Oponentský posudek doktorské disertační práce

Drahomíra „johanka" Spoustová: Kombinované statisticko-pravidlové metody značkování češtiny

Doktorská disertační práce D. Spoustové se zabývá hybridními metodami pro morfologické značkování českých textů.
Práce je tvořena pěti kapitolami: první charakterizuje výchozí situaci, autorka v ní věnuje pozornost taggerům pro morfologickou desambiguaci v češtině. Ve druhé kapitole najdeme popis valenčního slovníku českých deverbativních adjektiv vzniklého převodem z povrchových valencí českých sloves. Třetí kapitola se zabývá kombinovanými metodami značkování, ve čtvrté kapitole autorka prezentuje rozšíření desambiguačních pravidel na syntax, pátá kapitola je závěrečná.

Práce je psána česky, výklad je přehledný a srozumitelný, text je typograficky kvalitní.

a) formální chyby: s potěšením konstatuji, že v práci je minimum překlepů, i když pár se jich najde (s. 16, 19, v seznamu literatury, s.82). Použití korektoru by pomohlo.
b) K terminologii je třeba konstatovat, že v bohemisticky orientované práci se vyskytuje naprosto zbytečná bastardizace, např. na s. 49 (feature-based tagger), s. 50 (recall morfologické analýzy). Vnucuje se otázka, v jakém jazyce tyto výrazy vlastně jsou? K termínům ‚recall', ‚precision' a ‚feature' přece existují přesné české ekvivalenty.

Přínos práce:

a) Problematika morfologické desambiguace zkoumaná v práci je aktuální, získané výsledky jsou nové. Lze konstatovat, že práce přináší tři hlavní výsledky
   - algoritmus převodu valenčních rámců sloves na valenční rámce deverbativních adjektiv, který pracuje s derivačními vzory, a vytvoření seznamu adjektiv s těmito rámci,
   - vytvoření hybridních (kombinovaných) technik morfologické desambiguace a jejich vyhodnocení,
   - pokus rozšířit desambiguační pravidla na syntax věty.

Problematické body a otázky:

a) Jak se řeší případy, kdy přiřazení značky je obecně sporné? Typicky jsou to případy, kdy značku nedokáže jednoznačně přiřadit ani člověk. Některé výrazy tohoto typu jsou navíc dosti frekventované, např. *jako* má v Syn2000 četnost 317 791 nebo *ale* s četností 387 457. Jsou pak zdrojem nepochybně nezanedbatelného procenta chyb. V práci o desambiguaci bych očekával aspoň stručnou charakteristiku tohoto problému, případně stanovisko k němu.
b) Dalším zdrojem problémů u statistických taggerů je fakt, že trénovací data jsou zpravidla jiná než data, na nichž tagger pracuje v konkrétní aplikaci. V práci se o tom autorka nezmiňuje, mohla by se při obhajobě pokusit tuto situaci charakterizovat s ohledem na chybovost?
c) Jakou roli hraje při desambiguaci subklasifikace slovních druhů ? Souvisí jistě s počtem použitých značek, který nepochybně ovlivňuje vyhodnocování úspěšnosti/chybovosti?

d) Pokud jde o algoritmus převodu slovesných valenčních rámců na adjektivní valenční rámce, je zjevné, že přegenerovává a poskytuje tvary, jež jsou snad správně utvořeny, ale nedoloženy v Syn2000, např. *zajedší, rozehřavší, vypásší, zmátší* (s. 32, 34) aj. Přegenerovávání nepokládám za nijak kritické, ale bylo by jistě vhodné nabídnout k němu potřebný komentář.

e) Za přínosný pokládám autorčin pokus aplikovat desambiguační pravidla na syntax, i když ona sama jej nehodnotí jako mimořádně úspěšný. Po metodologické stránce jej pokládám za motivující pro další směry výzkumu.

Hodnocení:
Autorka si v práci vytyčila některé metodologické a teoretické cíle – její soubor hybridních technik se jeví jako slibné řešení, které může překlenout propast mezi pravidlovým a statistickým přístupem v oblasti morfologické desambiguace (a nejen v ní).

Závěr:
Autorka svou prací prokázala, že se dovede samostatně vyrovnat se složitými problémy v oblasti počítačového zpracování přirozeného jazyka. Předloženou disertační práci pokládám za **výborný podklad** pro získání stupně Ph. D.

V Brně, 10. 8. 2007

Karel Pala
Katedra informačních technologií
Fakulta informatiky MU