

KONTEKSTNO-NEODVISNE GRAMATIKE ZA KODIRANJE IN STISKANJE PODATKOV

JANEZ PODLOGAR

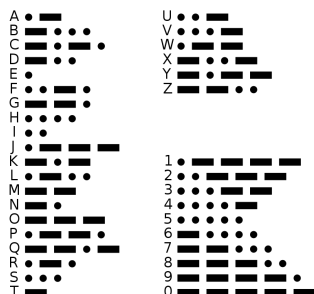
POVZETEK. V delu predstavimo motivacijo in definicije, ki so potrebni za obravnavanje stiskanja podatkov s kontekstno-neodvisnimi gramatikami.

1. KODIRANJE PODATKOV

Zapis informacije v neki obliki ni primeren za vsakršno rabo. Besedilo, zapisano z pismenkami, je neberljivo za slepe osebe, saj je komunikacijski kanal v tem primeru vid. Prav tako pisanega besedila v pravotni obliki ni moč poslati z telegrafom. V tem primeru je komunikacijski kanal žica in pismenke se po njej ne morejo sprehoditi. V obeh primerih je informacija, ki bi jo radi prenesli v neprimerni obliki. V prvem primeru je potrebno besedilo zapisati z Braillovo pisavo. V drugem primeru pa je potrebno besedilo pretvoriti v električni signal. Spreminjanje zapisa sporočila pravimo *kodiranje*, sistemu pravil, po katerem se kodiranje opravi, pa *kod*.

Primer 1.1. *Morsejeva abeceda* je kodiranje črk, števil in ločil s pomočjo zaporedja kratkih in dolgih signalov:

- Dolžina kratkega signala je ena enota.
- Dolgi signal je trikrat daljši od kratkega signala.
- Razmiki med signali znotraj znaka so dolžine kratkega signala.
- Presledki med znaki so dolgi tri kratke signale oz. en dolgi signal.
- Presledki med besedami so dolgi sedem kratkih signalov.



SLIKA 1. Mednarodna Morsejeva abeceda

Prvotni namen Morsejeve abecede je komunikacija preko telegrama, saj komunikacijski kanal dovoljuje le električne signale in tišino med njimi. Kodiranje črk je takšno, da imajo črke z višjo frekvenco (v angleškem jeziku) krajši zapis, s tem se dolžina kodiranega sporočila skrajša in posledično tudi čas prenosa.

◇

Definicija 1.2. *Abeceda* je neprazna množica Σ . Elementom abecede pravimo *črke*. *Množica vseh končnih besed abecede* Σ je

$$\Sigma^* = \{ a_1 a_2 a_3 \cdots a_n \mid n \in \mathbb{N}_0 \wedge \forall i : a_i \in \Sigma \},$$

kjer za $n = 0$ dobimo prazno besedo, ki jo označimo z ε . *Dolžina besede* w označimo z $|w|$ in je enaka številu črk v besedi $w \in \Sigma^*$. *Jezik na abecedi* Σ je poljubna podmnožica množice vseh besed Σ^* .

Opomba 1.3. V splošnem ima lahko abeceda poljubno kardinalnost, v diplomski nalogi se bomo srečali le s končnimi abecedami.

Opomba 1.4. *Kleenejeva zvezdica* ali *Kleenejevo zaprtje* je operacija, ki abecedi Σ priredi najmanjšo nadmnožico Σ^* , ki vsebuje *prazno besedo* ε in je zaprta za konkatenacijo oziroma veriženje. Z drugimi besedami, Σ^* je množica vseh končnih besed oziroma končnih nizov, ki jih lahko generiramo z veriženjem črk abecede Σ .

Za abecedo Σ definirajmo

$$\Sigma^0 = \{\varepsilon\}$$

$$\Sigma^1 = \Sigma$$

ter za vsak $i > 0$ rekurzivno

$$\Sigma^{i+1} = \{ wa \mid w \in \Sigma^i \text{ in } a \in \Sigma \},$$

potem je Kleenejeva zvezdica na Σ enaka

$$\Sigma^* = \bigcup_{i \geq 0} \Sigma^i$$

Primer 1.5. Naj bo $\Sigma = \{a, b, c\}$ abeceda, potem je

$$ab \in \Sigma^*$$

$$ccc \in \Sigma^*$$

$$cababccababcccab \in \Sigma^*$$

◇

Definicija 1.6. *Kodiranje besed abecede* Σ je obrnljiva funkcija $K: \Sigma^* \rightarrow \Sigma_c^*$, kjer je Σ_c *kodirna abeceda*. $K(w)$ imenujemo *koda besede* w . *Dekodiranje* je obraten postopek, ki za podano kodo $c \in \Sigma_c^*$ vrne prvotno besedo $w \in \Sigma^*$, ki smo jo kodirali.

Primer 1.7. Formalizirajmo Morsejevo abecedo iz Primera 1.1. Abecedi sta

$$\Sigma = \{A, B, \dots, Z\} \cup \{0, 1, \dots, 9\} \cup \{ _ \}$$

$$\Sigma_c = \{., -, _, ____ \}$$

Definirajmo kodno funkcijo črk abecede $\kappa: \Sigma \rightarrow \Sigma_c^*$, ki vsakei črki iz abecede priredi besedo iz črk kodirne abecede. Funkcija je definirana s tabelo iz Slike 1, dodatno $\kappa(_) = ____$. Za besedo $w = a_1 a_2 \dots a_n \in \Sigma^*$ definiramo kodirno funkcijo K po črkah

$$K(w) = \kappa(a_1) \kappa(a_2) \cdots \kappa(a_n)$$

Opazimo, da je Morsejeva abeceda kodiranja brez izgube.

◇

2. STISKANJE PODATKOV

Eden izmed namen kodiranja je tudi doseči čim večjo ekonomičnost zapisa. Želimo, da bi bilo naše sporočilo čim krajše. Kodiranje, ki skrajša zapis podatkov, imenujemo *stiskanje podatkov*.

Definicija 2.1. *Stiskanje* je kodiranje K za katerega velja

$$\forall w \in \Sigma^*: |K(w)| \ll |w|$$

Primer 2.2. Za abecedo vzemimo $\Sigma = \{a, b, c\}$ in pogledimo besedo

$$w = cababcccababcccab$$

Opazimo, da se nam v besedi w večkrat ponovita vzorca ab in ccc . Zato uvedemo novi spremenljivki $A_1 = ab$ in $A_2 = ccc$. Sedaj lahko zapišemo w kot

$$w = cA_1A_1A_2A_1A_1A_2A_1$$

Ponovno se nam pojavi vzorec, tokrat $A_1A_1A_2$. Uvedemo novo spremenljivko $A_3 = A_1A_1A_2$ in zapišemo w kot

$$w = cA_3A_3A_1$$

Prvotno besedilo smo z novimi spremenljivkami uspešno skrajšali. Kot bomo videli, smo pretvorili besedo w v kontekstno neodvisno gramatiko G_w s produkcijskimi pravili

$$A_0 \rightarrow cA_3A_3A_1,$$

$$A_1 \rightarrow ab,$$

$$A_2 \rightarrow ccc,$$

$$A_3 \rightarrow A_1A_1A_2$$

◇

3. KONTEKSTNO-NEODVISNE GRAMATIKE

Definicija 3.1. *Formalna gramatika* G so pravila, ki nam iz abecede Σ tvorijo jezik $L(G)$

Definicija 3.2. *Kontekstno-neodvisna gramatika* je četverica $G = (V, \Sigma, P, S)$, kjer je V končna množica spremenljivk, abeceda Σ množica končnih simbolov tako, da $\Sigma \cap V = \emptyset$, $P \subseteq V \times (V \cup \Sigma)^*$ množica produkcijskih pravil in $S \in V$ začetna spremenljivka.

Definicija 3.3. Naj bo $G = (V, \Sigma, P, S)$ kontekstno-neodvisna gramatika. Naj bodo $\alpha, \beta, \gamma \in (V \cup \Sigma)^*$ nizi spremenljivk in končnih simbolov, $A \in V$ spremenljivka ter naj bo $(A, \beta) \in P$ produkcijsko pravilo, označimo ga z $A \rightarrow \beta$. 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 *porodi* β , če je $\alpha = \beta$ ali če za $k \geq 0$ obstaja zaporedje $\alpha_1, \alpha_2, \dots, \alpha_n \in (V \cup \Sigma)^*$ tako, da

$$\alpha \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \dots \Rightarrow \alpha_n \Rightarrow \beta$$

in pišemo $\alpha \xRightarrow{*} \beta$.

Posledica 3.4. *Jezik kontekstno neodvisne gramatike* G *je*

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

Opomba 3.5. Ime kontekstno-neodvisna gramatika izvira iz oblike produkcijskih pravil. Na levi strani produkcijskega pravila mora vedno stati samo spremenljika. Torej vsebuje samo pravila oblike

$$A \rightarrow \alpha,$$

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

$$\alpha A \gamma \rightarrow \alpha \beta \gamma,$$

kjer je $A \in V$ in so $\alpha, \beta, \gamma \in (V \cup \Sigma)^*$, saj je pravilo odvisno od predhodnega konteksta. Odvisno je od tega ali pred njim stoji α in za njim β .

Primer 3.6. Formalizirajmo gramatiko iz Primera 2.2, ki smo jo generirali z besede $w = cababcccababcccab$. Označimo jo z $G_w = (V, \Sigma, P, S)$, kjer je

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

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

$$P = \{ A_0 \rightarrow cA_3A_3A_1, A_1 \rightarrow ab, A_2 \rightarrow ccc, A_3 \rightarrow A_1A_1A_2 \},$$

$$S = A_0$$

Vidimo, da G_w ustreza naši definiciji kontekstno-neodvisne gramatike in res kodira w , saj je

$$L(G_w) = w$$

◇