Word Level Analysis

Pavel Šmerk

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

Workshop of Natural Language Processing Centre
May 28, 2013

Motivation

chladnička s mrazákem

Přibližný počet výsledků: 1 940 000 (0,16 s)

Volně stojící **chladničky s mrazákem** dole | MALL.CZ

Nabízíme zajímavou nabídku volně stojících chladniček s **mrazákem** dole značek AEG, Amica, baumatic, Fagor, LG a další. Vše pohodlně do 24 hodin u vás.

DATART | Ledničky

Chladničky kombinované s mrazákem nahoře. Americké chladničky. Auto, přenosné **...** NO FROST **chladnička s mrazákem**. Doprava Datart. Přidat do porovnání **...**

Lednice A++ - Heureka.cz

A++, kombinované, volně stojící, 230 l , 180 cm , **Mrazák**. Úsporná kombinovaná **chladnička** Gorenje RK 61820 W je řazena do energetické třídy A++. Police ve **...**

Bílé zboží > **Chladničky** > **Mrazák** dole - Srovnání cen - Zalevno.cz

Porovnejte si ceny zboží z kategorie Bílé zboží > **Chladničky** > **Mrazák** dole - najděte nejlevnější obchod.

Motivation

- Many applications need a tool for "clustering" of word forms appearing in texts:
  - *chladniček*
  - *chladničky*
  - *chladničkách* $\iff$ *chladnička*
  - *chladničce*
  - *...*
- Indexing, searching, keyword extraction, ...
- And almost all NLP tools

Word Level Processing Data for Czech

- For almost 12 M word forms (incl. colloquial forms):
  - lemma (canonical form, dictionary form)
  - grammatical information: part of speech, number, case etc.
- Word form *stroj* has 3 interpretations:
  - lemma *stroj*, nominative
  - lemma *stroj*, accusative
    - noun, masculine animated, singular
  - lemma *strojit*,
    - verb, 2^{nd} person, singular, imperative mood

Possible Applications

- Various types of analyses:
  - word form $\Rightarrow$ lemma (many types of searching/indexation)
    - *nebral* $\Rightarrow$ *brát/nebrat (úplatky)*
    - *nejstaršího* $\Rightarrow$ *nejstarší/starý (člověk)*
    - *chladnička* $\Rightarrow$ *chladničky* (as a class)
    - *bavlna* $\Leftrightarrow$ *bavlněný* (word derivation)
  - word form/lemma + gram. info. $\Rightarrow$ word form
    - e.g. salutation generation: *pane Procházko*
  - word form/lemma $\Rightarrow$ all word forms
  - word form $\Rightarrow$ lemma + full/partial grammatical information
- The analysis is very fast
  - approx. 1 million word forms per second

Processing Unknown Words

- Some word forms in processed texts are unknown:
  - terms *polydaktylie*, neologisms *klausoviny*, typos *bizardního*, colloquial words *plaťáky*, etc.
- An ending of the word form is able to determine e.g.
  - lemma: *klausoviny* $\Rightarrow$ *klausovina*
  - grammatical information: *bizardního* $\Rightarrow$ genitive, etc.
  - derivational relations: *plaťáky* $\Rightarrow$ *plaťákový*
- Texts from a particular domain allows grouping of unknown word forms:
  - *polydaktylie, polydaktiliích, polydaktylií, . . .* $\Leftrightarrow$ *polydaktylie*
  - $\Rightarrow$ extension of data or more precise "guessing"

Resolving Ambiguities Using Context

- An extreme case *Stroj ženu holí.*
  - *Já stroj ženu holí, ty stroj ženu holí, ten stroj ženu holí.*
- Usual case is e.g. *stát*
  - noun: *Stát jsem já.*
  - verb: *Celá továrna musela hodinu stát.*
  - at the part of speech level, it is a bigger problem for English
- The context of the word determines its interpretation
  - rules and/or statistical data describe typical contexts of nouns, verbs, etc.
  - using such information one can tell that *stát* is noun/verb

Example of Contexts — Word Sketches

stát podstatné jméno

a_modifier	938517	-0.8
spojený	223381	12.28
členský	137993	11.83
americký	29942	9.01
demokratický	12202	8.46

gen_2	274456	-0.7
hlava	20922	8.7
zastupování	2716	8.24
složka	5263	7.9
majetek	5793	7.85

post_verb	143087	-0.8
dotovat	433	6.3
moct	15773	5.93
hodlat	528	5.87
dlužit	342	5.87

stát sloveso

has_subj	942837	-3.7
zázrak	4433	7.12
nehoda	4438	6.87
socha	3587	6.72
kostel	3714	6.39

post_v	184481	-1.5
čelo	11624	9.36
pozadí	2507	7.83
fronta	2654	7.72
přepočet	1098	7.35

is_subj_of	127156	-0.5
zavázat	469	6.58
hospodařit	517	6.56
zůstat	3245	6.5
přispívat	1021	6.46

Spellchecking and Diacritics Restoration

- Data also allow spellchecking and diacritics restoration:

Universality

- All the mentioned processes can be
  - tuned for a specific domain
    - using texts from this domain
  - applied to a language other than Czech
    - (Slovak, Polish, German, English, ...)

Latest Applications

- Seznam.cz, Yandex.ru, Aukro.cz, Václav Havel Library
  - indexing and searching
- Information System of Masaryk University
  - other universities and schools (FHS UK, JAMU, VŠFS, . . .)
  - affiliate projects (theses.cz, odevzdej.cz, repozitar.cz)
  - indexing, searching and plagiarism detection
- "Internetová jazyková příručka"
  - online source on Czech orthography and grammar
  - NLP Centre data were a starting point for word form tables

Conclusions

- Word level processing of texts allows:
  - various types of base word determining
    which forms are to be grouped together
  - ambiguity resolution according to the context
  - word form generation
  - spellchecking, diacritics restoration
- The tools/data can be domain specific and for various languages