Ročníkový projekt LS 2007/08, ZS 2008/09

Zadání ročníkového projektu

Téma: Doplňování tvaroslovné informace k chybějícím slovům textu

Anotace: Český akademický korpus (http://ufal.mff.cuni.cz/rest/CAC/cac_10.html) je poměrně rozsáhlá banka českých textů.U každého slova korpusu je doplněna informace mj. o slovním druhu a jeho kategoriích (rod, číslo, pád, osoba aj.). V textech jsou věty, ve kterých chybí více či jedno slovo. Detekce těchto podezřelých míst byla provedena ručně. Cílem rp je implementace automatické procedury, která u chybějících a ručně lokalizovaných slov doplní informace o slovním druhu a dalších kategoriích.

Student: Jan Václ, *e-mail*: janvacl@centrum.cz

Vedoucí: Barbora Vidová Hladká, ÚFAL MFF UK, *e-mail*: hladka@ufal.mff.cuni.cz

ŘEŠENÍ PROJEKTU

Časový harmonogram

LS 2007/2008

1. nastudovat formát CSTS, prostudovat ukázková data z Českého akademického korpusu 2.0 (viz přiložené CD-ROM)
2. seznámit se s metodami tagování
3. seznámit se s nástrojem tool_chain – viz přiložené CD-ROM,
4. zpracovat data z CD-ROM nástrojem tool_chain s parametry pro tagování,
5. navrhnout proceduru pro doplňování morfologických značek,
6. zvolit testovací data – výběr z dat Pražského závislostního korpusu 2.0
7. sepsat podrobnou specifikaci,
8. implementovat pilotní verzi,
9. testovat proceduru na testovací množině,

ZS 2008/2009

10. vylepšit pilotní verzi,
11. otestovat ji na testovací množině,
12. sepsat závěrečnou dokumentaci.

Studijní materiály

Monografie

Barbora Vidová Hladká a kol. *Průvodce Českým akademickým korpusem 1.0*, Karolinum, 2007. http://ufal.mff.cuni.cz/rest/CAC/cac_10.html

Hans van Halteren. *Syntactic wordclass tagging*. Kluwer Academic Publishers 1999.

Tokenizace, morfologická analýza, tagování -
http://ufal.mff.cuni.cz/morce/cac/?chapter=3#nastroje-zprac

Formát CSTS

- stručný popis - http://ufal.mff.cuni.cz/morce/cac/?chapter=3#data-format
- podrobná dokumentace - http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/cz/html/ch03.html#a-data-formats-csts

Ukázka souborů v reprezentaci CSTS – viz přiložené CD-ROM; soubory obsahují ručně doplněné morfologické značky

Pražský závislostní korpus 2.0 – http://ufal.mff.cuni.cz/pdt20

Nástroj tool_chain

- stručný popis - http://ufal.mff.cuni.cz/morce/cac/?chapter=3#nastroje-zprac
- dále viz přiložené CD-ROM

Obsah podrobné specifikace

1. Detailní popis problematiky, podle něhož by jiný programátor napsal "tentýž" program.
2. Návrh struktury programu (moduly, knihovny, vzájemná provázanost).
3. OS, jazyk, vývojové prostředí.

Obsah pilotní verze
Značkovací procedura s úspěšností alespoň 80% správně přiřazených značek.

Obsah závěrečné dokumentace

1. Instalační příručka - popis instalace a spuštění programu.
2. Uživatelská příručka - popis ovládání programu.
3. Programátorská dokumentace - postup překladu programu, popis implementace (co, v čem, jak), popis netriviálních algoritmů.

ORGANIZAČNÍ ZÁLEŽITOSTI

Termíny
30. června 2008 – odevzdání pilotní verze

Projektová wiki stránka – https://wiki.ufal.ms.mff.cuni.cz/user:hladka:jan-vacl