Kontekstno-neodvisne gramatike za kodiranje in stiskanje podatkov

Janez Podlogar

Univerza v Ljubljani, Fakulteta za matematiko in fiziko

16. 9. 2024

Kontekstno-neodvisne gramatike
za kodiranje in stiskanje podatkov

Janez Podlogar

Univerza v Ljubljani, Fakulteta za matematiko in fiziko

16. 9. 2024

- Kodiranje je spreminjanje zapisa sporočila.
- Stiskanje podatkov je zapis sporočilo v zgoščeni obliki.
- Gramatike je nabor pravil za tvorjenje nizov.

- Kodiranje je spreminjanje zapisa sporočila.
- Stiskanje podatkov je zapis sporočilo v zgoščeni obliki.
- Gramatike je nabor pravil za tvorjenje nizov.

- Ideja je: Stistniti niz z "majhno" gramatiko.

Primer

Naj bo $V = \{S\}, \Sigma = \{a, b, c\}, P = \{S \to aSb, S \to \epsilon\}$ in $S = S$.
Izpeljemo nize

$$S \Rightarrow \epsilon,\ S \Rightarrow aSb \Rightarrow ab,\ S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aabb,\ \ldots$$

- Elementom P pravimo *prepisovalna pravila*.
- Prazen niz ε je enota za operacijo stik.
- Niz $\alpha A \gamma$ se *prepiše s pravilom* $A \to \beta$ v $\alpha\beta\gamma$, pišemo $\alpha A \gamma \Rightarrow \alpha\beta\gamma$.
- Niz α *izpelje* niz β, če lahko α prepišemo v β z uporabo končno mnogo prepisovalnih pravil.

Kontekstno-neodvisna gramatika

Definicija

Kontekstno-neodvisna gramatika je četverica $G = (V, \Sigma, P, S)$, kjer je

- V abeceda *nekončnih simbolov*;
- Σ abeceda *končnih simbolov* taka, da $\Sigma \cap V = \emptyset$;
- $P \subseteq V \times (V \cup \Sigma)^*$ celovita relacija;
- $S \in V$ *začetni simbol*.

Kontekstno-neodvisna gramatika
Definicija
Kontekstno-neodvisna gramatika je četverica $G = (V, \Sigma, P, S)$, kjer je
- V abeceda *nekončnih simbolov*;
- Σ abeceda *končnih simbolov* taka, da $\Sigma \cap V = \emptyset$;
- $P \subseteq V \times (V \cup \Sigma)^*$ celovita relacija;
- $S \in V$ *začetni simbol*.

- Abeceda je končna neprazna množica.
- $(V \cup \Sigma)^*$ je množica vseh nizov končne dolžine končnih in nekončnih simbolov. Vsebuje tudi prazen niz ε, saj je to niz dolžine 0.
- Okrajšamo z KNG.

Primer

Jezik je $\{a^n b^n \mid n \geq 0\}$.

Kontekstno-neodvisna gramatika

Definicija

Kontekstno-neodvisna gramatika je četverica $G = (V, \Sigma, P, S)$, kjer je

- V abeceda *nekončnih simbolov;*
- Σ abeceda *končnih simbolov* taka, da $\Sigma \cap V = \emptyset$;
- $P \subseteq V \times (V \cup \Sigma)^*$ celovita relacija;
- $S \in V$ *začetni simbol.*

Definicija

Jezik KNG je množica nizov, ki jih lahko izpeljemo iz začetnega simbola in ne vsebujejo nekončnih simbolov.

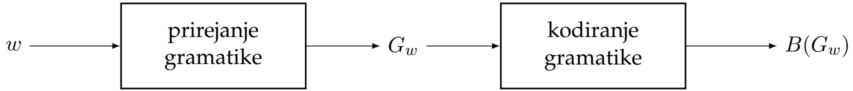
Kontekstno-neodvisna gramatika
Definicija
Kontekstno-neodvisna gramatika je četverica $G = (V, \Sigma, P, S)$, kjer je
V abeceda nekončnih simbolov;
Σ abeceda končnih simbolov taka, da $\Sigma \cap V = \emptyset$;
$P \subseteq V \times (V \cup \Sigma)^*$ celovita relacija;
$S \in V$ začetni simbol.
Definicija
Jezik KNG je množica nizov, ki jih lahko izpeljemo iz začetnega simbola in ne vsebujejo nekončnih simbolov.

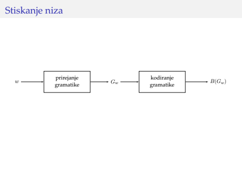
- Abeceda je končna neprazna množica.
- $(V \cup \Sigma)^*$ je množica vseh nizov končne dolžine končnih in nekončnih simbolov. Vsebuje tudi prazen niz ε, saj je to niz dolžine 0.
- Okrajšamo z KNG.

Primer

Jezik je $\{a^n b^n \mid n \geq 0\}$.

Stiskanje niza





- Glavna ideja je, da je jezik G_w je $\{w\}$, saj lahko enolično rekonstruramo niz w.
- Alternativni pristop: KNG G je poznana tako kodirniku kot dekodirniku in jezik KNG G vsebuje vse nize, ki jih želimo stistniti. Posamezni niz stisnemo tako, da stistnemo izpeljavo niza iz začetnega simbola.

Dopustna gramatika

Definicija

KNG je *deterministična*, če vsak nekončen simbol $A \in V$ nastopa natanko enkrat kot leva stran nekega prepisovalnega pravila.

Dopustna gramatika

Definicija
KNG je *deterministična*, če vsak nekončen simbol $A \in V$ nastopa natanko enkrat kot leva stran nekega prepisovalnega pravila.

- Prejšnji primer ni deterministična gramatika.
- Če odstranimo odstarnimo pravilo $S \to aSb$, postane deterministična, a je jezik $\{\varepsilon\}$.
- Če odstranimo $S \to \varepsilon$ postane deterministična, a je njen jezik prazna množica.
- Determinizem sam po sebi ni dovolj močan, da prepreči praznost jezika. Zato bomo od dopustnih gramatik zahtevali, da je njihov jezik neprazen in da ne vsebuje pravila oblike $A \to \varepsilon$.

Dopustna gramatika

Definicija

KNG je *deterministična*, če vsak nekončen simbol $A \in V$ nastopa natanko enkrat kot leva stran nekega prepisovalnega pravila.

Trditev

Jezik deterministične KNG je enojec ali prazna množica.

Dopustna gramatika
Definicija
KNG je *deterministična*, če vsak nekončen simbol $A \in V$ nastopa natanko enkrat kot leva stran nekega prepisovalnega pravila.
Trditev
Jezik deterministične KNG je enojec ali prazna množica.

- Prejšnji primer ni deterministična gramatika.
- Če odstranimo odstarnimo pravilo $S \to aSb$, postane deterministična, a je jezik $\{\varepsilon\}$.
- Če odstranimo $S \to \varepsilon$ postane deterministična, a je njen jezik prazna množica.
- Determinizem sam po sebi ni dovolj močan, da prepreči praznost jezika. Zato bomo od dopustnih gramatik zahtevali, da je njihov jezik neprazen in da ne vsebuje pravila oblike $A \to \varepsilon$.

Definicija

KNG *ne vsebuje neuporabnih simbolov*, če se vsak simbol $y \in V \cup \Sigma,\ y \neq S$ pojavi vsaj enkrat v izpeljavi niza, ki je v jeziku KNG.

Definicija
KNG *ne vsebuje neuporabnih simbolov*, če se vsak simbol $y \in V \cup \Sigma,\ y \neq S$ pojavi vsaj enkrat v izpeljavi niza, ki je v jeziku KNG.

- Povedano drugače: simbol je neupraven, če se ne pojavi v nobeni izpeljavi niza, ki je v jeziku KNG.

Definicija

KNG je *dopustna gramatika*, če je:

- deterministična,
- ne vsebuje neuporabnih simbolov,
- ima neprazen jezik,
- prazen niz ne nastopa kot desna stran kateregakoli prepisovalnega pravila.

Definicija
KNG je *dopustna gramatika*, če je:
- deterministična,
- ne vsebuje neuporabnih simbolov,
- ima neprazen jezik,
- prazen niz ne nastopa kot desna stran kateregakoli prepisovalnega pravila.

Definicija

KNG je *dopustna gramatika*, če je:

- deterministična,
- ne vsebuje neuporabnih simbolov,
- ima neprazen jezik,
- prazen niz ne nastopa kot desna stran kateregakoli prepisovalnega pravila.

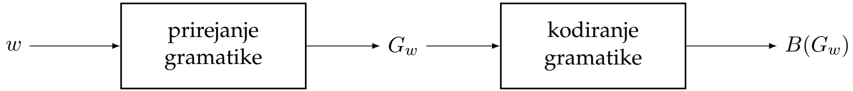
Posledica

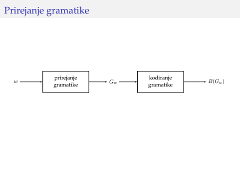
Jezik dopustne gramatike je enojec.

Definicija
KNG je *dopustna gramatika*, če je:
- deterministična,
- ne vsebuje neuporabnih simbolov,
- ima neprazen jezik,
- prazen niz ne nastopa kot desna stran kateregakoli prepisovalnega pravila.

Posledica
Jezik dopustne gramatike je enojec.

Prirejanje gramatike





Asimptotsko kompaktno prirejanje gramatike

Definicija

Prirejanje gramatike nizu abecede $\mathcal{A}$ je *asimptotsko kompaktno*, če za vsak niz $w \in \mathcal{A}^+$ velja

$$\lim_{n\to\infty} \max_{w \in \mathcal{A}^n} \frac{|G_w|}{|w|} = 0.$$

Asimptotsko kompaktno prirejanje gramatike

Definicija

Prirejanje gramatike nizu abecede $\mathcal{A}$ je *asimptotsko kompaktno*, če za vsak niz $w \in \mathcal{A}^+$ velja

$$\lim_{n\to\infty} \max_{w \in \mathcal{A}^n} \frac{|G_w|}{|w|} = 0.$$

- Z $|G_w|$ označimo vsoto dolžin desnih strani prepisovalnih pravil KNG
- Primer je bisekcijsko prirejanje gramatike.

Primer

Naj bo $w = 000101$. Podčrtamo podnize $\underline{0}\,\underline{0}\,\underline{0}\,\underline{1}\,\underline{0}\,\underline{1}$.
$A_w \to A_{0001}A_{01},\; A_{0001} \to A_{00}A_{01},\; A_{00} \to A_0 0,\; A_{01} \to A_0 1,\; A_0 \to 1,$ $A_1 \to 1.$

Neskrčljivo prirejanje gramatike

Definicija

Pravimo, da je KNG G *neskrčljiva gramatika*, če:

1. Vsak $A \in V$, $A \neq S$ nastopa vsaj dvakrat v desni strani prepisovalnih pravil;
2. Ne obstajata $y_1, y_2 \in V \cup \Sigma$, da niz $y_1 y_2$ nastopa kot podniz desne strani kateregali prepisovalnega pravila več kot enkrat na neprekrivajočih se mestih.

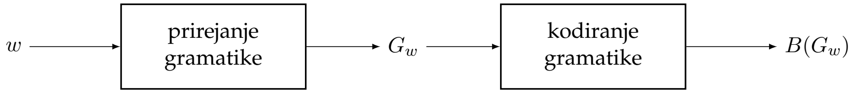
Neskrčljivo prirejanje gramatike
Definicija
Pravimo, da je KNG G *neskrčljiva gramatika*, če:
1. Vsak $A \in V$, $A \neq S$ nastopa vsaj dvakrat v desni strani prepisovalnih pravil;
2. Ne obstajata $y_1, y_2 \in V \cup \Sigma$, da niz $y_1 y_2$ nastopa kot podniz desne strani kateregali prepisovalnega pravila več kot enkrat na neprekrivajočih se mestih.

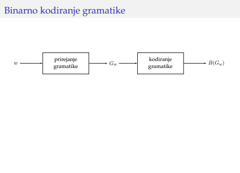
- Primer prekrivajočih mest je $\underline{11}1$ in $1\underline{11}$.
- Primer neprekriavjočih mest je $\underline{11}\ \underline{11}$.
- Primer je metodo najdaljšega ujemajočega podniza.

Primer

Naj bo $w = 00101010001$. Začnemo z trivialno dopustno gramatiko in podčratmo najdaljši podniz, ki se ponovi vsaj dvakrat.
$S \to \underline{001}01010\underline{001}$.
$S \to A\underline{01}\ \underline{01}0A,\ A \to 0\underline{01}$.
$S \to ABB0A,\ A \to 0B,\ B \to 01$.

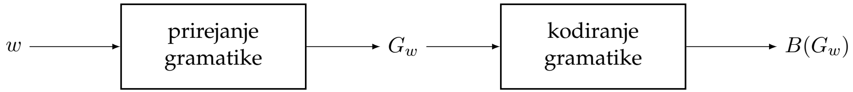
Binarno kodiranje gramatike





- $H(\mathcal{G})$ je entropija KNG G.

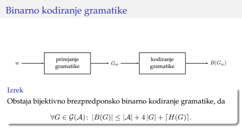
Binarno kodiranje gramatike



Izrek

Obstaja bijektivno brezpredponsko binarno kodiranje gramatike, da

$$\forall G \in \mathcal{G}(\mathcal{A}) \colon \ |B(G)| \le |\mathcal{A}| + 4\,|G| + \lceil H(G) \rceil .$$



- $H(\mathcal{G})$ je entropija KNG G.

Definicija

Stiskanje niza abecede $\mathcal{A}$ z gramatikami je par preslikav kodne in dekodne preslikave $\Phi = (\kappa, \delta)$. Kodna preslikava je

$$\begin{aligned} \kappa \colon \mathcal{A}^+ &\to \{0,1\}^+, \\ w &\mapsto B(\pi(w)), \end{aligned}$$

kjer je π prirejanje gramatike nizu abecede $\mathcal{A}$ in B binarno kodiranje dopustne gramatike.

Definicija
Stiskanje niza abecede $\mathcal{A}$ z gramatikami je par preslikav kodne in dekodne preslikave $\Phi = (\kappa, \delta)$. Kodna preslikava je
$\kappa \colon \mathcal{A}^+ \to \{0,1\}^+$, $w \mapsto B(\pi(w))$,
kjer je π prirejanje gramatike nizu abecede $\mathcal{A}$ in B binarno kodiranje dopustne gramatike.

- Odvečnost meri količino ponavljajočih se ali predvidljivih podatkov znotraj sporočila, ki jih je mogoče odstraniti, da se prihrani prostor, brez izgube informacije.
- Odvečnost stiskanja z asimptotsko kompaktnim prirejanjem konvergira proti 0 v odvisnosti od izbire kodiranja znotraj razreda.
- Stiskanje z neskrčljivim prirejanjem tudi stiskanje z asimptotsko kompaktnim prirejanjem in odvečnost konvergira enakomerno proti 0 za vsa stiskanja z neskrčljivim prirejanjem, vsaj tako hitro kot $\frac{\log_2 \log_2(n)}{\log_2(n)}$ pomnoženo z neko konstanto.