Zadanie 2 z pierwszej części egzaminu 2016

Antoni Kościelski

1 Treść zadania

Przyjmijmy, że CFL_1 oznacza te języki bezkontekstowe, dla których istnieje gramatyka mająca, oprócz symbolu początkowego S_0, który nie może pojawić się po prawej stronie żadnej reguły, jeszcze jeden symbol nieterminalny S, zaś L_3 – język reprezentowany przez wyrażenie regularne $aa^*bb^*cc^*$.

Część I. Udowodnij następujący lemat:

Lemat 1.1 (o pompowaniu dla języków CFL_1) *Dla każdego języka $L \in CFL_1$ istnieje stała k taka, że dla każdego słowa $w \in L$ o długości większej od k istnieje podział $v_0 s_1 v_1 s_2 \ldots s_l v_l$ słowa w i podział $x_i y_i z_i$ każdego ze słów s_i takie, że*

1) $l \leqslant k$,

2) każde v_i i każde y_i ma długość co najwyżej k,

3) dla każdych $1 \leqslant i, j \leqslant l$ słowo $v_0 s_1 v_1 s_2 \ldots v_{i-1} x_i s_j z_i v_i \ldots s_l v_l$, powstałe ze słowa w przez zastąpienia fragmentu y_i słowa s_i w w przez s_j również należy do L.

Część II. Użyj powyższego lematu o pompowaniu aby pokazać, że $L_3 \notin CFL_1$.

Uwaga. Aby ułatwić trochę czytanie wzorów, wprowadźmy następujące oznaczenia: dla danego podziału słowa w takiego, jak w lemacie o podstawianiu, postaci

$$v_0 s_1 v_1 s_2 \ldots v_{i-1} x_i y_i z_i v_i \ldots s_l v_l,$$

symbolem $w_i[u]$ będziemy oznaczać słowo

$$v_0 s_1 v_1 s_2 \ldots v_{i-1} x_i u z_i v_i \ldots s_l v_l.$$

Tak więc punkt 3) ze sformułowania lematu o pompowaniu można też sformułować w następujący sposób:

3) dla każdych $1 \leqslant i, j \leqslant l$ słowo $w_i[s_j]$ również należy do L.

2 Dowód lematu o pompowaniu

Załóżmy, że język L jest generowany przez gramatykę G z produkcjami

$$S_0 \to P_i \text{ dla } i = 1, \ldots, m, \quad P_i \in \{a, b, c, S\}^*$$

oraz

$$S \to u_i \text{ dla } i = 1, \ldots, n, \quad u_i \in \{a, b, c, S\}^*.$$

Przyjmijmy, że k oznacza większą z liczb

$$\max\{|\,P_i\,| : i = 1, 2, \ldots, m\} \text{ oraz } \max\{|\,u_i\,| : i = 1, 2, \ldots, n \wedge u_i \in \{a, b, c\}^*\}.$$

Weźmy dowolne słowo $w \in L$ o długości $> k$. Gdyby w było można wyprowadzić w G w jednym kroku, to w byłoby równe P_i dla pewnego i oraz

$$| w | = | P_i | \leqslant k.$$

Wyprowadzenie w wymaga więc przynajmniej dwóch kroków i ma postać[1]

$$S_0 \Rightarrow P \Rightarrow w$$

dla pewnego $i = 1, \ldots, m$ oraz $P = P_i$.

Słowo $P \in \{a, b, c, S\}^*$ można przedstawić w postaci

$$P = v_0 S v_1 S \ldots v_{l-1} S v_l$$

dla pewnych słów $v_0, v_1, \ldots, v_l \in \{a, b, c\}^*$. Nietrudno zauważyć, że długości tych słów oraz liczba l spełniają nierówności

$$| v_i | \leqslant | P | \leqslant k \text{ oraz } l = | P |_S \leqslant k.$$

W gramatykach bezkontekstowych, jeżeli tylko słowo w daje się wyprowadzić ze słowa $v_0 S v_1 S \ldots v_{l-1} S v_l$, to istnieją słowa $s_1, s_2, \ldots, s_l$ oraz wyprowadzenia $S \Rightarrow s_i$ dla $i = 1, 2, \ldots, l$ takie, że

$$w = v_0 s_1 v_1 s_2 \ldots v_{l-1} s_l v_l.$$

W ten sposób zdefiniowaliśmy podział słowa w wymagany przez tezę lematu. Teraz musimy odpowiednio podzielić słowa s_i.

Jeżeli słowo s_i zostało wyprowadzone z S w jednym kroku, to przyjmujemy $x_i = z_i = \varepsilon$ oraz $y_i = s_i$. W przeciwnym razie wyprowadzenie s_i wymaga przynajmniej dwóch kroków i – po wyróżnieniu ostatniego kroku wyprowadzenia – ma postać

$$S \Rightarrow x_i S y_i \Rightarrow x_i u_j z_i = s_i$$

dla pewnego u_j takiego, że $S \to u_j$. W tej sytuacji przyjmujemy, że $y_i = u_j$, a x_i i z_i są takie, jak wyżej. Oczywiście, dla wszystkich i mamy $| y_i | \leqslant k$, a także

$$S \Rightarrow x_i S z_i \text{ oraz } S \Rightarrow y_i. \tag{1}$$

Pozostało dowieść, że zdefiniowane podziały mają własność 3) z lematu o pompowaniu. Weźmy więc dowolne i, j takie, że $1 \leqslant i, j \leqslant l$. Wyprowadzając w gramatykach bezkontekstowych, symbole nieterminalne możemy zastępować w dowolnej kolejności. Wobec tego, wyprowadzając ze słowa P możemy najpierw przekształcać wszystkie symbole S z wyjątkiem i-tego, zostawiając wyprowadzanie z tego ostatniego symbolu na koniec. Ze sposobu konstruowania podziału i wzorów (1) wynika, że

$$S_0 \Rightarrow v_0 s_1 v_1 \ldots v_{i-1} S v_i \ldots v_{l-1} s_l v_l \Rightarrow w_i[S] \text{ dla } i = 1, 2 \ldots, l,$$

a słowo w można też wyprowadzić w następujący sposób:

$$S_0 \Rightarrow w_i[S] \Rightarrow w_i[y_i] = w.$$

Zmieniając samą końcówkę tego wyprowadzenia możemy uzyskać słowo podane w tezie lematu:

$$S_0 \Rightarrow w_i[S] \Rightarrow w_i[s_j].$$

[1]Będziemy stosować następujące, uproszczone oznaczenia związane z gramatykami: symbol $\to$ oznacza produkcję gramatyki, a symbol $\Rightarrow$ relację wyprowadzenia w rozważanej gramatyce, albo w jednym, albo też w wielu krokach.

3 Dwa dające do myślenia fakty

Najpierw zauważmy, że lemat o pompowaniu mówi coś tylko o długich słowach. Mamy bowiem następujący

Fakt 3.1 Weźmy dowolną liczbę naturalną k. Jeżeli słowo $w \in L_3$ ma długość niewiększą niż $k^2 + k$, to dla tej liczby k daje się ono podzielić tak, jak to zostało podane w lemacie o pompowaniu.

Aby się o tym przekonać, wystarczy dane słowo $w \in L_3$ jakkolwiek podzielić na $l \leqslant k^2 + k$ podsłów $v_0, v_1, \ldots, v_l$, każde długości $\leqslant k$. Przyjmując wtedy, że $s_i = x_i = y_i = z_i = \varepsilon$ otrzymujemy podział słowa

$$w = v_0 v_1 \ldots v_l = v_0 s_1 v_1 s_2 \ldots s_l v_l$$

spełniający wszelkie warunki z lematu o pompowaniu. W szczególności, wszystkie słowa $w_i[s_j] \in L_3$, gdyż $w_i[s_j] = w_i[s_i] = w$.

Drugi fakt podpowiada, jak można budować gramatykę typu CFG_1 generującą język L_3. Przy okazji podpowiada, z jakich powodów język L_3 nie jest typu CFG_1.

Fakt 3.2 Każdy język $L \subseteq L_3$ reprezentowany przez wyrażenie regularne postaci $aa^*(b^{m_1} + \ldots + b^{m_n})cc^*$ jest generowany przez gramatykę z produkcjami

$$S \to aS, \ \ S \to Sc \ \text{ oraz } \ S \to ab^{m_i}c \ \text{ dla } \ i = 1, 2, \ldots, n.$$

4 Rozwiązanie części II

Dla dowodu nie wprost załóżmy, że $L_3 \in CFG_1$. Możemy więc skorzystać z lematu o pompowaniu. W tym celu weźmy stałą k z tezy lematu i słowo $w = a^n b^n c^n \in L_3$ o długości $3n > k^2 + k$. Możemy dodatkowo założyć, że $k \geqslant 2$. Wtedy n będzie większe od k ($n > k$, jeżeli chcemy posłużyć się wzorem).

Zgodnie z lematem 1.1 słowo w można przedstawić w postaci

$$v_0 s_1 v_1 s_2 v_2 \ldots s_l v_l,$$

gdzie dla $i = 1, 2, \ldots l$ mamy

$$s_i = x_i y_i z_i$$

oraz

1) $l \leqslant k$;

2) $|\, v_i \,| \leqslant k$ dla $i = 0, 1, \ldots, l$,

3) $|\, y_i \,| \leqslant k$ dla $i = 1, 2, \ldots, l$,

4) wszystkie słowa postaci $w_i[s_j]$, dla dowolnych i, j, $1 \leqslant i, j \leqslant l$, należą do L_3.

Krok 1. Przynajmniej jedno ze słów $s_1, s_2, \ldots, s_l$ jest niepuste. W przeciwnym razie, długość słowa $w = v_0 v_1 v_2 \ldots v_{l-1} v_l$ byłaby nie większa niż $(k+1) \cdot k$. Byłoby to sprzeczne z wyborem w.

Krok 2. Litera a występuje w jednym ze słów $s_1, s_2, \ldots, s_l$. Słowo v_0 jest zbyt krótkie, by były w nim zawarte wszystkie litery a ze słowa w. Wobec tego, a jest pierwszą literą słowa $s_1 v_1 s_2 v_2 \ldots s_l v_l$. Jeżeli słowo s_1 jest niepuste, to a jest pierwszą literą s_1. W przeciwnym razie, $s_1 = \varepsilon$ i a jest pierwszą literą słowa $v_1 s_2 v_2 \ldots s_l v_l$. Przypuśćmy, że s_i jest niepustym słowem. W słowie

$$w_1[s_i] = v_0 s_i v_1 s_2 \ldots s_l v_l \in L_3.$$

wszystkie litery słowa s_i poprzedzają jedno z wystąpień litery a. Ponieważ jest to słowo z języka L_3, więc słowo s_i składa się wyłącznie z liter a, a ponieważ jest niepuste, więc występuje w nim litera a.

Krok 3. Analogicznie dowodzimy, że litera c występuje w jednym ze słów $s_1, s_2, \ldots, s_l$.

Krok 4. W żadnym ze słów s_i nie występują jednocześnie litery a i c. Przypuśćmy, że jednak w słowie s_i występują litery a i c. Słowo $w_i[s_i] \in L_3$ zawiera podsłowo $x_i s_i z_i$. W słowach z L_3 wszystkie litery znajdujące się za literą c są równe c. Wobec tego, słowo z_i składa się tylko z liter c. Analogicznie, słowo x_i składa się tylko z liter a. Każde podsłowo słowa z $w \in L_3$, zawierające litery a i c, zawiera także wszystkie wystąpienia litery b występujące w w. Wobec tego, wszystkie litery b występujące w w znajdują się w słowie $s_i = x_i y_i z_i$, a ponieważ nie ma ich w x_i oraz z_i, wszystkie znajdują się w y_i. Stąd $|\, y_i \,| \geqslant n > k$, co jednak nie jest możliwe.

Krok 5 i ostatni. Przyjmijmy, że litera a występuje w słowie s_i, a litera c występuje w słowie s_j. Wiemy już, że $i \neq j$. Gdyby i było większe od j, to w słowie w jedna z liter a znajdowałaby się za literą c, co nie jest możliwe. Tak więc $i < j$. Zgodnie z lematem o pompowaniu, słowo $w_i[s_j]$ należy do L_3. W słowach z L_3 wszystkie litery znajdujące się za literą c są równe c. Wobec tego, słowo $z_i v_{i+1} s_{i+2} \ldots v_l$, a także jego podsłowo s_j, składają się tylko z liter c. Analogicznie, słowo $w_j[s_i]$ należy do L_3, a stąd słowa $v_0 \ldots s_{j-1} v_{j-1} x_j$ oraz s_i składają się tylko z liter a. Wspólny fragment tych słów, czyli słowo $z_i v_{i+1} \ldots v_{j-1} x_j$, składa się zarówno tylko z liter a, jak i wyłącznie z liter c. Jest więc słowem pustym, także słowa z_i i x_j są puste, a słowo w jest równe

$$v_0 \ldots s_{i-1} v_{i-1} x_i y_i y_j z_j v_{j+1} s_{j+2} \ldots v_l = v_0 \ldots s_{i-1} v_{i-1} s_i \; s_j v_{j+1} s_{j+2} \ldots v_l.$$

W ten sposób okazało się, że słowo jest konkatenacją dwóch słów, złożonych odpowiednio z liter a i c. Nie ma więc w nim liter b, mimo że definiując słowo w wpisaliśmy do niech wiele takich liter. Ta sprzeczność świadczy o stwierdzenia z części II.

5 Inne rozwiązania części II i ciekawe argumenty.

Zakładamy teraz, że mamy słowo $w \in L_3$, w którym każda z liter a, b i c występuje więcej niż k razy, a także podziały

$$w = v_0 s_1 v_1 s_2 \ldots s_l v_l \;\text{ oraz }\; s_i = x_i y_i z_i \;\text{ dla }\; i = 1, 2, \ldots, l,$$

spełniające tezę lematu o pompowaniu ze stałą k.

Fakt 5.1 Jeżeli $l = 1$, to w słowie s_1 występują litery a i c, a w konsekwencji także litera b.

Tak jest, ponieważ $w = v_0 s_1 v_1$ i słowa v_0 oraz v_1 są zbyt krótkie, aby zawierały odpowiednio wszystkie litery a i c.

Fakt 5.2 W żadnym słowie s_i nie występują jednocześnie litery a i c.

To już zostało pokazane, w kroku 4. Alternatywny dowód jest następujący: w słowie s_i są też wszystkie litery b występujące w w, ale nie wszystkie mogą należeć do y_i. Wobec tego przynajmniej jedna z nich należy albo do x_i, albo do z_i. Słowo $w_i[s_i]$ zawiera podsłowo $x_i s_i z_i$, w którym – w przypadku, gdy b występuje w x_i – litera b poprzedza literę a w słowie s_i. W drugim przypadku, litera b ze słowa z_i znajduje się za literą c. Tak więc $w_i[s_i] \notin L_3$.

Fakt 5.3 Jeżeli $l > 1$, to w żadnym słowie s_i nie występują dwie różne litery.

Przypuśćmy, że w słowie s_i są dwie litery: a i b, oraz liczba j różna od i spełnia $1 \leqslant j \leqslant l$. Słowo $w_j[s_i]$ ma postać $\ldots s_i \ldots s_i \ldots$. w tym słowie litera b z pierwszego s_i poprzedza literę a z drugiego wystąpienia tego słowa. Wobec tego, $w_j[s_i] \notin L_3$.

Fakt 5.4 Istnieje $i \leqslant l$ takie, że w słowie s_i występuje litera b, ewentualnie s_i jest puste, ale wtedy znajduje się między literami b.

Pierwsza litera b w słowie w należy do pewnego podsłowa z podziału słowa w. Jeżeli tym podsłowem jest v_{i-1}, zbyt krótkie, by zawierać wszystkie litery b, to b jest pierwszą literą słowa $s_i v_i \ldots v_l$. Wtedy albo s_i zawiera literę b, albo jest puste. W tym drugim przypadku, litery b w słowie w znajdują się zarówno przed, jak i po fragmencie s_i. Jeżeli zaś pierwsza litera b ze słowa w należy do podsłowa s_i, to fakt jest oczywisty.

Z powyższych faktów łatwo wyprowadzić sprzeczność. Wystarczy wziąć słowo s_i z literą a i zastąpić nim fragment y_j ze słowa s_j zawierającego literę c. Po wykonaniu tego zastępowania otrzymyjemy słowo, w którym litera b poprzedza literę a.

Na koniec zauważmy, że biorąc długie słowo w, w którym każda litera występuje więcej niż $k^2 + k$ razy, bardzo łatwo wykazać, że każdy jego podział zawiera podsłowa s_i, s_j i s_k zawierające odpowiednio litery a, b i c.

6 Bardzo krótkie rozwiązanie części II

Dla dowodu nie wprost załóżmy, że L_3 jest językiem typu CFG_1 i zachodzi dla niego lemat o pompowaniu ze stałą k. Weźmy słowo $w = a^{k^2+k+1} b^{2k^2+k+1} c^{k^2+k+1}$. Zgodnie z lematem o pompowaniu, słowo w można przedstawić w postaci

$$w = v_0 s_1 v_1 s_2 \ldots s_l v_l, \quad \text{gdzie} \quad s_i = x_i y_i z_i \quad \text{dla} \quad i = 1, 2, \ldots, l.$$

W tym przedstawieniu długość słowa $v_0 v_1 \ldots v_{l-1} v_l$ nie przekraczała $k^2 + k$. Stąd wynika, że dla pewnego i jedna z liter a ze słowa w występuje w słowie s_i. Podobnie, dla liter b. Jest ich jednak więcej, ich liczba przekracza długość słowa $v_0 y_1 v_1 y_2 \ldots v_{l-1} y_l v_l$. Tak więc, dla pewnego j jedna z liter b ze słowa w występuje w słowie $x_j z_j$. Z literami c jest podobnie, jak z a, i jest takie k, że w słowie s_k występuje c.

Jeżeli w słowie x_j znajduje się litera b, to rozważamy słowo $w_j[s_i]$. W tym słowie litera b z x_j poprzedza literę a występującą w s_i. Słowo $w_j[s_i]$ nie należy więc do L_3. Jeżeli litera b jest w słowie z_j, to znajduje się w słowie $w_j[s_k]$ za literą c. W tym przypadku słowo $w_j[s_k]$ też nie należy do L_3. W obu przypadkach dochodzimy do wniosków sprzecznych z lematem o pompowaniu, a to z kolei świadczy o słuszności uzasadnianej tezy.