

Velké textové korpusy v praxi

Karel Pala, Pavel Rychlý

Centrum zpracování přirozeného jazyka
Fakulta informatiky
Masarykova Univerzita

14. října 2007

Obsah

- 1 Co jsou korpusy?
- 2 Formáty korpusů
- 3 Značkování korpusů
- 4 Korpusové manažery
- 5 Pokročilé statistické zpracování kontextů

Obsah

- 1 Co jsou korpusy?
- 2 Formáty korpusů
- 3 Značkování korpusů
- 4 Korpusové manažery
- 5 Pokročilé statistické zpracování kontextů

Co jsou korpusy?

- velké ucelené soubory textů (psaných i mluvených) daného jazyka
- přirozená podoba - opakované použití
- reprezentativnost a vyváženost (účel)
- rozsah: desítky a stovky miliónů (dnes miliardy) slovních tvarů
- čím více, tím lépe (velká čísla)
- vše se zpracovává na počítačích
- pro potřeby NLP
 - **textový korpus**

Typy a příklady korpusů

- psané (textové) korpusy - obecné, specializované
- BNC - cca 100 mil. tokenů,
- SYN2000 - cca 100 mil. tokenů, SYN2005,
- DESAM - manuálně značkováný korpus FI MU, SNK, další národní korpusy: maďarský, chorvatský, polský, ruský a dalších cca 30 jazyků
- mluvené (řečové) korpusy - součást BNC (10)
- paralelní - vícejazyčné - zarovnané texty ve dvou a více jazycích
- multimediální (audio-video) korpusy

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

BNC

British National Corpus

- klasika mezi textovými korpusy
- britská angličtina, 10% mluva
- první velký korpus pro lexikografy
- vydavatelé slovníků (OUP) + univerzity
- 1991–1994, World Edition 2000
- ≈3000 textů, 100 mil. slov
- gramatické značkování automatickým nástrojem

Nástroje pro práci s korpusy

- nástroje pro zpracování textů
- konkordanční programy
- korpusové manažery - v jazyce C, Linuxu - X-Windows
- CQP, GCQP, Bonito/Manatee, SARA
- značkovače (taggery)
- desambiguátory (viz níže)

Motivace: proč potřebujeme korpusy?

- v lingvistice došlo ke změně paradigmatu ve zkoumání přirozeného jazyka
- introspektivní pohled na jazyk (Chomsky a generativní gramatiky)
- nyní empirický pohled - velké soubory jazykových dat
- interdisciplinarita - korpusy jsou užitečné
 - pro sociology, sociolingvisty a psychology
 - lexikografy a lingvisty, překladatele (strojový překlad)
 - pro učitele a studenty cizích jazyků (autentické texty)
 - výzkumné pracovníky v oblasti AI a NLP

Obsah

- 1 Co jsou korpusy?
- 2 Formáty korpusů
- 3 Značkování korpusů
- 4 Korpusové manažery
- 5 Pokročilé statistické zpracování kontextů

Formáty korpusů

- archiv/kolekce
 - různé formáty, podle zdroje/typu
- textové banky
 - jednotný formát a základní struktura
 - dokumenty/texty, základní metainformace
- vertikální text
- binární data v aplikaci
 - pomocné data pro rychlejší zpracování
 - indexy
 - statistiky

Kódování znaků

- 8 bitů \approx 256 znaků
 - ASCII – základ 7 bitů
 - kódování pro češtinu
 - ISO-Latin-2, Windows-1250, 852
- Unicode
 - 32bitů na znak
 - UTF-8
 - 1 až 4 byty na znak
 - UTF-16
 - 2 až 4 byty na znak

Kódování metainformací

- escape-sekvence
 - speciální znak mění význam následujících znaků
 - \n, \t, &#x26; , <tag>
- SGML
 - Standard Generalised Markup Language
 - ISO 8879:1986(E)
- XML
 - Extensible Markup Language
 - W3C, 1998

XML

- struktura popsána v DTD
- elementy
 - počáteční, koncová značka
 - <doc>, <head>, </head>, <g/>
- atributy elementů/značek
 - <doc title="Jak pejsek ..." author="Čapek">
 - <head type="main">
- entity
 - >, <, &, ´

Standardy pro ukládání

- SGML/XML
- TEI
 - Text Encoding Initiative
 - TEI Guidelines for Electronic Text Encoding and Interchange
- CES, XCES
 - Corpus Encoding Standard

Obsah korpusu

Co je v korpusu uloženo?

- text
- metainformace
- struktura dokumentu
 - odstavce, nadpisy, verše, věty
- značkování
 - informace o slovech
 - morfologie, základní tvary

Tokenizace

Rozdělení textu do pozic

- token (pozice) = základní prvek korpusu
- většinou slovo, číslo, interpunkce
 - bude-li, don't
- může silně ovlivnit výsledky

Vertikální text

- jednoduchý formát i jeho zpracování
 - každý token na samostatném řádku
 - struktury formou XML elementů
 - značkování odděleno tabulátorem
- podrobnosti
 - <http://www.fi.muni.cz/nlp/>
 - Informace pro současné a potenciální spolupracovníky
 - Textové korpusy
 - Popis vertikálů

Zpracování textů na UNIXu

- coreutils
 - cat, head, tail, wc, sort, uniq, comm
 - cut, paste join, tr
- grep
- awk
- sed / perl

Příklady použití coreutils

- slovník z vertikálního textu

Example

```
cut -f 1 -s desam.vert |sort |uniq -c \  
|sort -rn >desam.dict
```

- jednoduchá tokenizace

Example

```
tr -cs 'a-zA-Z0-9' '\n' <GPL >GPL.vert  
cat GPL.vert |sort |uniq -c |sort -rn >GPL.dict
```

Obsah

- 1 Co jsou korpusy?
- 2 Formáty korpusů
- 3 Značkování korpusů
- 4 Korpusové manažery
- 5 Pokročilé statistické zpracování kontextů

Značkování (anotování) korpusů

- vkládání (meta-)informace o jazyce do korpusů
- strukturní značkování
- informace o struktuře textu
- označení nadpisů, odstavců, event. Vět
- uvedení zdrojů textu, autorů, času vzniku aj.
- viz níže příklad pro BNC

Morfologické značkování

- určení základních tvarů - lemmat
- přiřazování značek slovních druhů slovním tvarům
- určení gramatických kategorií nesených jednotlivými slovními tvary
- potřebujeme soubory značek - tagsety (pro češtinu cca 1600)
- příklady značek: brněnské (atributový princip), pražské (poziční)
- vertikální formát - ukázka (samostatný obrázek)
 - 'korpus' k1gInSc1, k1gInSc4,
 - 'je' k5eAaImIp3nS
 - 'je' k3p3gMnPc4, k3p3gInPc4, k3p3gNnSc4, k3p3gNnPc4, k3p3gFnPc4
 - 'nový' k2eAgMnSc1d1 plus dalších 22 značek
- víceznačnost slovnědruhová, v gramatických kategoriích, u lemmat

Příklad vertikálu

```
Václav <l>Václav <c>k1gMnSc1
Havel <l>Havel <c>k1gMnSc1
přišel <l>přijít <c>k5eApMnStMmPaP,k5eApInStMmPaP

naopak <l>naopak <c>k6xMeA
s <l>s <c>k7c7
vlastním <l>vlastní <c>k2eAgMnSc67d1,k2eAgXnPc3d1,k2eAgU
<c>k5eApInStPmIaI
volebním <l>volební <c>k2eAgMnSc67d1,k2eAgXnPc3d1,k2eAgU
programem <l>program <c>k1gInSc7 ,
který <l>který <c>k3xQgMnSc15,k3xQgInSc145
nikomu <l>nikdo <c>k3xNnSc3
neubližuje <l>ubližovat <c>k5eNpMnStPmTaI,k5eNp3nStPmIaI
.
```

Syntaktické značkování

- přiřazování syntaktických reprezentací (stromových struktur) větám v korpusu
- PDT v.1 a 2 (manuálně, věty z ČNK)
- závislostní stromové struktury
- složkové stromy (Penn Treebank, PropBank)

Sémantické značkování

- přiřazování významů slovním tvarům v korpusu (podle slovníku)
- vztah k Word Sense Disambiguation (WSD)
- určení anaforické a referenční struktury textu
- sémantické reprezentace - tektogramatické struktury přiřazují se manuálně, viz např. PDT v.2)

Nástroje pro korpusy

- morfologické analyzátoři, značkovače (taggery)
- ajka – ukázka: příklady
- desambiguace – desambiguátory: pravidlové, statistické, hybridní
- hodnocení pomocí parametrů: pokrytí, přesnost
- viz např. pro SYN2000 – hybridní desambiguátor MORČE: 96 %

Budování korpusů - textových a mluvených

- konverze z médií a sazebních souborů
- využití technik OCR
- klasické manuální techniky - přepisování
- získávání textů z webu (BootCat, Corpus Builder)
- problémy s autorskými právy
- standardizace korpusů - iniciativa EAGLE

Reprezentativnost a vyváženost - na příkladu BNC

- přírodní vědy a čistá věda 5%
- aplikované vědy 5%
- sociální vědy 15%
- politická publicistika 15%
- publicistika obchodní a finanční 10%
- publicistika umělecká (rock & pop, divadlo,...) .. 10%
- publicistika náboženská a filosofická 5%
- publicistika zábavná (sport, zahrádkáři, ...) 15%

Klasifikační rysy použité v BNC

- identifikátor vzorku
- rozsah vzorku (počet slov), začátek a konec vzorku
- rozsah textu příslušného typu (počet slov)
- kompozice textu (hladký, složený, sbírka)
- standardní bibliografický odkaz
- datum vzniku
- předmětná oblast
- autorství (individuální, společné, institucionální, neznámé)
- pohlaví autora
- věková skupina autora
- etnick skupina autora

Vnitřní struktura korpusu

- atributy poziční
- atributy strukturní (hranice vět, odstavců)

slovo	lemma	gr.značky	sem.značky
ženu	hnát/žena	k5/k1gFnSc1	HUM+FEM/POHYB
ovce	ovce	k1gFnPc4	ANIM
na	na	k7c4	DIRECT
pastvu	pastva	k1gFnSc4	LOC

Obsah

- 1 Co jsou korpusy?
- 2 Formáty korpusů
- 3 Značkování korpusů
- 4 Korpusové manažery
- 5 Pokročilé statistické zpracování kontextů

Korpusové manažery

nástroje na zpracování korpusů

- uložení textu
- editace/příprava textu
- značkování
- rozdělení do pozic (tokenizace)
- vyhledávání (konkordance)
- statistiky

Systém Manatee

- korpusový manažer
- přímo podporuje
 - uložení textu
 - vyhledávání (konkordance)
 - statistiky
- externí nástroje
 - značkování
 - rozdělení do pozic

Systém Manatee

hlavní zaměření

- velké korpusy
- rozsáhlé značkování
 - morfologické, syntaktické, metainformace
- návaznost na další aplikace/nástroje
 - korpusový editor, tvorba slovníků
- univerzálnost
 - různé jazyky, kódování, systémy značek

Klíčové vlastnosti

- modulární systém
- přístup z různých rozhraní
 - grafické uživatelské rozhraní (Bonito)
 - aplikační programové rozhraní (API)
 - příkazový řádek
- rozsáhlá data
 - až 2 mld. pozic
 - neomezeně atributů a metainformací
- rychlost
 - vyhledávání, statistiky

Klíčové vlastnosti

- multihodnoty
 - zpracování víceznačných značkování
- dynamické atributy
 - vyhledávání a statistiky na počítaných datech
- subkorpusy
- silný dotazovací jazyk
 - dotazy na všechny atributy, metainformace
 - pozitivní/negativní filtry

Klíčové vlastnosti

- frekvenční distribuce
 - víceúrovňová
 - všechny atributy a metainformace
- kolokace
 - různé statistické funkce

Dotazovací jazyk

úroveň jenoho tokenu

- regulární výrazy nad znaky
- logické kombinace (and, or, not)

úroveň posloupnosti tokenů

- regulární výrazy nad tokeny
- posloupnost
- opakování
- alternativa
- omezení na struktury

Příklady dotazů

```
[word="dream.*"]  
[word="[dD]ream"]  
[word="[0-9]*"] [lc="dreams"]  
[tag="NN."] [lempos="dream-v"]  
[word="[0-9]5,"] [word="\."]   
[tag="DPS"] [] [lemma="dream"]  
[tag="DPS"] [tag="AJ0"]? [lemma="dream"]  
[tag="AJ0"]2 [lemma="dream"]  
[word="the"] []0,3 [lempos="dream-n"]  
[lemma="dream"] within <bnccdoc id="A0.">
```

Obsah

- 1 Co jsou korpusy?
- 2 Formáty korpusů
- 3 Značkování korpusů
- 4 Korpusové manažery
- 5 Pokročilé statistické zpracování kontextů

Co to je kontext?

Kontext jsou slova v okolí klíčového slova.

- Jaké okolí?:
 - následující slovo
 - předcházející slovo
 - okno, +1 až +5
 - okno, -5 až -1
- Ne všechna slova v okolí jsou důležitá.
- Jak určíme důležitost?
 - nejčastější kolokace – ale to je “the”
 - (statisticky) nejvýznamnější – jaký vzorec?

Word Sketch

Jednostránkový souhrn chování slova

Jak jej lze vytvořit?

- Velký vyvážený korpus
- Vyhledáme závislé prvky (subjects, objects, heads, modifiers, ...)
- Seznam kolokací pro každou gramatickou relaci
- Statistika pro třídění každého seznamu

The Sketch Engine

- Vstup:
 - libovolný korpus, libovolný jazyk
 - lemmatizovaný, značkováný
 - specifikace gramatických relací
- Výstup:
 - Word sketches
 - Thesaurus
 - Dotazovací systém

Koeficient výnačnosti

- počty výskytů ($word_1, gramrel, word_2$)
- $AScore(w_1, R, w_2) = \log \frac{\|w_1, R, w_2\| \cdot \|*, *, *\|}{\|w_1, R, *\| \cdot \|*, *, w_2\|} \cdot \log(\|w_1, R, w_2\| + 1)$

Koeficient podobnosti

- porovnání profilů slov w_1 a w_2
- pouze důležité (význačné) kontexty
- jaký je překryv
- počty ($word_1, (gramrel, word_i)$) a ($word_2, (gramrel, word_i)$)

$$Sim(w_1, w_2) = \frac{\sum_{(tup_i, tup_j) \in \{tup_{w_1} \cap tup_{w_2}\}} AS_i + AS_j - (AS_i - AS_j)^2 / 50}{\sum_{tup_i \in \{tup_{w_1} \cup tup_{w_2}\}} AS_i}$$

Velikosti dat

Velikosti korpusů, jejich slovníků a počty slov v kontextech

Korpus	Velikost	Slov	Lemat	Různé k.	Všechny k.
BNC	111m	776k	722k	23m	63m
SYN2000	114m	1,65m	776k	19m	58m
OEC	1,12g	3,67m	3,12m	84m	569m
Itwac	1,92g	6,32m	4,76m	67m	587m

Velikosti slovníků i počty různých kontextů rostou sublineárně s velikostí korpusu.

Velikost matice

- Podobnost všech dvojic lemat
- Matice velikosti N^2 , kde N je 700k – 5m
- Počet prvků v řádech tera (10^{12})
- Matice je naštěstí velice řídká
- Většina hodnot je 0 nebo “skoro” 0
- Dokonce většina celých řádků/sloupců je prázdných

Praktické velikosti dat

- Výpočet pouze pro slova s minimální četností
- Lépe limitovat počty kontextů než prostých výskytů
- Z kontextů brát pouze statisticky významné

Korpus	MIN	Lemmat	KWIC	CTX
BNC	1	152k	5.7m	608k
BNC	20	68k	5.6m	588k
OEC	2	269k	27.5m	994k
OEC	20	128k	27.3m	981k
OEC	200	48k	26.7m	965k
Itwac	20	137k	24.8m	1.1m

Praktické velikosti dat

- Matice velikosti N^2 , kde N je 50k – 200k
- Počet prvků v řádech giga (10^{10})
- Hodnota každého prvku vznikne aplikací funkce podobnosti na vektory délky $K = 500k - 1m$.
- Přímočarý algoritmus pro výpočet celé matice má časovou složitost $O(N^2K)$.
- Složitost je polynomiální, ale algoritmus je prakticky nepoužitelný pro dané rozsahy hodnot.
- Odhadované doby výpočtu jsou v měsících až letech.
- Heuristiky snižují velikosti N a K na úkor přesnosti výsledných hodnot.
- Doba výpočtu je potom v řádech dnů s chybou 1–4%.

Efektivní algoritmus

- I menší matice je velice řídká
- Není potřeba počítat podobnost pro slova, která nemají nic společného,
- tedy nemají žádný společný kontext.
- Hlavní cyklus algoritmu tedy nevedeme přes slova, ale přes kontexty.

Efektivní algoritmus

- Vstup: seznam všech možných slov v kontextech, $\langle w, r, w' \rangle$, s četnostmi výskytů v korpusu
- Výstup: matice podobností slov $sim(w_1, w_2)$

```
for  $\langle r, w' \rangle$  in CONTEXTS:  
  WLIST = set of all  $w$  where  $\langle w, r, w' \rangle$  exists  
  for  $w_1$  in WLIST:  
    for  $w_2$  in WLIST:  
       $sim(w_1, w_2) + = f(\text{frequencies})$ 
```

Optimalizace

- Pokud $|WLIST| > 10000$, daný kontext přeskočíme.
- Matici $sim(w_1, w_2)$ během výpočtu nedržíme celou v paměti.
- Opakovaný běh hlavního cyklu pro omezený rozsah w_1 .
- Místo $sim(w_1, w_2) + x$ generujeme na výstup $\langle w_1, w_2, x \rangle$.
- Výstupní seznam potom setřídíme a sčítáme jednotlivé x .
- Využití TMMS (Two Phase Multi-way Merge Sort) s průběžným sčítáním.
- Místo několika stovek GB třídíme jednotky GB.

Výsledky

- Algoritmus je řádově rychlejší než přímočarý algoritmus. (18 dnů × 2 hodiny)

Korpus	MIN	Lemmat	KWIC	CTX	čas
BNC	1	152k	5.7m	608k	13m 9s
BNC	20	68k	5.6m	588k	9m 30s
OEC	2	269k	27.5m	994k	1h 40m
OEC	20	128k	27.3m	981k	1h 27m
OEC	200	48k	26.7m	965k	1h 10m
Itwac	20	137k	24.8m	1.1m	1h 16m

- Bez omezení přesnosti.
- Možnost snadné paralelizace.