Nástroje pro automatickou anotaci

Jiří Materna

9. května 2007

Připomenutí PDT 2.0

- český, ručně anotovaný korpus, celkem 2 miliony slov
- anotace na 3 úrovních:
  - Morfologická
  - Analytická
  - Tektogramatická
- automatická anotace na morfologické a analytické úrovni

Tokenizace a segmentace (1)

- **Tokenizace** – rozdělení vstupu na slova a interpunkci
- **Segmentace** – rozdělení tokenů do vět
- problémy s rozpoznáním konce věty:
  - řadové číslovky
  - zkratky
  - apod.
- řešeno seznamem výjimek

Tokenizace a segmentace (2)

- pokrytí 91.4%
- přesnost 98%
- F-míra 94.6%
- **Vstup** – plain text v ISO 8859-2
- **Výstup** – soubor ve formátu CSTS

Morfologická analýza a značkování (1)

- **morfologická analýza** – přiřazení možných lemmat a tagů slovům
- **morfologické značkování** – vybrání správného lemmatu a značky
- morfologická analýza založená na slovních kmenech (350.000 záznamů)
- analyzuje okolo 12 milionů českých slov

Morfologická analýza a značkování (2)

- pokrytí analyzátoru 97.5%
- desambiguace – bere se v úvahu krátký kontext, statistické metody
- přesnost desambiguace – 93.08%
- **Vstup** – soubor ve formátu CSTS
- **Výstup** – desambiguovaný CSTS

Parsing a přiřazení analytické funkce (1)

- **Parsing** –vytvoření závislostního stromu
- **přiřazení analytické funkce** – přiřazení sémantických hodnot jednotlivým uzlům
- parser upraven z parseru pro angličtinu Michaela Collinse
- věty delší než 60 slov neznačkovány
- přesnost parseru 82%

Parsing a přiřazení analytické funkce (2)

- přiřazení analytické funkce jako klasifikační problém
- vytvoření rozhodovacího stromu algoritmus C5
- naučená pravidla využita v perlovském programu pro přiřazení hodnot
- přesnost přiřazení 92%
- **Vstup** – CSTS výstup z morfolog. analyzátoru
- **Výstup** – označkovaný CSTS

- testování na 9MB souboru
- tokenizace – 17 sec
- morfologicka analyza 4 min 55 sec
- syntaktická analýza 1 min 6 sec