Korpusová lingvistika a počítačové zpracování přirozeného jazyka

Vladimír Petkevič & Alexandr Rosen

Ústav teoretické a komputační lingvistiky
Filozofické fakulty Univerzity Karlovy v Praze

Korpusový seminář
5. května 2016

Osnova

Osnova

(Počítačové zpracování přirozeného jazyka, neboli:

- natural language processing (NLP)

vs.

- počítačová lingvistika (computational linguistics)

Na pomezí:

- lingvistiky
- informatiky
- umělé inteligence
- akustiky, ...

Zkoumá:

- komunikaci mezi lidmi a počítači v přirozeném jazyce
- typicky: analýzu a syntézu textů nebo řeči

Co NLP zvládne

- NLP jako to dělá člověk: úkol, který je *AI-complete*
- Stejně těžké jako naučit počítat zvládnout to, co člověk (*strong AI*
- Počítači něco chybí, co to je?

Korpusy a NLP

- NLP jako (spolu)tvůrce
- NLP jako uživatel
- NLP jako pomocník uživatele

- NLP: i jednoduché metody a nástroje pro zpracování textu na počítači

Osnova

Historie

- 1950: Turingův test
- 1954: Georgetownský experiment
- 1966: Zpráva ALPAC (Automatic Language Processing Advisory Committee)
- 1966: ELIZA
- 1970: SHRDLU (Terry Winograd)
- 1990– strojové učení, statistické modely

Osnova

Metody NLP

- modelování přirozeného jazyka
- pravidly (rule-based)
- statisticky (stochastic)
- hybridně

Osnova

Čím se NLP zabývá 1/3

- strojový překlad (machine translation)
- porozumění přirozenému jazyku (NL understanding)
- syntéza/generování přirozeného jazyka (NL generation)
- rozpoznávání řeči (speech recognition)
- syntéza řeči (text-to-speech), včetně znakového jazyka
- tokenizace, segmentace textu i řeči
- dialogové systémy, telematika, počítač pro smyslově postižené
- morfologická analýza
- morfologická desambiguace (part-of-speech tagging)
- syntaktická analýza (parsing)
- analýza diskursu (discourse analysis)
- analýza textů (text analytics, big data)

Čím se NLP zabývá 2/3

- určování koreferencí (coreference resolution)
- lexikální sémantická disambiguace (word sense disambiguation)
- postojová analýza (sentiment analysis)
- rozpoznávání pojmenovaných entit (named entity recognition)
- identifikace jazyka (language identification)
- optické rozpoznávání znaků (optical character recognition)
- rozpoznávání spamu
- detekce plagiátů, kyberšikany a trollingu
- automatická administrace internetových diskusí
- forenzní lingvistika, stylometrie
- výukový software

Čím se NLP zabývá 3/3

- korektury pravopisu a gramatiky (spell and grammar checking)
- zodpovídání dotazů (question answering)
- rozvíjení dotazů (question expansion)
- excerpce informací, naplňování báze znalostí z textu (information retrieval, extraction)
- indexace textu, řeči, videa (http://ufal.mff.cuni.cz/cvhm/about-us.html)
- automatická sumarizace (automatic summarization)
- zjednodušování textu (text simplification)
- segmentace podle tématu, určování tématu (topic segmentation and recognition)

Osnova

Vnitřní vs. vnější

- etalon (gold standard)

vs.

- srovnatelné výsledky jiného nástroje

Černá vs. skleněná skříňka

- výsledky, rychlost, nároky

vs.

- design, metody, algoritmy, lingvistické zdroje

Automaticky vs. ručně

- etalon s dostatečnou mezianotátorskou shodou (IAA)

vs.

- posuzovatelé

Osnova

Odkazy

- http://en.wikipedia.org/wiki/Natural_language_processing
- http://en.wikipedia.org/wiki/Outline_of_natural_language_processing
- http://en.wikipedia.org/wiki/Outline_of_natural_language_processing
- http://www.systems.ethz.ch/node/354
- http://ufal.mff.cuni.cz/course/popj1
- http://blog.algorithmia.com/2015/09/getting-started-with-natural-language-processing/
- http://www.nltk.org
- http://www.geneea.com
- http://jasnopis.pl