \{a, b\}$. Zbiór produkcji to :

$$S \to aB,\ S \to bA,\ A \to a,\ A \to aS$$

$$A \to bAA,\ B \to b,\ B \to bS,\ B \to aBB$$

Gramatyki bezkontekstowe opisują własności tekstów w sposób rekurencyjny, dowody poprawności często indukcyjne ze względu na długość generowanego słowa.

Aby udowodnić poprawność ostatniej gramatyki dowodzimy indukcyjnie własności języków generowanych przez poszczególnie nieterminale:

- S generuje słowa w takie, że $\#_a(w) = \#_b(w)$,
- A takie, że $\#_a(w) = \#_b(w) + 1$
- B takie, że $\#_a(w) + 1 = \#_b(w)$.

Niech D_k będzie tzw. język Dycka k-tego rzędu, składa sie on z napisów będących poprawnymi wyrażeniami nawiasowymi k typów nawiasów.
D_k jest zbiorem napisów które można zredukować do słowa pustego zamieniając sąsiednie pary nawiasów, otwierający i zamykający tego samego typu, na słowo puste.
W szczególności D_1 jest generowany przez gramatykę

$$S \to SS \mid (S) \mid ( ) \mid \epsilon$$

Natomiast D_2 jest generowany przez:

$$S \to SS \mid (S) \mid [S] \mid ( ) \mid [ ] \mid \epsilon$$

Języki nawiasowe mają charakter "cebulkowy", niech

$$onion(X, Y, Z) = \bigcup_i X^i Y Z^i$$

Jeśli Y jest słowem pustym to piszemy $onion(X, Z)$. Jeśli X, Y, Z są określone przez gramatyki bezkontekstowe to łatwo wygenerować gramatykę na cebulkę. Rozważmy język

$$L = \{ uv : uv^R \in D_1 \}$$

Niech D_1' będzie językiem D_1 w którym zamieniliśmy nawiasy otwierające z zamykającymi. Niech

$$X = D_1 \cdot'' ('' \cdot D_1, \; Z = D_1' \cdot'' ('' \cdot D_1'$$

Wtedy

$$L = onion(X, Z)$$

Niech L będzie zbiorem słów długości nad $\{a, b\}$, które nie są postaci xx.

Niech $X = \{a, b\}$. Wtedy

$$L = onion(X, a, X) \cdot onion(X, b, X) \cup onion(X, b, X) \cdot onion(X, a, X)$$

Teraz już łatwo jest skonstruować gramatykę dla języka L.

$$S \to C1 \cdot C2 \mid C2 \cdot C1;\ C1 \to X \cdot C1 \cdot X;\ \mid a$$

$$C2 \to X \cdot C2 \cdot X;\ \mid b;\ X \to a \mid b$$

Nie wiadomo czy język słów pierwotnych, tzn. takich które nie są postaci x^k dla jakiegokolwiek naturalnego $k > 1$, jest bezkontekstowy.

Nieskończone słowo Fibonacciego $\mathcal{F}$ można zdefiniować jako nieskończony ciąg nad alfabetem $\{a, b\}$ spełniający warunki:

1. bb nie występuje,
2. pierwszym symbolem jest a;
3. dla każdego n zachodzi: n pozycji po n-tym wystąpieniu a występuje symbol b.

Niech L będzie zbiorem słów które nie są prefiksami (są nie-prefiskami) ciągu $\mathcal{F}$. Gramatykę dla L można skonstruować jako kompozycję trzech gramatyk, negujących kolejne warunki (w sensie alternatywy).
Negacja ostatniego warunku sprawdza sie do pewnej cebulki gdzie

$$X = b^*ab^*, \ Z = \{a, b\}$$

Postępujemy podobnie jak dla zbioru nie-prefiksów słowa Fibonacciego. Słowo Thue-Morse'a $\mathcal{T}$ spełnia warunki:

1. $\mathcal{T}$ zaczyna się od 0;
2. $\mathcal{T}_{2i} = \mathcal{T}_i$;
3. $\mathcal{T}_{2i+1}$ jest negacją $\mathcal{T}_i$.

Konstruujemy trzy gramatyki, z których każda niezależnie sprawdza negację jednego z trzech warunków. Konstrukcja dla ostatnich dwóch warunków jest *cebulkowa*.

Nie wiadomo czy język nie-podsłów słowa $\mathcal{T}$ jest bezkontekstowy. Udowodnimy potem, że język podsłów nie jest bezkontekstowy.

Gramatyka bezkontekstowa jest liniowa jesli po prawej stronie produkcji występuje co najwyżej jeden symbol nieterminalny.

Gramatyki jednostronnie liniowe to takie w których symbol nieterminalny wystepuje zawsze jako ostatni w produkcjach gramatyki.
W szczególności gramatyka jest prawostronnie liniowa gdy produkcje sa postaci $A \to wB$ lub $A \to \epsilon$ gdzie $w \in T^*$

Gramatyki lewostronnie liniowe definiujemy analogicznie.

Twierdzenie

Język jest regularny wtedy i tylko wtedy gdy jest generowany przez pewną gramatykę jednostronnie liniową.

Nieterminal A jest zbędny gdy nie występuje w żadnym wyprowadzeniu jakiegokolwiek słowa terminalnego z początkowego nieterminala gramatyki. W szczególności $L(G) = \emptyset$ wtw. gdy wszystkie nieterminale są zbędne.

Twierdzenie

Można sprawdzić w czasie wielomianowym, które nieterminale są zbędne. Przy okazji daje to wielomianowy algorytm na niepustść $L(G)$ (czy symbol początkowy jest zbędny).

Twierdzenie

Jeśli słowo puste nie jest generowane przez gramatykę to można przekształcić tę gramatykę do gramatyki równoważnej nie używającej słowa pustego ale rozmiar gramatyki może wzrosnąć wykładniczo.