

Základy matematiky a statistiky pro humanitní obory II

Pavel Rychlý Vojtěch Kovář

Fakulta informatiky, Masarykova univerzita
Botanická 68a, 602 00 Brno, Czech Republic

{pary, xkovar3}@fi.muni.cz

3.5.2011

Obsah přednášky

- 1 Statistika a zpracování jazyka
- 2 Vyhledávání kolokací
- 3 N-gramové jazykové modely

Statistika a zpracování jazyka

- Statistika je nástroj, který
 - umožňuje uchopit velké množství dat
 - na základě dat vyvozovat informace o zkoumané oblasti
 - → pravděpodobnosti jevů, predikce
- Velké soubory dat o přirozeném jazyce
 - **jazykové korpusy**
 - v současnosti velikost až 10 miliard slov
 - umožňují statistický popis jevů v jazyce
- Využití statistiky v NLP je obrovské
 - přiblížíme si to dvěma ukázkami

Vyhledávání kolokací

■ Kolokace

- různé definice
- fráze, jejíž význam se neskládá z významů jejích částí
- nějakým způsobem „významné“ spojení dvou slov
- např. idiomy, ale nejen
- základní škola, silný čaj, ...

■ Jakým způsobem vyhledat v korpusu kolokace?

- případně statisticky určit „sílu“ libovolné kolokace na základě dat?
- odlišit „strong tea“ od „powerful tea“

Jakým způsobem vyhledat v korpusu kolokace?

- Prosté frekvence sekvencí slov v korpusu?
 - → „of the”, „in the”, ...
- Frekvence filtrovaných sekvencí slov?
 - na základě slovních druhů jednotlivých slov
 - → „New York”, „United States”, ...
 - ale třeba i „last week”
- T-test
 - aplikace testování hypotéz
 - předpokládáme, že se slova chovají standardně (nulová hypotéza) = podle svých obvyklých pravděpodobnostních rozložení
 - vyvrácení nulové hypotézy = kolokace
- Další – vzájemná informace, logdice, ...

N-gramové jazykové modely

■ N-gramový jazykový model

- „hádáme další slovo“ (značku) na základě předchozích
- $P(w_n | w_1, \dots, w_{n-1})$
- z dat odvodíme pravděpodobnostní rozložení všech možných w_n

■ Použití

- strojový překlad, morfologické značkování, rozpoznávání řeči...

■ Problémy

- pro $N > 4$ většinou výpočetně nezvládnutelné
- „**Snědl** jsem velkou zelenou ...”
- **Data sparseness** – pro slova, která se vyskytují méně často, není dost dat → špatný model