Morfo, Překlad

David Kolovratník

April 9, 2009

Témata

Zobecnění a reimplementace české morfologie

- sada programů pro práci s morfologickým slovníkem
- navazuje na studentský softwarový projekt Morfo
- podporováno tříletým GAUKem (běží druhý rok)

Výběr lexikálního ekvivalentu v automatickém překladu

- téma doktorského studia

Zobecnění a reimplementace české morfologie

Morfologie (pro účely projektu)

- práce s tvaroslovím – podchycení vztahu tvarů a lemmatu
- slovo – základní tvar = slovníkové heslo = lemma
- (ohebné) slovo se ohýbá – množina tvarů `žena` → `žena, ženy, ženě, ženu, ...`
- tvaru přísluší gramatické kategorie – podstatné jméno: rod, číslo, pád, (vzor, stylová platnost, ...)

Zobecnění a reimplementace české morfologie (2)

Řešené úlohy

- generování
  - k zadanému základnímu tvaru vypsat tvary ohnuté + gramatické kategorie
- analýza
  - k zadanému (ohnutému) tvaru vypsat možné základní tvary + gramatické kategorie
- vztahy jsou podchyceny ve **slovníku** vytvořeném na ÚFALu

Aplikace

- běžná součást v úlohách zpracování přirozeného jazyka
- vyhledávání informací, porozumění textu, spell & grammar checking

Výběr lexikálního ekvivalentu v automatickém překladu

automatický / strojový překlad

- z cizího (přirozeného) jazyka do vlastního
- různé přístupy, avšak kvůli velikosti / složitosti jazyka zpravidla založeny na datech a strojovém učení
- oblíbený přístup – statistický strojový překlad

Cíle

- kvalita výstupu „lze publikovat“ se zdá zatím nedosažitelná
- cíl do článků: zlepšovat hodnocení automatickou metrikou BLEU
- praktický cíl: snižovat práci nutnou k opravě výstupních textů

Výběr lexikálního ekvivalentu v automatickém překladu (2)

typické podproblémy

- sběr dvou/vícejazyčných paralelních textů
- čištění textů, sjednocení kódování, formátu, rozpoznání jazyka
- hledání párování odpovídajících si vět

- volba překladového modelu
  - konstrukce hypotézy – přeložené věty
  - ohodnocení (částečné) hypotézy
- hledání nejlepšího řešení vůči modelu – dekódování
- vypořádání se s řídkostí trénovacích dat
- vážení součástí překladového modelu

- jazykový model
- vyhodnocení, zpětná vazba

Konec