Ogdenova lema

Peter Kostolányi

5. marca 2024

1 Obmedzenia pumpovacej lemy pre bezkontextové jazyky

Pripomeňme si znenie pumpovacej lemy pre bezkontextové jazyky.

Veta 1. *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) $\forall i \in \mathbb{N} : xu^iyv^iz \in L$.

Pokiaľ teda pre nejaký jazyk L konštanta p s uvedenými vlastnosťami neexistuje, jazyk L nemôže byť bezkontextový. Na tomto pozorovaní je založená metóda dokazovania negatívnych výsledkov o bezkontextovosti jazykov z minulého semestra. Ukážeme teraz, že napriek svojej nespornej užitočnosti táto metóda nie je univerzálne použiteľná, keďže pumpovacia lema hovorí iba o *nutnej*, ale nie o postačujúcej podmienke bezkontextovosti jazyka. Existujú teda aj jazyky $L \notin \mathscr{L}_{CF}$, pre ktoré konštanta p z pumpovacej lemy existuje – vyvrátiť bezkontextovosť takéhoto jazyka pomocou pumpovacej lemy logicky nie je možné.

Príklad 1. Uvažujme jazyk $L = \{a^nb^nc^nd^k \mid n, k \in \mathbb{N}\} \cup \{a, b, c\}^*$. Tento jazyk nie je bezkontextový, pretože inak by pre *regulárny* jazyk $R = a^*b^*c^*d$ a homomorfizmus $h\colon \{a, b, c, d\}^* \to \{a, b, c\}^*$ s $h(a) = a$, $h(b) = b$, $h(c) = c$ a $h(d) = \varepsilon$ musel byť bezkontextový aj jazyk

$$h(L \cap R) = \{a^nb^nc^n \mid n \in \mathbb{N}\},$$

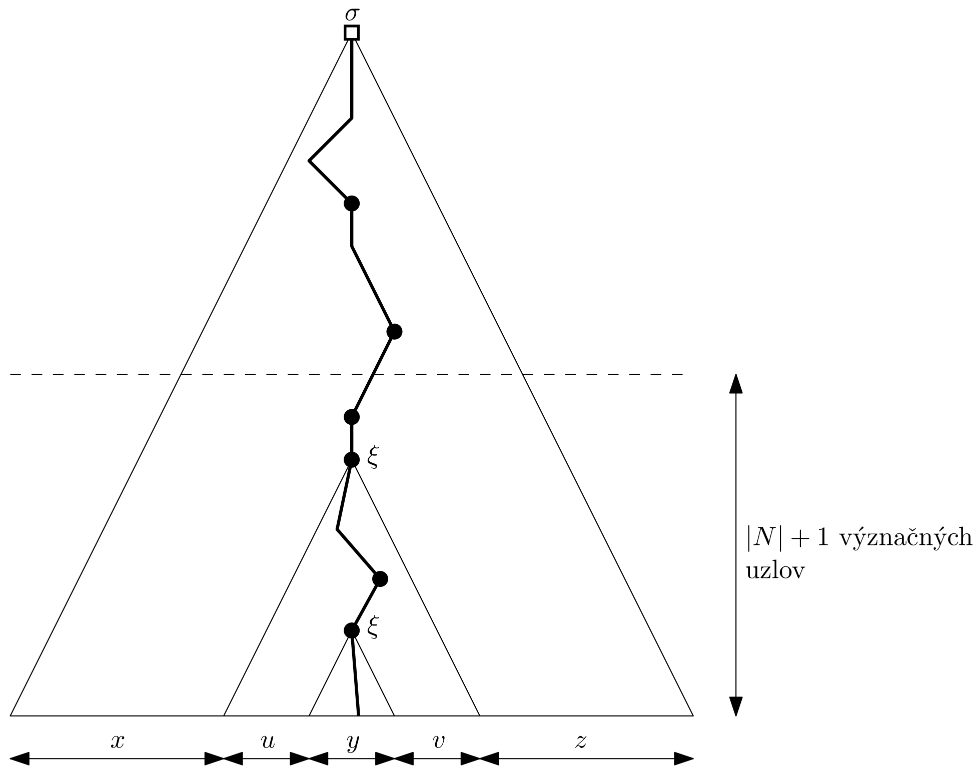
o ktorom vieme, že bezkontextový nie je. Jazyk L ale napriek tomu spĺňa podmienky z pumpovacej lemy pre bezkontextové jazyky pre $p = 1$: ak $w \in \{a, b, c\}^*$ a $|w| \geq 1$, musí byť $w = et$ pre nejaké $e \in \{a, b, c\}$ a $t \in \{a, b, c\}^*$ a možno tak vziať napríklad $x = u = y = \varepsilon$, $v = e$ a $z = t$; ak $w = a^nb^nc^nd^k$ pre nejaké $n \in \mathbb{N}$ a $k \in \mathbb{N} - \{0\}$, možno zase vziať $x = a^nb^nc^nd^{k-1}$, $u = d$ a $y = v = z = \varepsilon$.

Ogdenova lema, ktorú v nasledujúcom odvodíme, je silnejším variantom pumpovacej lemy pre bezkontextové jazyky – bude ňou možné napríklad aj vyvrátiť bezkontextovosť jazyka z predchádzajúceho príkladu. Stále však pôjde iba o *nutnú* podmienku bezkontextovosti jazyka.

2 Ogdenova lema

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. Položme $p := k^{|N|+1}$.

Nech $w \in L$ je *ľubovoľné* slovo také, že $|w| \geq p$. Označme v slove w aspoň p *ľubovoľných* po dvoch rôznych symbolov. Vezmime strom nejakého odvodenia slova w v gramatike G a nazvime vnútorný uzol tohto stromu *význačným*, ak aspoň dva spomedzi podstromov zakorenených v synoch tohto uzla obsahujú list zodpovedajúci označenému symbolu[1] slova w. Každý uzol môže mať najviac k synov – ľahko sa teda presvedčíme o tom, že ak každá cesta z koreňa do listu obsahuje najviac M význačných uzlov, môže takýto strom zodpovedať iba odvodeniu slova s nanajvýš k^M označenými symbolmi. Keďže vo výsledku je potrebné vygenerovať aspoň $p = k^{|N|+1}$ označených symbolov, musí z uvedeného dôvodu v strome existovať cesta z koreňa do niektorého z listov obsahujúca aspoň $|N|+1$ význačných uzlov. Vezmime spomedzi takýchto ciest nejakú s maximálnym počtom význačných uzlov. Na nej musia aspoň dva spomedzi najnižších $|N|+1$ význačných uzlov zodpovedať rovnakému neterminálu ξ. Situácia je znázornená na obrázku 1.



Obr. 1: Schematické znázornenie stromu odvodenia v gramatike G. Uvažovaná cesta z koreňa do listu s maximálnym počtom význačných uzlov je znázornená hrubou lomenou čiarou; význačné uzly sú na nej zvýraznené čiernymi kruhmi (koreň, ktorý môže, ale nemusí byť význačným uzlom, je zvýraznený prázdnym štvorcom). Spomedzi $|N| + 1$ najnižších význačných uzlov na tejto ceste musia aspoň dva zodpovedať rovnakému neterminálu ξ. Tieto jeho dva výskyty určujú faktorizáciu $w = xuyvz$ vygenerovaného slova.

Uvažované odvodenie slova w v gramatike G možno s odkazom na obrázok 1 zapísať ako

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

Pre slovo w teda *existujú* slová x, u, y, v, z také, že $w = xuyvz$. Podslovo uyv je vygenerované v rámci podstromu zakoreneného vo význačnom uzle zodpovedajúcom neterminálu ξ, pričom vďaka voľbe nami uvažovanej cesty a výskytov neterminálu ξ na nej je zrejmé, že každá cesta z koreňa tohto podstromu do niektorého listu môže obsahovať najviac $|N| + 1$ význačných uzlov. V dôsledku toho môže podslovo uyv slova w obsahovať najviac $k^{|N|+1} = p$ označených symbolov. Keďže je vyšší spomedzi dvoch uvažovaných uzlov zodpovedajúcich neterminálu ξ význačný, musí podslovo uyv obsahovať aspoň jeden označený symbol, pričom všetky tieto označené symboly nemôžu byť obsiahnuté v podslove y – v opačnom prípade by vyšší uzol zodpovedajúci ξ mohol mať iba jedného syna s označeným symbolom v podstrome. Označený tak musí byť aj aspoň jeden zo symbolov podslova u alebo v.

[1]Význačný uzol teda zodpovedá neterminálu ξ, na ktorý bolo v odvodení použité pravidlo $\xi \to x$, pričom z aspoň dvoch rôznych (neterminálnych alebo terminálnych) symbolov slova x sa v odvodení vygeneruje aspoň jeden označený symbol slova w.

Keďže napokon možno 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. Tým dostávame nasledujúce tvrdenie známe ako *Ogdenova lema.*

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 $|w| \geq p$ a ľubovoľné označenie najmenej p rôznych symbolov slova w existujú $x, u, y, v, z \in \Sigma^*$ také, že:*

(*i*) $w = xuyvz$;

(*ii*) *podslovo uyv obsahuje nanajvýš p označených symbolov slova w;*

(*iii*) *aspoň jedno z podslov u, v slova w obsahuje označený symbol;*

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

Ogdenova lema je skutočným zovšeobecnením pumpovacej lemy pre bezkontextové jazyky – tá totiž zodpovedá špeciálnemu prípadu Ogdenovej lemy, keď sú v uvažovanom slove w označené všetky symboly.

3 Riešené úlohy

Aplikujme najprv Ogdenovu lemu na vyvrátenie bezkontextovosti jazyka z príkladu 1 – tým ukážeme, že Ogdenova lema je skutočne silnejším nástrojom, než pumpovacia lema pre bezkontextové jazyky.

Úloha 1. Pomocou Ogdenovej lemy dokážte, že jazyk $L = \{a^nb^nc^nd^k \mid n, k \in \mathbb{N}\} \cup \{a, b, c\}^*$ nie je bezkontextový.

Riešenie. Za účelom sporu predpokladajme, že $L \in \mathscr{L}_{CF}$. Podľa Ogdenovej lemy prislúcha jazyku L konštanta $p \in \mathbb{N}$. Uvažujme slovo $w = a^pb^pc^pd$ – evidentne $w \in L$ a $|w| \geq p$ – a označme v ňom práve všetky výskyty písmena a. Z Ogdenovej lemy potom vyplýva existencia slov $x, u, y, v, z \in \{a, b, c, d\}^*$ s vlastnosťami (*i*) až (*iv*).

Keby niektoré zo slov u, v obsahovalo dva rôzne symboly, muselo by vďaka vlastnosti (*iv*) pre $i = 2$ do jazyka L patriť aj slovo xu^2yv^2z, ktoré však obsahuje najmenej jedno písmeno d a súčasne nie je prvkom jazyka $a^*b^*c^*d^*$ – nemôže teda byť ani prvkom jazyka L.

Každé zo slov u, v je teda mocninou niektorého z písmen a, b, c, d – a vďaka podmienke (*iii*) je $\#_a(uv) \geq 1$. Nutne potom existuje $e \in \{b, c\}$ také, že $\#_e(uv) = 0$. Vďaka (*iv*) pre $i = 2$ potom $xu^2yv^2z \in L$ – avšak

$$\#_a(xu^2yv^2z) = \#_a(w) + \#_a(uv) > \#_a(w) = \#_e(w) = \#_e(w) + \#_e(uv) = \#_e(xu^2yv^2z)$$

a súčasne $\#_d(xu^2yv^2z) \geq \#_d(w) = 1$, čo je spor. □

Úloha 2. Zistite, či je jazyk $L = \{a^ib^ic^j \mid i, j \in \mathbb{N};\ i \neq j\}$ bezkontextový. Svoje tvrdenie dokážte.

Riešenie. Dokážeme, že jazyk L nie je bezkontextový. Predpokladajme opak; jazyku L potom podľa Ogdenovej lemy prislúcha konštanta $p \in \mathbb{N}$. Uvažujme slovo $w = a^{p+p!}b^{p+p!}c^p$; evidentne $w \in L$ a $|w| \geq p$. Označme v slove w práve všetky výskyty písmena c. Z Ogdenovej lemy dostávame existenciu slov $x, u, y, v, z \in \{a, b, c\}^*$ spĺňajúcich podmienky (*i*) až (*iv*).

Žiadne zo slov u, v nemôže obsahovať výskyty dvoch rôznych symbolov, pretože v opačnom prípade by slovo xu^2yv^2z nepatrilo do jazyka $a^*b^*c^*$, čo by odporovalo podmienke (*iv*), podľa ktorej toto slovo musí byť prvkom L.

Keďže podľa (*iii*) musí byť $\#_c(uv) \geq 1$, nutne nastane jeden z nasledujúcich troch prípadov:

a) $u = c^r$ a $v = c^s$ pre nejaké $r, s \in \mathbb{N}$ také, že $1 \leq r + s \leq p$ (pokiaľ pritom $r = 0$, slovo y nemusí byť mocninou písmena c);

b) $u = a^r$ a $v = c^s$ pre nejaké $r, s \in \mathbb{N} - \{0\}$;

c) $u = b^r$ a $v = c^s$ pre nejaké $r, s \in \mathbb{N} - \{0\}$.

V prvom z prípadov môžeme v (*iv*) zvoliť $i = p!/(r + s) + 1$; z toho $xu^iyv^iz = a^{p+p!}b^{p+p!}c^{p+p!} \in L$, čo je spor. V zostávajúcich dvoch prípadoch stačí vziať $i = 2$ a všimnúť si, že pre slovo $xu^2yv^2z \in L$ je $|\#_a(xu^2yv^2z) - \#_b(xu^2yv^2z)| = r > 0$. To je opäť spor. □

Poznámka 1. Nie je ťažké ukázať, že podobne ako v prípade jazyka z príkladu 1, aj pre jazyk L z predchádzajúcej úlohy existuje konštanta $p \in \mathbb{N}$, pre ktorú platí tvrdenie z pumpovacej lemy pre bezkontextové jazyky. Na vyvrátenie bezkontextovosti jazyka L teda pumpovacia lema nie je postačujúcim nástrojom.

4 Nejde o postačujúcu podmienku bezkontextovosti

Podobne ako v prípade pumpovacej lemy pre bezkontextové jazyky, aj v prípade Ogdenovej lemy ide len o nutnú podmienku bezkontextovosti jazyka, ktorá nie je postačujúcou. Existujú teda aj jazyky, ktoré nie sú bezkontextové, ale pomocou Ogdenovej lemy to o nich dokázať nemožno. Ukážme si teraz príklad takéhoto jazyka, s ktorého o niečo všeobecnejším variantom prišli roku 1978 L. Boasson a S. Horváth [1].

Príklad 2. Uvažujme abecedu $\Sigma = \{a, b\}$ a jazyk $L = \{(ab)^{2^n} \mid n \in \mathbb{N}\} \cup \Sigma^*\{aa, bb\}\Sigma^*$. Tento jazyk nemôže byť bezkontextový, pretože pre homomorfizmus $h\colon a^* \to \Sigma^*$ daný ako $h(a) = ab$ je

$$h^{-1}(L) = \{a^{2^n} \mid n \in \mathbb{N}\} \notin \mathscr{L}_{CF}.$$

Napriek tomu ale uvedený jazyk vyhovuje tvrdeniam Ogdenovej lemy pre $p = 3$. Ak totiž $w = (ab)^{2^n}$ pre $n \in \mathbb{N} - \{0\}$, pričom sú v tomto slove označené aspoň tri symboly, možno vziať za slovo u ktorýkoľvek označený symbol, ktorý nie je na začiatku ani na konci slova a slovo v možno vziať prázdne; slová x, y, z ďalej zvolíme tak, aby bolo $xuyvz = w$. Pre takto zvolené $x, u, y, v, z \in \Sigma^*$ sú potom evidentne splnené podmienky (*i*) až (*iii*) z Ogdenovej lemy; vypustením alebo niekoľkonásobným zopakovaním podslova u dĺžky 1 ďalej zo slova w určite dostaneme slovo obsahujúce dva po sebe idúce výskyty niektorého písmena – pre $i \in \mathbb{N} - \{1\}$ tak $xu^iyv^iz \in \Sigma^*\{aa, bb\}\Sigma^*$, kým pre $i = 1$ je $xuyvz \in \{(ab)^{2^n} \mid n \in \mathbb{N}\}$; je teda splnená aj podmienka (*iv*) Ogdenovej lemy.

Ak naopak $w \in \Sigma^*\{aa, bb\}\Sigma^*$ je slovo dĺžky aspoň 3, v ktorom sú označené aspoň tri symboly, zvoľme ľubovoľné podslovo aa alebo bb slova w a spomedzi označených symbolov *mimo tohto podslova* zvoľme ľubovoľný za u. Za slovo v vezmime ε a slová x, y, z zvoľme tak, aby bolo $xuyvz = w$. Zrejme potom $xyz \in \Sigma^*\{aa, bb\}\Sigma^*$ Podmienky (*i*) až (*iii*) z Ogdenovej lemy sú opäť evidentne splnené – a splnená je aj podmienka (*iv*), pretože z $xyz \in \Sigma^*\{aa, bb\}\Sigma^*$, $w = xuyvz \in \Sigma^*\{aa, bb\}\Sigma^*$ a $u \in \{a, b\}$ vyplýva $xu^iyv^iz \in \Sigma^*\{aa, bb\}\Sigma^*$ pre všetky $i \in \mathbb{N}$.

Literatúra

[1] Boasson, L., Horváth, S.: On Languages Satisfying Ogden's Lemma. In *RAIRO – Informatique théorique*. 1978, vol. 12, no. 3, pp. 201–202.