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 nizev, 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 $*$ asociatna binarna operacija ma množici vseh končnih nizov Σ^* tako, da je prazen niz ε nevtralen 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: 6. februar 2023.

Definicija 1.5. *Kontektsno-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

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

Naj bo G kontekstno-neodvisna gramatika. Ker je relacija P celovita, za vsak nekončni simbol $A \in V$ obstaja končen niz nekončnih in končnih simbolov $\alpha \in (V \cup \Sigma)^*$, da je (A, α) produkcijsko pravilo. Torej $(A, \alpha) \in P$, pišemo

$$A \to \alpha.$$

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 simobolov, ki jih lahko izpeljemo z uporabo produkcijskih pravil 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. 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 pogosto imenovali samo gramatike.

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" 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, ki za svoj jezik generirajo le 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 z *Izpeljevalnim grafom gramatike* G in $D0L$- *sistemom gramatike* G.

Deterministične gramatike.

Definicija 2.1. Za kontekstno-neodvisna gramatika G označimo z $V(G)$ množico nekončnih simbolov, s $\Sigma(G)$ množico končnih simbolo, s $P(G)$ množico produkcijskih pravil in z $S(G)$ začetni simbol.

Definicija 2.2. Kontekstno-neodvisna gramatika G je *deterministična*, če za vsak nekončen simbol $A \in V(G)$, nastopa natanko enkrat kot levi član nekega produkcijskega pravila $P(G)$, oziroma

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

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

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

Trditev 2.3. *Naj bo G deterministična kontekstno-neodvisna gramatika. Potem je jezik gramatike G enojec ali prazni niz.*

Dokaz. □

Primer 2.4. Naj bo G kontekstno-neodvisna gramatika in

$$V(G) = \{S, A, B, C\},$$
$$\Sigma(G) = \{a, b, c\},$$
$$P(G) = \{S \to Ac, A \to Bc, A \to Ac, B \to bb, B \to cc, C \to B, C \to b\},$$
$$S(G) = S.$$

Podana gramatika ni deterministična, saj imamo dve produkcijski pravili, katerih levi član je A.

♢

Definicija 2.5. Pravimo, da kontekstno-neodvisna gramatika G *ne vsebuje neuporabnih simbolov*, ko za vsak simbol $T \in V(G) \cup \Sigma(G),\ T \neq S$ 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

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

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 2.6. 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 $P(G)$.

Opomba 2.7. Naj bo G kontekstno-neodvisna gramatika. Prazen niz ne nastopa kot desni član kateregakoli produkcijska pravila v $P(G)$, ko

$$\nexists A \in V(G) \colon (A, \varepsilon) \in P(G).$$

Posledica 2.8. *Jezik dopustne kontekstno-neodvisne gramatike je enojec.*

Da določimo dopustno kontekstno-neodvisno gramatiko je dovolj, da podamo le produkcijska pravila, saj lahko iz njih enoločno določimo $V(G)$, $\Sigma(G)$ in $S(G)$. 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 2.9. 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,

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

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

$$\Sigma(G) = \{a, b\}.$$

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

$$S(G) = A_0.$$

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

♢

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 pozornost zaporedju nizov, oziroma prepisovanju niza z produkcijskimi pravili, kot statičnim množicam nizov. V matematičnem smislu se bomo posvetili endomorfizmu definiranem na prostem monoidu.

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

Definicija 2.10. Naj bo Σ abeceda. *Endomofizem 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 2.11. Endomorfizem f na Σ^* je natančno določen ko za vsako črko $a \in \Sigma$ podamo njeno preslikao $f(a) \in \Sigma^*$.

Definicija 2.12. *D0L sistem* je je trojica $D = (\Sigma, f, w)$, kjer je Σ abeceda; f endomorfizem množice Σ^*; in $w \in \Sigma^*$ imneujemo *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 2.13. 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 determinstičnost L-sistema je definirana enako kot pri gramatikah.

L-sistemi se uporabljajo za modeliranje morfologije bitji prav tako z njimi generiramo fraktale. Za generirane 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 karakterzirali preko fiksne točke pripadajočega $D0L$-sistema.

Definicija 2.14. Naj bo G deterministična kontekstno-neodvisna gramatika v kateri prazen niz ne nastopa kot desni član kateregakoli produkcijska pravila. Na $(V(G) \cup \Sigma(G))^*$ definiramo endomorfizem f_G tako, da

$$\forall a \in \Sigma \colon f_G(a) = a;$$
$$\text{če je } A \to \alpha \text{ produkcijsko pravilo, potem je } f_G(A) = \alpha.$$

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

Izrek 2.15. *Naj bo G dopustna kontekstno-neodvisna gramatika. Potem jezik gramatike G ustreza fiksni točki $D0L$-sistema prirejenga gramatiki G.*

Dokaz. □