Cvičenie č. 9
Bezkontextové jazyky

Peter Kostolányi

16. novembra 2022

Vieme už, že zásobníkové automaty sú – bez ohľadu na mód akceptácie – z hľadiska opisnej sily ekvivalentné bezkontextovým gramatikám. Obidva modely teda realizujú rovnakú triedu jazykov, ktoré nazývame *bezkontextovými*.

Definícia 1. *Bezkontextový jazyk* je jazyk L, pre ktorý existuje bezkontextová gramatika G taká, že $L(G) = L$. Triedu všetkých bezkontextových jazykov označujeme $\mathscr{L}_{CF}$.

Veta 1. *Nech L je jazyk. Nasledujúce tvrdenia sú ekvivalentné:*

(*i*) *Jazyk L je bezkontextový.*

(*ii*) *Existuje zásobníkový automat A taký, že $L(A) = L$.*

(*iii*) *Existuje zásobníkový automat A taký, že $N(A) = L$.*

1 Pumpovacia lema pre bezkontextové jazyky

Nech L je *ľubovoľný* bezkontextový jazyk. Potom *existuje* bezkontextová gramatika $G = (N, T, P, \sigma)$ taká, že $L(G) = L$. Nech má najdlhšia pravá strana pravidla z P dĺžku k; bez ujmy na všeobecnosti predpokladajme, že $k \geq 2$. V každom kroku odvodenia potom možno jeden neterminál prepísať na najviac k symbolov. Strom odvodenia s hĺbkou menšou alebo rovnou n tak môže zodpovedať iba slovu dĺžky najviac k^n a každý strom odvodenia slova dĺžky aspoň k^n preto musí mať hĺbku najmenej n. Položme $p := k^{|N|+1}$.

Nech $w \in L$ je *ľubovoľné* slovo také, že $|w| \geq p$. Vezmime strom najkratšieho[1] odvodenia slova w v gramatike G. Ten má hĺbku aspoň $|N| + 1$. Najdlhšia cesta z koreňa do niektorého z listov preto obsahuje aspoň $|N| + 2$ uzlov, pričom v liste je terminál alebo prázdne slovo a vo vnútorných uzloch, ktorých je aspoň $|N|+1$, sú neterminály. Na tejto ceste sa teda musí zopakovať niektorý neterminál ξ. Ak je cesta dlhšia ako $|N|+1$, môžeme predpokladať, že k obom výskytom neterminálu ξ dôjde v rámci najnižších $|N| + 2$ úrovní. Situácia je znázornená na obrázku 1.

Odvodenie slova w v gramatike G tak možno s odkazom na obrázok 1 zapísať nasledovne:

$$\sigma \Rightarrow^* x\xi z \Rightarrow^+ xu\xi vz \Rightarrow^* xuyvz = w.$$

Pre slovo w teda *existujú* slová x, u, y, v, z také, že $w = xuyvz$. Keďže k podslovu uyv prislúcha podstrom hĺbky najviac $|N| + 1$, nutne $|uyv| \leq k^{|N|+1} = p$. Keby boli obidve slová u, v prázdne, bolo by možné časť odvodenia – konkrétne $x\xi z \Rightarrow^+ xu\xi vz$ – vynechať bezo zmeny vygenerovaného slova; to by ale odporovalo nášmu predpokladu, podľa ktorého ide o najkratšie odvodenie slova w. Preto $|uv| \geq 1$. Keďže napokon možno uvedenú časť odvodenia – $x\xi z \Rightarrow^+ xu\xi vz$ – vynechať alebo niekoľkokrát zopakovať, pre každé $i \in \mathbb{N}$ patrí aj slovo xu^iyv^iz do jazyka generovaného gramatikou G. Dokázali sme teda nasledujúce tvrdenie – *pumpovaciu lemu pre bezkontextové jazyky*.

Veta 2. *Nech Σ je abeceda a $L \subseteq \Sigma^*$ bezkontextový jazyk. Potom existuje $p \in \mathbb{N}$ také, že pre všetky $w \in L$ s $|w| \geq p$ existujú slová $x, u, y, v, z \in \Sigma^*$ také, že:*

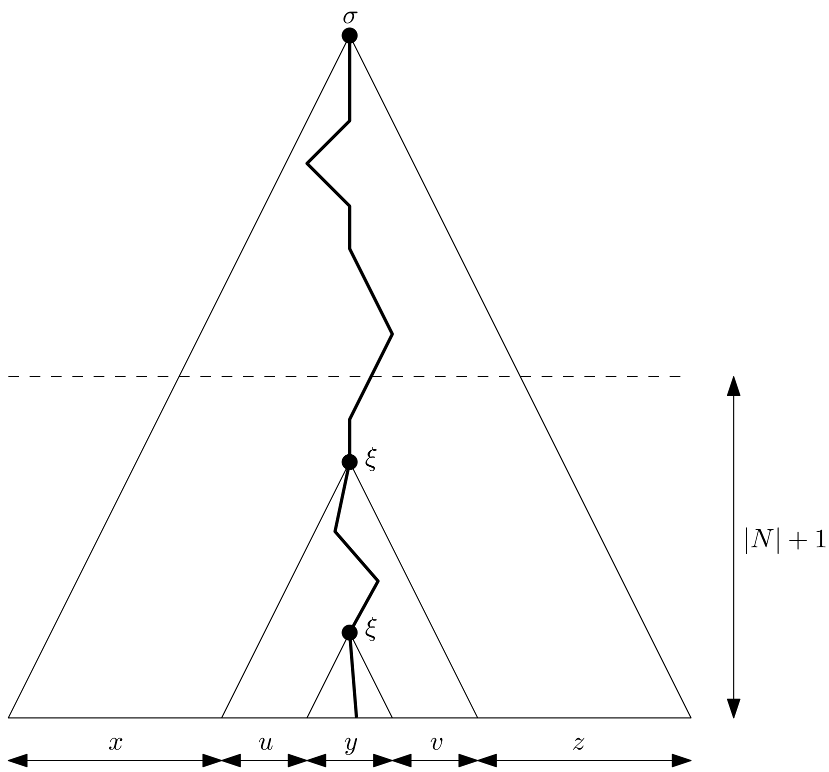
(*i*) $w = xuyvz$,

(*ii*) $|uyv| \leq p$,

(*iii*) $|uv| \geq 1$,

(*iv*) $\forall i \in \mathbb{N} : xu^iyv^iz \in L$.

[1] Ak existuje viacero odvodení minimálnej dĺžky, môžeme vziať ľubovoľné z nich.



Obr. 1: Schematické znázornenie stromu odvodenia v gramatike G. Ak má slovo dĺžku aspoň p, jeho strom odvodenia je hĺbky najmenej $|N|+1$. Najdlhšia cesta z koreňa do listu je znázornená hrubou čiarou. V rámci posledných $|N|+2$ uzlov – čiže v podstrome hĺbky najviac $|N|+1$ – sa na tejto ceste musí zopakovať niektorý neterminál ξ. Tieto jeho dva výskyty určujú faktorizáciu $w = xuyvz$ vygenerovaného slova.

1.1 Riešené úlohy

Pumpovacia lema pre bezkontextové jazyky je užitočným nástrojom pri dokazovaní, že nejaký jazyk L nie je bezkontextový. Jej použitie je podobné ako pri pumpovacej leme pre regulárne jazyky.

Úloha 1. Dokážte, že jazyk $L = \{wcw \mid w \in \{a,b\}^*\}$ nie je bezkontextový.

Riešenie. Za účelom sporu predpokladajme, že $L \in \mathscr{L}_{CF}$. Nech $p \in \mathbb{N}$ je konštanta prislúchajúca k jazyku L podľa pumpovacej lemy. Uvažujme slovo $w = a^p b^p c a^p b^p$; evidentne $w \in L$ a $|w| \geq p$. Podľa pumpovacej lemy teda existujú slová x, u, y, v, z, pre ktoré sú splnené podmienky (i) až (iv).

Keby bolo $\#_c(u) > 0$ alebo $\#_c(v) > 0$, muselo by byť aj $\#_c(xu^2yv^2z) \geq 2$, pričom z podmienky (iv) by sme pre $i = 2$ dostali $xu^2yv^2z \in L$ – to je očividný spor s definíciou jazyka L.

Môžeme teda predpokladať, že $\#_c(u) = \#_c(v) = 0$. Jediný výskyt písmena c v slove w tak musí byť súčasťou niektorého z faktorov x, y, z.

a) Nech $\#_c(x) = 1$. Potom $x = a^p b^p c x'$ pre nejaké $x' \in \{a,b\}^*$ také, že $x'uyvz = a^p b^p$. Vďaka (iii) ďalej $|u| + |v| \geq 1$. Z podmienky (iv) pre $i = 2$ dostávame $xu^2yv^2z = a^p b^p c x' u^2 y v^2 z \in L$. To ale odporuje definícii jazyka L, pretože $|x'u^2yv^2z| = |x'uyvz| + |u| + |v| \geq |a^p b^p| + 1$ – nemôže teda platiť $a^p b^p = x'u^2yv^2z$.

b) Nech $\#_c(z) = 1$. Potom $z = z' c a^p b^p$ pre nejaké $z' \in \{a,b\}^*$ také, že $xuyvz' = a^p b^p$. Vďaka (iii) je $|u| + |v| \geq 1$ a zo (iv) pre $i = 2$ dostávame $xu^2yv^2z = xu^2yv^2z' c a^p b^p \in L$. To je spor, lebo $|xu^2yv^2z'| = |xuyvz'| + |u| + |v| \geq |a^p b^p| + 1$ – nemôže teda platiť $xu^2yv^2z' = a^p b^p$.

c) Nech $\#_c(y) = 1$. Keďže vďaka podmienke (ii) je $|uyv| \leq p$, pre nejaké $k, \ell, m, n \in \mathbb{N}$ musí byť $x = a^p b^k$, $u = b^\ell$, $y = b^{p-k-\ell} c a^m$, $v = a^n$ a $z = a^{p-m-n} b^p$. Vďaka podmienke (iii) pritom platí aspoň jedna z nerovností $\ell \geq 1$ alebo $n \geq 1$. Z podmienky (iv) pre $i = 2$ napokon dostávame $xu^2yv^2z = a^p b^k b^{2\ell} b^{p-k-\ell} c a^m a^{2n} a^{p-m-n} b^p = a^p b^{p+\ell} c a^{p+n} b^p \in L$ – to je spor, pretože v prípade platnosti aspoň jednej z nerovností $\ell \geq 1$ a $n \geq 1$ nemôže byť $a^p b^{p+\ell} = a^{p+n} b^p$. $\square$

Úloha 2. Dokážte, že jazyk $L = \{a^{2^n} \mid n \in \mathbb{N}\}$ nie je bezkontextový.

Riešenie. Budeme postupovať veľmi podobne ako pri dokazovaní, že $L \notin \mathscr{R}$. Za účelom sporu predpokladajme, že $L \in \mathscr{L}_{CF}$. Nech $p \in \mathbb{N}$ je konštanta prislúchajúca k jazyku L podľa pumpovacej lemy. Vezmime ľubovoľné slovo $w = a^{2^m}$ také, že $2^m > p$ – evidentne $w \in L$ a $|w| \geq p$. Existujú teda slová x, u, y, v, z, pre ktoré sú splnené podmienky (i) až (iv) z pumpovacej lemy.

Vďaka (i) je $w = xuyvz$. Z (ii) a (iii) ďalej vyplýva, že pre nejaké $r, s, t, k \in \mathbb{N}$ s $s + k \geq 1$ a $s + t + k \leq p$ je $x = a^r$, $u = a^s$, $y = a^t$, $v = a^k$ a $z = a^{2^m - r - s - t - k}$. Z podmienky (iv) pumpovacej lemy napokon pre $i = 2$ vyplýva $xu^2yv^2z = a^r a^{2s} a^t a^{2k} a^{2^m - r - s - t - k} = a^{2^m + s + k} \in L$. Keďže ale $1 \leq s + k \leq s + t + k \leq p < 2^m$, je $2^m < 2^m + s + k < 2^{m+1}$, a teda $2^m + s + k$ nie je mocnina dvoch – dostávame teda spor s definíciou jazyka L. $\square$