Centrum zpracování přirozeného jazyka

Fakulta informatiky, Masarykova univerzita

Pavel Rychlý

26. dubna 2012

Co znamená ***zpracovávat*** přirozený jazyk?

- snažit se popsat/formalizovat přirozený jazyk tak, aby s ním bylo možné (alespoň do určité míry) strojově manipulovat, například:
  - analyzovat strukturu jazyka
  - překládat texty mezi různými jazyky
  - zachytit význam textů
  - vylepšit různé aplikace znalostmi o jazyku (např. information retrieval, rozhraní k databázím/informačním systémům)
- hlavní motivací je plnohodnotná (obousměrná) komunikace s počítačem v přirozeném jazyce ($\rightsquigarrow$ umělá inteligence)

Jak se zpracovává přirozený jazyk?

- **korpus** – rozsáhlý soubor textů
  - umožňuje nám nahlédnout, jak se jazyk používá
  - lze na něm zkoumat různé pravidelnosti/zákonitosti (která slova se často používají spolu s jinými slovy atp.)
- **statistické metody a strojové učení**
  - máme-li dostatečně rozsáhlá ručně zpracovaná data, je možné v nich obsaženou “znalost”/informaci použít pomocí metod statistiky a strojového učení na zpracování nových dat
- **jazykové zdroje**
  - slovníky všeho druhu
  - ručně či automaticky vytvářené
- a mnoho dalšího ...

Czaccent

- program pro doplňování diakritiky do textů psaných bez háčků a čárek
- Příklady:
  mesicu $\rightsquigarrow$ *měsíců*
  zadany $\rightsquigarrow$ *zadaný*
  $\rightsquigarrow$ *žádaný*
- Doplňuje diakritiku s úspěšností cca. 97%.
- přístupné přes webové rozhraní:
  `http://nlp.fi.muni.cz/cz_accent/`
- tisíce uživatelů

Ajka

- morfologický analyzátor Ajka
- pro slovní tvar určí základní tvar a hodnoty gramatických kategorií
  - ženu $\rightsquigarrow$ hnát, k5eAaImIp1nS
  - ženu $\rightsquigarrow$ žena, k1gFnSc4
- další varianty – všechny tvary k základnímu, možné oháčkování, . . .
- software je jazykově nezávislý
- nová implementace Majka – mnohonásobně rychlejší (6-10x, 5000x)
- využívá např. Seznam.cz, Yandex, IS MU

Platforma DEB II

- DEBDict
  - prohlížeč elektronických slovníků
  - umožňuje pracovat současně s libovolným počtem elektronických slovníků uložených ve formátu XML
  - ve slovnících lze hledat pomocí komplexních dotazů a výsledky různými způsoby třídit
- WordNet – DebVisDic – sémantická síť
- Jazyková příručka
  - projekt společně s ÚJČ AVČR
  - 20 tisíc přístupů každý den
- desítka dalších systémů/slovníků

Velké korpusy

- ve spolupráci s Lexical Computing Ltd.
- vytváření rozsáhlých jazykových korpusů pro libovolný jazyk
- rozsah miliardy slov
- crawling, detekce jazyka, čištění, deduplikace
- SpiderLing, chared, JusText, onion
- Manatee, Bonito – vyhledávání, statistiky
- seznamy slov, n-gramy, kolokace, thesaurus
- Seznam.cz, ÚČNK, OUP, CUP, …

Plánované projekty

- Archiv českého webu + Search API
  - stahování a indexování českého webu
  - vyhledávání podle různých kritérií
- Strojový překlad
  - blízké jazyky (čeština–slovenština)
  - překlad v omezené doméně (např. popisy produktů)

Uvítáme jakoukoliv formu spolupráce.

Další informace

- web: http://nlp.fi.muni.cz/
- projekty NLP Centra většinou dostupné na http://nlp.fi.muni.cz/projekty/
- projekty s větší aktivitou: http://nlp.fi.muni.cz/trac/

pary@fi.muni.cz