KONTEKSTNO-NEODVISNE GRAMATIKE ZA KODIRANJE IN STISKANJE PODATKOV

JANEZ PODLOGAR

1. Kontekstno-neodvisne gramatike

V jezikoslovju pravopis določa pravila o rabi črk in ločil. S slovnico poimenujemo sistem pravil za tvorjenje povedi in sestavljanje besedil. Slovenska slovnica, Slovenski pravopis in Slovar slovenskega knjižnega jezika natančno določajo Slovenski knjižni jezik, ki je poglavitno sredstvo javnega in uradnega sporazumevanja v Sloveniji.

Podobno je v teoriji formalnih jezikov gramatika sistem pravil, ki nam pove, kako iz dane abecede tvorimo nize. Gramatika nam torej določa neko podmnožico nizov, ki jo imenujemo formalni jezik. Gramatike in formalni jeziki imajo široko teoretični in praktično uporabo. Uporabljajo se za modeliranje naravnih jezikov, so osnova programskih jezikov, formalizirajo matematično logiko in sisteme aksiomov ter se uporabljajo tudi za kompresijo podatkov.

Definicija 1.1. *Abeceda* je končna neprazna množica Σ. Elementom abecede pravimo *črke*. *Množica vseh končnih nizov abecede* Σ je

$$\Sigma^* = \{a_1a_2a_3\cdots a_n \mid n \in \mathbb{N}_0 \wedge \forall i : a_i \in \Sigma\},$$

kjer za $n = 0$ dobimo prazen niz, ki ga označimo z ε. *Množica vseh končnih nizov abecede brez praznega niza* označimo s Σ^+. *Dolžino niza* w označimo z $|w|$ in je enaka številu črk v nizu $w \in \Sigma^*$. *Množico vseh nizov dolžine* ℓ, kjer je ℓ pozitivno celo število, označimo s Σ^ℓ. *Jezik na abecedi* Σ je poljubna podmnožica množice Σ^*.

Definicija 1.2. Naj bo Σ abeceda. Naj bo $*$ asociativna binarna operacija na množici vseh končnih nizov Σ^* tako, da je prazen niz ε nevtralni element in za niza $w, u \in \Sigma^*$ velja

$$w * u = w_1w_2\cdots w_nu_1u_2\cdots u_m,$$

kjer sta $w_1w_2\cdots w_n$ in $u_1u_2\cdots u_m$ predstavitvi nizov w in u s črkami abecede Σ. Znak za operacijo $*$ spustimo in krajše pišemo wu. Monoid $(\Sigma^*, *)$ imenujemo *prost monoid nad* Σ.

Opomba 1.3. Dvočlena operacija $\circ$ na množici A je preslikava:

$$A \times A \to A,$$
$$(x, y) \mapsto x \circ y.$$

Monoid $(A, \circ)$ je neprazna množica A z dvočleno asociativno operacijo $\circ$, ki ima nevtralni element. Ime prosti monoid izhaja iz Teorije kategorij.

Definicija 1.4. Naj bo Σ abeceda. *Členitev niza* $w \in \Sigma^*$ je vsako zaporedje $(w_1, w_2, \ldots, w_m)$ tako, da so $w_1, w_2, \ldots, w_m \in \Sigma^*$ in je

$$w_1w_2\cdots w_n = w.$$

Date: 28. februar 2023.

Definicija 1.5. *Kontekstno-neodvisna gramatika* je četverica $G = (V, \Sigma, P, S)$, kjer je V končna množica *nekončnih simbolov*; Σ množica *končnih simbolov* tako, da $\Sigma \cap V = \emptyset$; $P \subseteq V \times (V \cup \Sigma)^*$ celovita relacija, elementom relacije pravimo *produkcijska pravila*; in $S \in V$ *začetni simbol*.

Opomba 1.6. Relacija $P \subseteq A \times B$ je celovita, če velja

$$\forall x \in A \; \exists y \in B \colon (x, y) \in P.$$

Definicija 1.7. Naj bo $G = (V, \Sigma, P, S)$ kontekstno-neodvisna gramatika. Naj bodo α, β, $\gamma \in (V \cup \Sigma)^*$ nizi nekončnih in končnih simbolov, $A \in V$ nekončni simbol ter naj bo $(A, \beta) \in P$ produkcijsko pravilo, označimo ga z $A \to \beta$, A imenujemo *levi član produkcijskega pravila* in β imenujemo *desni član produkcijska pravila*. Pravimo, da se $\alpha A \gamma$ *prepiše s pravilom* A v $\alpha\beta\gamma$, pišemo $\alpha A \gamma \Rightarrow \alpha\beta\gamma$. Pravimo, da α *izpelje* β, če je $\alpha = \beta$ ali če za $n \geq 0$ obstaja zaporedje $\alpha_1, \alpha_2, \ldots \alpha_n \in (V \cup \Sigma)^*$ tako, da

$$\alpha \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \cdots \Rightarrow \alpha_n \Rightarrow \beta,$$

pišemo $\alpha \overset{*}{\Rightarrow} \beta$. *Jezik kontekstno-neodvisne gramatike* G je množica vseh nizov končnih simbolov, ki jih lahko izpeljemo s produkcijskimi pravili gramatike, označimo ga z $L(G)$.

Posledica 1.8. *Jezik kontekstno neodvisne gramatike* G *je*

$$L(G) = \{w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w\}.$$

Opomba 1.9. Standardno z velikimi tiskanimi črkami $A, B, C, \ldots$ označujemo nekončne simbole, z malimi tiskanimi črkami $a, b, c, \ldots$ označujemo končne simbole in z grškimi črkami $\alpha, \beta, \gamma, \ldots$ označujemo končne nize nekončnih in končnih simbolov. Simbol, ki je nekončen ali končen, bomo označili z y. Če ni navedeno drugače, je S vedno začetni simbol.

Pravila, ki imajo za levega člana isti simbol, $A \to \alpha$, $A \to \beta$, $\ldots$, kompaktno zapišemo

$$A \to \alpha \mid \beta \mid \ldots$$

Ime kontekstno-neodvisna gramatika izvira iz oblike produkcijskih pravil. Na levi strani produkcijskega pravila mora vedno stati samo spremenljivka. Torej vsebuje samo pravila oblike

$$A \to \alpha,$$

kjer je $A \in V$ in $\alpha \in (V \cup \Sigma)^*$. Ne sme pa vsebovati pravila oblike

$$\alpha A \gamma \to \alpha\beta\gamma,$$

kjer je $A \in V$ in so $\alpha, \beta, \gamma \in (V \cup \Sigma)^*$, saj je uporaba pravila odvisno od konteksta nekončnega simbola A. Kontekst določa niza α in β, ki se nahajata neposredno pred in po nekončnim simbolom A.

Pripomnimo, da so vse gramatike v delu kontekstno-neodvisne gramatike, zato jih bomo imenovali tudi samo gramatike.

Ko imamo opravka s kontekstno-neodvisnimi gramatikami je priročno, da so zapisane v preprosti obliki. Preprosta in zelo uporabna je *normalna oblika Chomskega*.

Definicija 1.10. Kontekstno neodvisna gramatika je v *normalni obliki Chomskega*, če so vsa produkcijska pravila oblike:

$$A \to BC,$$

$$A \to a,$$

$$S \to \varepsilon,$$

kjer so $A, B, C \in V$; $B, C \neq S$ in $a \in \Sigma$.

Izrek 1.11. *Za vsak kontekstno-neodvisen jezik, obstaja kontekstno-neodvisna gramatika v normalni obliki Chomskega, ki generira ta jezik.*

Dokaz. Začnemo z neko kontekstno-neodvisno gramatiko, ki generira ta jezik in jo preoblikujemo po sledečem postopku.

Najprej dodamo nov začetni simbol S_0 in pravilo $S_0 \to S$. Tako zagotovimo, da se začetni simbol S_0 ne pojavi kot levi člane kateregakoli pravila.

Sedaj poskrbimo za pravila, ki vsebujejo ε. Odstranimo pravila oblike $A \to \varepsilon$, kjer $A \neq S_0$ in za vsako pojavitev simbola A kot desnega člana pravila dodamo novo pravilo, kjer odstranimo tisto pojavitev simbola A. Korak ponavljamo, dokler ne odstranimo vseh takšnih pravil.

Nato odstranimo pravila oblike $A \to B$ in za vsako pravilo $B \to \alpha$ dodamo pravilo $A \to \alpha$, razen če smo to pravilo pred tem že odstranili. Korak ponavljamo, dokler ne odstranimo vseh takšnih pravil.

Nazadnje še preostala pravila preoblikujemo v primerno obliko. Vsako pravilo oblike $A \to y_1 y_2 \dots y_k$, kjer je $k \geq 3$ in je y_i nekončni ali končni simbol za vsak $i = 1, 2, \dots k$, zamenjamo s pravili $A \to y_1 A_1$, $A_1 \to y_2 A_2$, $\dots$, $A_{k-3} \to y_{k-2} A_{k-2}$ in $A_{k-2} \to y_{k-1} y_k$. V dodanih pravilih končne simbole y_i zamenjamo z nekončnimi simboli C_i in dodamo pravila $U_i \to y_i$. Nekončne simbole pustimo pri miru. □

Opomba 1.12. Ko odstranjujemo pravila oblike $A \to \varepsilon$, moramo upoštevati vsako pojavitev simbola A kot desnega člana pravila. Torej, če odstranimo pravilo oblike $A \to \varepsilon$ in imamo pravilo $B \to \alpha A \beta$, dodamo pravilo $B \to \alpha\beta$. Če imamo pravilo oblike $B \to \alpha A \beta A \gamma$, dodamo pravila $B \to \alpha\beta A\gamma$, $B \to \alpha A \beta\gamma$, $B \to \alpha\beta\gamma$. Če imamo pravilo $B \to A$, dodamo pravilo $B \to \varepsilon$, razen če smo pred tem že odstranili pravilo $B \to \varepsilon$.

Primer 1.13. ♢

2. Dopustne gramatike

Recimo, da želimo stisniti niz w. Ideja je, da poiščemo gramatiko G_w, ki generira enojec $\{w\}$ za svoj jezik. Med njimi poiščemo "najmanjšo" oziroma "najlepšo" in jo kodiramo. Ker gramatike G_w generira w in je "majhna", je kodirana v "kratko" kodo. Tako bomo posredno preko gramatike "dobro" stisnili niz w.

V razdelku definiramo podmnožico kontekstno-neodvisnih gramatik, katerih jezik je le neprazen enojec, in jih imenujemo *dopustne gramatike*.

Pri določanju ali je dana gramatika dopustna in kaj je jezik te gramatike si bomo pomagali še z dvema konceptoma iz teorije formalnih jezikov in sicer s *D0L sistemom gramatike G* in *Izpeljevalnim grafom gramatike G*.

Deterministične gramatike.

Definicija 2.1. Kontekstno-neodvisna gramatika G je *deterministična*, če vsak nekončen simbol $A \in V$, nastopa natanko enkrat kot levi član nekega produkcijskega pravila P, oziroma

$$\forall A \in V \ \exists! \alpha \in (V \cup \Sigma)^* \colon (A, \alpha) \in P.$$

Kontekstno-neodvisna gramatika, ki ni deterministična, je *nedeterministična*.

Deterističnost gramatike nam zagotovi, da ko preberem vhodni niz in se odločimo, da uporabimo produkcijsko pravilo, katerega levi član je $A \in T(G)$, je niz, ki ga prepišemo s pravilom, točno določen. V nedeterministični gramatiki naslednji niz ni nujno že določen, saj je lahko več pravil, ki imajo A za levega člana, med katerimi izbiramo.

Trditev 2.2. *Naj bo G deterministična kontekstno-neodvisna gramatika. Potem je jezik gramatike G enojec ali pa prazna množica.*

Dokaz. □

Poglejmo si dva preprosta primera deterministične kontekstno-neodvisne gramatike, ki ima za jezik le prazno množico.

Primer 2.3. Naj bo G kontekstno-neodvisna in

$$V = \{S\},$$
$$\Sigma = \{a\},$$
$$P = \{S \to S\},$$
$$S = S.$$

Gramatika je očitno deterministična Jezik gramatike je $L(G) = \emptyset$. saj ne moremo izpeljati nobenega niza, ki bi vseboval končne simbole. ♢

Primer 2.4. Naj bo G kontekstno-neodvisna in

$$V = \{S\},$$
$$\Sigma = \{a\},$$
$$P = \{S \to Aa, A \to Ba, B \to A\},$$
$$S = S.$$

Gramatika je očitno deterministična. Jezik gramatike je $L(G) = \emptyset$, saj ne moremo izpeljati nobenega niza, ki bi vseboval le končne simbole. Z uporabo končno mnogo produkcijskih pravil se le ciklamo med A in B, pri tem nam število ponovitev končnega simbola a pove kolikokrat smo uporabili pravila. ♢

D0L-sistem. Podobno kot nas je pri uvedbi gramatike motivirala slovnica, nas sedaj motivira biologija. Procesi v biologiji potekajo istočasno, recimo proces razmnoževanja bakterij ali rast rastlin. Poizkušamo opisati dinamičen proces, ki je odvisen od časa. Takšne procese opišemo z *Lindenmayerjevim sistemom*, krajše *L-sistemom*, ki posveča več pozornost zaporedju nizov, oziroma prepisovanju niza s produkcijskimi pravili, kot statičnim množicam nizov. V matematičnem smislu se bomo posvetili endomorfizmu definiranem na prostem monoidu.

Spoznali bomo poseben primer determinističnega kontekstno-neodvisnega L-sistema, ki se imenuje *D0L sistem*.

Definicija 2.5. Naj bo Σ abeceda. *Endomorfizem množice* Σ^* je preslikava $f\colon \Sigma^* \to \Sigma^*$ tako, da

$$f(\varepsilon) = \varepsilon,$$
$$\forall w, u \in \Sigma^*\colon f(w)f(u).$$

Za endomorfizem f množice Σ^* induktivno definiramo

$$f^0(w) = w,$$
$$f^1(w) = f(w),$$
$$f^k = f(f^{k1}(w)),$$

kjer je $w \in \Sigma^*$ in $k \geq 2$ celo število.

Opomba 2.6. Endomorfizem f na Σ^* je natančno določen, ko za vsako črko $a \in \Sigma$ podamo njeno preslikavo $f(a) \in \Sigma^*$.

Definicija 2.7. *D0L sistem* je trojica $D = (\Sigma, f, w)$, kjer je Σ abeceda; f endomorfizem množice Σ^*; in $w \in \Sigma^*$ *aksiom*. Sistem generira zaporedje nizov $\{f^k(w) \mid k = 0, 1, 2 \ldots\}$, ki ima *fiksno točko* w^*, če velja

$$w^* \in \{f^k(w) \mid k = 0, 1, 2 \ldots\},$$
$$f(w^*) = w^*.$$

Opomba 2.8. Za splošni L-sistem v zgornji definiciji zamenjamo homomorfizem f z množico produkcijskih pravil P in predpostavimo, da vsebuje produkcijsko pravilo identitete. Kontekstna-neodvisnost in determinističnost L-sistema je definirana enako kot pri gramatikah.

L-sistemi se uporabljajo za modeliranje morfologije bitji prav tako z njimi generiramo fraktale. Za generiranje realističnih modelov so zanimivi stohastični L-sistemi, ki v vsakem koraku zaporedja nizov z neko verjetnostjo uporabijo produkcijsko pravilo.

S pomočjo naslednjega endomorfizma bomo jezik deterministične kontekstno-neodvisne gramatike karakterizirali preko fiksne točke pripadajočega D0L sistema.

Definicija 2.9. Naj bo G deterministična kontekstno-neodvisna gramatika v kateri prazen niz ne nastopa kot desni član kateregakoli produkcijska pravila. Na $(V \cup \Sigma)^*$ definiramo endomorfizem f_G tako, da

$$\forall a \in \Sigma \colon f_G(a) = a;$$
$$A \to \alpha \Rightarrow f_G(A) = \alpha.$$

D0L sistem $(V \cup \Sigma, f_G, S)$ označimo z $D(G)$ in ga imenujemo *D0L sistem prirejen gramatiki G*.

Izrek 2.10. *Naj bo G dopustna kontekstno-neodvisna gramatika. Potem jezik gramatike G ustreza fiksni točki D0L sistema prirejenega gramatiki G.*

Dokaz. □

Izpeljevalni graf.

Definicija 2.11. Naj bo G kontekstno-neodvisna gramatika v kateri prazen niz ne nastopa kot desni član kateregakoli produkcijska pravila. *Izpeljevalni graf gramatike G* je usmerjen graf. Množica vozlišč grafa ustreza $V \cup \Sigma$. Naj bo produkcijsko pravilo

$$A \to \alpha = y_1 y_2 \cdots y_m,$$

kjer so $y_1, y_2, \ldots, y_m \in V \cup \Sigma$, potem iz vozlišča A izvirajo usmerjene povezave v vozlišča $y_1, y_2, \ldots, y_m$.

Primer 2.12. Poglejmo si izpeljevali graf gramatike iz primera. ♢

Dopustne gramatike. Ker bomo gramatiko stiskali, želimo da je čim "manjša". Ne želimo odvečnih simbolov. Gramatika, ki ne vsebuje neuporabnih simbolov, nam zagotavlja, da je vsak nekončni in vsak končni simbol prisoten vsaj v eni izpeljavi niza, ki je v jeziku gramatike.

Definicija 2.13. Pravimo, da kontekstno-neodvisna gramatika G *ne vsebuje neuporabnih simbolov*, ko za vsak simbol $T \in V \cup \Sigma$, $T \neq S$ obstaja končno mnogo nizov $\alpha_1, \alpha_2, \ldots, \alpha_n \in (V \cup \Sigma)^*$ tako, da je T vsebovan vsaj v enem izmed nizov in velja

$$S = \alpha_1 \Rightarrow \alpha_2 \Rightarrow \cdots \alpha_n \in L(G).$$

Definicija 2.14. Kontekstno-neodvisna gramatika G je *dopustna*, če je deterministična, ne vsebuje neuporabnih simbolov, $L(G) \neq \emptyset$ in prazen niz ne nastopa kot desni član kateregakoli produkcijska pravila v P.

Posledica 2.15. *Jezik dopustne kontekstno-neodvisne gramatike je enojec.*

Da določimo dopustno kontekstno-neodvisno gramatiko je dovolj, da podamo le produkcijska pravila, saj lahko iz njih enolično določimo V, Σ in S. Nekončni simboli gramatike so levi člani produkcijskih pravil, končni simboli gramatike so desni člani produkcijski pravil, ki niso tudi levi člani kateregakoli produkcijska pravila in začetni simbol je nekončni simbol, ki ne nastopa kot desni član kateregakoli produkcijskega pravila.

Primer 2.16. Podana imamo produkcijska pravila

$$A_0 \to aA_1A_2A_3$$
$$A_1 \to ab$$
$$A_2 \to A_1b$$
$$A_3 \to A_2b$$

Levi člani produkcijskih pravil so nekončni simboli gramatike,

$$V = \{A_0, A_1, A_2, A_3\}.$$

Desni člani produkcijskih pravil, ki niso tudi levi člani, so končni simboli gramatike,

$$\Sigma = \{a, b\}.$$

Izmed nekončnih simbolov je A_0 edini, ki ne nastopa kot desni član kateregakoli produkcijskega pravila, torej je začetni simbol,

$$S = A_0.$$

Pripomnimo, da je gramatika podana s temi produkcijskimi pravili dopustna kontekstno-neodvisna gramatika, kar bomo preverili kasneje, in da je $L(G) = \{aababbabbb\}$. ♢

Sedaj lahko karakteriziramo dopustne gramatike tudi preko pripadajočega D0L sistema in izpeljevalnega grafa gramatike.

Izrek 2.17. *Naj bo G kontekstno-neodvisna gramatika v kateri prazen niz ne nastopa kot desni član kateregakoli produkcijska pravila. Potem so naslednje trditve ekvivalentne*

(1) *Gramatika G je dopustna.*

(2) *Izpeljevalno drevo $D(G)$ je aciklično in koren S.*

(3) $f_G^{|V|}(S) \in T^+$ *in vsak simbol iz* $V \cup T$ *nastopa v vsaj enem izmed nizov* $f_G^i(S)$ *za* $i = 0, 1, \ldots, |V|$

Dokaz. □

Sledeči izrek, ki sledi iz izreka 2.17, nam poda algoritem za izračun enojca dopustne gramatike.

Izrek 2.18. *Jezik dopustne kontekstno-neodvisna gramatike G je*

$$L(G) = \{f_G^{|V|}(S)\}.$$