IB047

Úvod do korpusové lingvistiky a počítačové lexikografie

Pavel Rychlý

pary@fi.muni.cz
Centrum zpracování přirozeného jazyka

23. února 2024

Technické informace

- http://www.fi.muni.cz/~pary/ib047/

Technické informace

- `http://www.fi.muni.cz/~pary/ib047/`
- zkouška
  - písemná
  - možnost získat extra body za praktický úkol
- praktické úkoly
  - morfologické značkování textu
  - konfigurace tokenizátoru
  - detekce problémů v textech pro strojový překlad
  - významy slov dle kolokace

Korpusová lingvistika

Předmět přednášky

- lingvistika
  - věda, která se zabývá přirozenými jazyky
- korpus
  - velký soubor textů
  - většinou v elektronické podobě

Korpusová lingvistika

Předmět přednášky

- lingvistika
  - věda, která se zabývá přirozenými jazyky
- korpus
  - velký soubor textů
  - většinou v elektronické podobě
- zkoumá jazyky na základě korpusů

Počítačová lexikografie

Předmět přednášky

- lexikografie
  - věda, která se zabývá slovníky
  - zejména budováním slovníků

Počítačová lexikografie

Předmět přednášky

- lexikografie
  - věda, která se zabývá slovníky
  - zejména budováním slovníků
- budování slovníků na základě korpusů
- jedno z nějčastějších použití korpusů

Obsah přednášky (1)

- Úvod, motivace, historie
- Typy a formáty korpusů, standardy
- Nástroje na tvorbu a údržbu korpusů
- Značkování, metastruktura
- Gramatické značkování
- Syntaktické značkování
- Paralelní korpusy

Obsah přednášky (2)

- Automatické značkování, desambiguace
- Statistické zpracování korpusů
- Využití korpusů, uživatelská rozhraní
- Typy a formáty slovníků
- Struktura hesla, popis významů
- Využití korpusů pro tvorbu slovníků
- Lexikografické stanice

Motivace

- Popis přirozeného jazyka
  - slovník
  - gramatika
- Zkoumání jazyka
  - tradičně pomocí introspekce/intuice
  - podpořené výpisky (citáty) (autorit)
  - často subjektivní

Motivace

- Tradiční přírodní vědy
  - hypotézy ověřeny měřením
- Lingvistika
  - Jak provést měření?

Motivace

- Tradiční přírodní vědy
  - hypotézy ověřeny měřením
- Lingvistika
  - Jak provést měření?
  - objektivní zkoumání reálných užití jazyka
  - korpus

Má lingvista dělat měření?

- Pravidla pravopisu
  - závazná norma

Má lingvista dělat měření?

- Pravidla pravopisu
  - závazná norma
- Studium cizího jazyka
  - je výhodnější učit se reálný jazyk, jak lidé mluví a píší

Má lingvista dělat měření?

- Pravidla pravopisu
  - závazná norma
- Studium cizího jazyka
  - je výhodnější učit se reálný jazyk, jak lidé mluví a píší
- Zpracování přirozeného jazyka
  - potřebujeme robustní aplikace

Co to je korpus?

- Co to je text, dokument?
  - leccos
- Různé typy korpusů
  - textové
  - mluvené

Co to je korpus?

- Co to je text, dokument?
  - leccos
- Různé typy korpusů
  - textové
  - mluvené
- Pro potřeby lingvistiky
  - textový korpus

Textový korpus

- soubor textů
- charakteristiky
  - rozsáhlý
  - v jednotném formátu
  - stukturovaný
  - v elektronické podobě

Co znamená rozsáhlý?

Co znamená *rozsáhlý*?

- první koprusy: 1 milion slov
  - příliš malé pro zajímavější výsledky
  - dostačující pro globální statistiky
  - délka věty/slova, nejčastější slova
- nyní běžně stovky milionů slov
  - průměrná rychlost čtení je 125–225 slov za minutu
  - $200 * 60 * 18 = 216000$ slov za den (18 hodin)
  - $\rightsquigarrow$ 79 milionů za rok (365 dní)
  - dost velká slovní zásoba
- dostupné jsou i giga-korpusy
  - více než miliarda slov
  - zhruba 50 let čtení při 4 hodinách denně
  - málokdo dokáže přečíst více

Typy korpusů

- vždy záleží na účelu a způsobu použití
- možnosti
  - jazyk
  - typy textů
  - zdroj dat
  - značkování
  - ...

První korpus

Brown

- americká angličtina (1961)
- Brown University, 1964
- gramatické značkování, 1979
- 500 textů, 1 mil. slov
- W. N. Francis & H. Kučera
  - první statistické charakteristiky angličtiny
  - relativní četnosti slov a slovních druhů

SUSANNE

SUSANNE

- Geoffrey Sampson
- English for the Computer
- část korpusu Brown
- nové gramatické značkování
- syntaktické značkování

BNC

British National Corpus

- britská angličtina, 10 % mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků + univerzity
- 1991–1994, World Edition 2000
- $\approx$3000 textů, 100 mil. slov
- gramatické značkování automatickým nástrojem

BoE

Bank of English

- britská angličtina
- COBUILD (HarperCollins), University of Birmingham
- 1991, stále rozšiřován
- 2005, $\approx$525 mil. slov

Další národní korpusy

Český národní korpus

- ÚČNK, FF UK
- SYN2000: 100 mil. slov (60% noviny)
- SYN2005: 100 mil. slov (40% beletrie)
- SYN2010, SYN2015, SYN2020: 100 mil. slov
- SYN2006PUB: 300 mil. slov
- SYN2009PUB (700), SYN2013PUB (940)
- dohromady SYN: 4,7 mld. slov (verze 9, 2010-2021)
- Litera, Synek, BMK, KSK, ...
- InterCorp – paralelní koprusy, více než 40 jazyků

Slovenský, Maďarský, Chorvatský, ...
Americký

Korpusy na FI

vytvořené na FI

Desam ručně značkovaný (desambiguovaný)
$\approx$ 1 mil. slov

WWW periodika z webu, z let 1996–1998
$\approx$ 100 mil.

I047 vytvářený studenty Úvodu do korp. ling.
$\approx$ 45 mil.

Chyby práce studentů předmětu Základy odb. stylu s vyznačenými chybami
$\approx$ 400 tis.

BiWeC obrovský korpus z webu, zatím angličtina
$\approx$ 3–9 miliard slov

enTenTen,deTenTen korpusy o velikostech 10^{10} slov

Korpusy na FI

spolupráce

- itWac, ukWac, deWac, ...
- Dopisy
- Mluv
- Kačenka
- ČNPK
- 1984
- Otto
- Italian
- Giga Chinese
- RapCor