KONTEKSTNO-NEODVISNE GRAMATIKE ZA KODIRANJE IN STISKANJE PODATKOV

JANEZ PODLOGAR

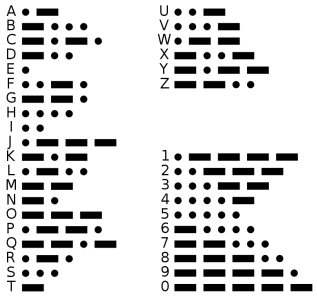
Povzetek. V delu predstavimo motivacijo in definicije, ki so potrebni za obravnavnje stiskanja podatkov s kontekstno-neodvisnimi gramatikami.

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 besedil 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.

♢

Date: 28. oktober 2022.

Definicija 1.2. *Abeceda* je neprazna množica Σ. *Množica vseh končnih besed na abecedi* Σ je

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

kjer ε prazna beseda. 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 z končnimi abecedami.

Primer 1.4. Naj bo $\Sigma = \{\, a, b, c \,\}$ abeceda, potem je

$$ab \in \Sigma^*$$
$$ccc \in \Sigma^*$$
$$cababcccababcccab \in \Sigma^*$$

♢

2. Stiskanje podatkov

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

Primer 2.1. Ponovno za abecedo vzemimo $\Sigma = \{\, a, b, c \,\}$ in poglejmo besedo

$$\mathbf{x} = cababcccababcccab$$

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

$$\mathbf{x} = cA_1A_1A_2A_1A_1A_2A_1$$

Ponovno se nam pojavi vzorec, tokrat $A_1A_1A_2$. Uvedemo novo spremeljivko $A_3 = A_1A_1A_2$ in zapišemo $\mathbf{x}$ kot

$$\mathbf{x} = cA_3A_3A_1$$

Prvotno besedilo smo z novimi spremeljivkami uspešno skrajšali. Kot bomo videli kmalu, smo pretvorili besedo $\mathbf{x}$ v kontekstno neodvisno gramatiko $\mathbf{G_x}$ z produkcijskimi pravili

$$\begin{aligned} A_0 &\to cA_3A_3A_1, \\ A_1 &\to ab, \\ A_2 &\to ccc, \\ A_3 &\to A_1A_1A_2. \end{aligned}$$

♢

3. Kontekstno-neodvisne gramatike

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

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

Definicija 3.3. Naj bo $G = (V, \Sigma, P, S)$ kontekstno-neodvisna gramatika. Naj bodo $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in (V \cup \Sigma)^*$ nizi spremenljivk in končnih simbolov, $A \in V$ spremenljivka ter naj bo $(A, \mathbf{y}) \in P$ produkcijsko pravilo, označimo ga z $A \to \mathbf{y}$. Pravimo, da $\mathbf{x}A\mathbf{z}$ *pridela* $\mathbf{xyz}$, pišemo $\mathbf{x}A\mathbf{z} \Rightarrow \mathbf{xyz}$. Pravimo, da $\mathbf{x}$ *porodi* $\mathbf{y}$, če je $\mathbf{x} = \mathbf{y}$ ali če za $k \geq 0$ obstaja zaporedje nizov $\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_n} \in (V \cup \Sigma)^*$ tako, da

$$\mathbf{x} \Rightarrow \mathbf{x} \Rightarrow \mathbf{x_1} \Rightarrow \ldots \Rightarrow \mathbf{x_n} \Rightarrow \mathbf{y}$$

in pišemo $\mathbf{x} \overset{*}{\Rightarrow} \mathbf{y}$.

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

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

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

$$A\mathbf{u} \to \mathbf{v},$$

kjer je $A \in V$ in $\mathbf{u}$, $\mathbf{v} \in \Sigma$, saj bo pravilo pridelalo $\mathbf{v}$ saj samo, če se niz končana $\mathbf{u}$. Torej je odvisno od predhodnega konteksta.

Primer 3.6. Formalizirajmo gramatiko iz Primera 2.1, ki smo jo generirali iz $\mathbf{x}$ Označimo jo z $\mathbf{G_x} = (V, \Sigma, P, S)$, kjer je

$$\begin{aligned}
&V = \{ A_0, A_1, A_2, A_3 \},\\
&\Sigma = \{ a, b, c \},\\
&P = \{ A_0 \to cA_3A_3A_1, A_1 \to ab, A_2 \to ccc, A_3 \to A_1A_1A_2 \},\\
&S = A_0.
\end{aligned}$$

Vidimo, da $\mathbf{G_x}$ ustreza naši definiciji kontekstno-neodvisne gramatike in res kodira $\mathbf{x}$, saj je

$$\mathbf{L}(\mathbf{G_x}) = \mathbf{x}$$

$\diamondsuit$