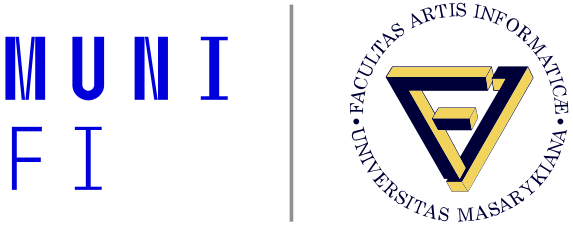

Zpracování přirozeného jazyka

Jindřich Matuška

Faculty of Informatics, Masaryk University

14. listopadu 2024

Čas na odpovědníky

Obsah

Obsah

Zdroje dat

Odkud můžeme čerpat data?

Zdroje dat

Odkud můžeme čerpat data?

- Weby (HTML stránky — Odstranění značek, nebo z nich lze vytáhnout více dat?)
- Knihy (Získání textu ze stránek. Lze vytáhnout více dat?)
- Sociální sítě (Můžeme? Anonymizace?)

Předzpracování dat

- Co je naším cílem?
- Co je pro náš cíl důležité, podstatné?
- Co z dat dokážeme vyčíst? Dokážeme vyčíst něco více než holý text?
- Jsou data dostatečně čistá? Je třeba je vyčistit? Lze je vyčistit?
- Anotace dat?
- V jaké formě budeme data vůbec ukládat?

Obsah

Bezkontextové gramatiky

- Množina terminálních symbolů $\Sigma = \{a, b, c, \ldots\}$
- Množina neterminálních symbolů $N = \{A, B, C, \ldots\}$
- Speciální neterminál (kořen) $S \in N$
- Soubor pravidel $\subseteq N \times V^*$, kde $V = \Sigma \cup N$
  - neterminál $\rightarrow$ libovolný řetězec
  - Pokud neterminál uvozuje více pravidel, používáme svislítko |

Příklad

$$
\begin{aligned}
S &\to NP\ VP, \\
NP &\to Noun \mid Ad\ NP, \\
VP &\to Verb, \\
Noun &\to \text{dítě} \mid \text{člověk} \mid \text{kapsa}, \\
Adj &\to \text{starý} \mid \text{cestující} \mid \text{nové}, \\
Verb &\to \text{píše} \mid \text{sedí} \mid \text{mluví},
\end{aligned}
$$

Syntaktický strom

- Kořenem je kořen gramatiky S
- Listy jsou terminály
- Potomci každého uzlu jsou seřazení
- Pro každý uzel:
  - uzel je terminál, nebo
  - potomci uzlu jsou pravidlem gramatiky
- Může existovat více různých odvození

Tvorba syntaktického stromu z věty se nazývá syntaktický analýza

Příklad

$$
\begin{aligned}
S &\to NP\ VP, \\
NP &\to Noun \mid Adj\ NP, \\
VP &\to Verb, \\
Noun &\to \text{dítě} \mid \text{člověk} \mid \text{kapsa}, \\
Adj &\to \text{starý} \mid \text{cestující} \mid \text{nové}, \\
Verb &\to \text{píše} \mid \text{sedí} \mid \text{mluví},
\end{aligned}
$$

1. „cestující sedí“
2. „nové nové kapsa píše“
3. „starý člověk mluví“

Příklad 9.2.1

Uvažte gramatiku s následujícími pravidly:

$$S \to NP\ VP \mid VP,$$
$$NP \to Noun \mid NP\ Conj\ NP,$$
$$VP \to NP\ Esse$$

a následujícím lexikonem:

$$Noun \to \text{Romulus} \mid \text{Remus} \mid \text{Danubius} \mid \text{fratellus} \mid \text{fratelli} \mid \text{fluvius},$$
$$Esse \to \text{sum} \mid \text{est} \mid \text{sunt} \mid \text{eram} \mid \text{erat} \mid \text{erant},$$
$$Conj \to \text{et}$$

a) Rozhodněte, které z následujících vět lze v gramatice vygenerovat.
   1. „Romulus et Remus fratelli erant“
   2. „Remus et Danubius et Romulus sum“
   3. „Danubius est fluvius“

b) Nalezněte ke každé větě její syntaktický strom, pokud existuje.

Pokrytí vs. přesnost

Pro zamýšlený jazyk L a gramatiku G generující jazyk $L(G)$:

- Pokrytí $\frac{|L \cup L(G)|}{|L|}$
- Přesnost $\frac{|L \cup L(G)|}{|L(G)|}$

Příklad 9.2.3

Uvažme gramatiku G s následujícími pravidly

$$
\begin{aligned}
S &\to AAA \mid BA \mid AB \mid C, \\
A &\to a, \\
B &\to bA \mid Ab, \\
C &\to BA \mid AB \mid cB \mid Bc
\end{aligned}
$$

a jazyk L vět délky 3 sestavených ze slov a, b, c, které obsahují slovo b právě jednou.

a) Je gramatika jednoznačná, neboli existuje pro každou věta jazyka $L(G)$ odvoditelnou v G právě jeden syntaktický strom?

b) Jaké je pokrytí gramatiky G vzhledem k zamýšlenému jazyku L?

c) Jaká je přesnost gramatiky G vzhledem k zamýšlenému jazyku L?