Počítače a přirozený jazyk
Statistický strojový překlad

Ondřej Bojar
bojar@ufal.mff.cuni.cz

1. březen 2006

Osnova

- Jak dopadl domácí úkol?
- Statistický strojový překlad (ngramový)
- Shrnutí potřebných součástí, témata k zápočtům, plán na semestr

Průlet statistickým překladem

viz výběry z prezentace ESSLLI SMT lecture 1, 2 a 3.

Souhrn: Co budeme potřebovat

- Paralelní texty (dodám, uvítám sběr dalších)
  Včetně zarovnání po větách (o tom jsem nemluvil, opět existuje ke stažení.)
- Zarovnání po slovech (zprovoznit GIZU)
- Spojování alignmentů a extrakce frází
- Jazykové modelování (SRI LM toolkit)
- Morfologie pro češtinu (Czech Free Morphology)
- Dekodér (překladač) (Pharaoh)
- Vyhodnocení (oficiální skript mt_eval)

Zdůvodnění "spěchu"

- Cílem semináře je dojít k provozuschopnému systému a "publikaci".
- Hodně času zabere:
  - Zprovoznění součástí
  - Kombinace součástí
  - Experimenty (vylepšování) na každé součásti – tj. podstata zápočtové úlohy
  - Sepsání výsledků, kritické zhodnocení metody a korektury publikace
- Uvedenou časovou potřebu nesmíme nechat na zkouškové období.
- Navíc by měl ještě v semestru být čas na diskusi o problémech, technických i principiálních slabinách metod.

Následující postup

- Příště podrobněji o tématech a dohoda nad rozdělením témat mezi skupinky.
- Přespříště diskuse nad problémy se zvolenými tématy:
  Vaše povídání (po skupinkách), čeho jste za ten týden chtěli dosáhnout (co zprovoznit), a jaké problémy jste vyřešili/nevyřešili.

Výhled

Zbytek první poloviny semestru bude uspořádán víceméně dle vašich přání:

- Mám povídat více o základních nástrojích jako je make, CVS?
- Mám zařadit Úvod do LaTeXu?
- Mám povědět více o tom, jak vevnitř pracují jednotlivé součásti systému?

Druhá polovina semestru už bude určena vašim souhrnům:

- jak to jde, co jste vyzkoušeli, jak to pomáhá a nepomáhá
- s čím stále narážíte

Domácí úkol

Není! ☺

. . . ale do příště si rozmyslete, na které části byste chtěli pracovat, a jak se sdružíte do skupinek.

Nezapomeňte: Příští hodina proběhne formou licitace témat.

References