Oponentský posudek doktorské disertační práce

Mgr. Jiří Semecký: Verb Valency Frames Disambiguation

Obsah práce

Předložená práce se zaměřuje na disambiguaci slovesných rámců v češtině. Autor experimentuje s nasazením různých statistických metod. Postupně analyzuje použití těchto algoritmů: naivní Bayesův klasifikátor, rozhodovací stromy, pravidlové systémy, support vector machines (SVM) a metodu maximální entropie. Každý z těchto algoritmů autor vyhodnocuje separátně a zároveň zkouší nalézt takové kombinace charakteristik (features), které by vylepšily chybovost disambiguace. Pro experimenty jsou použity korpusy VALEVAL a Prague Dependency Treebank.

Práce je rozdělena do osmi kapitol. Po úvodní kapitole autor v krátkosti představí algoritmy, které bude pro disambiguaci používat (kap. 2). Třetí kapitola vysvětluje motivaci, která stojí za disambiguací valenčních rámců a dále vysvětluje vztahy mezi významy slovesa a valenčními rámci. Autor zde přijímá (pro práci klíčovou) premisu, že význam slovesa lze aproximovat valenčním rámcem.
Ve čtvrté kapitole autor shrnuje použité korpusy a další zdroje dat. Pátá kapitola sumarizuje vědecký pokrok v oboru disambiguace významů slovesa.
Šestá a sedmá kapitola jsou nosnými kapitolami celé práce a obsahují nové vědecké poznatky v oboru disambiguace valenčních rámců. Velká péče je věnována výběru charakteristik a dále je popsáno přesné nasazení výše zmíněných disambiguačních algoritmů. Autor na obou použitých korpusech nachází optimální metody/charakteristiky a přichází s velmi zajímavými pozorováními.
Poslední kapitola shrnuje dosažená vylepšení oproti „baseline". Zde bych autorovi lehce vytknul přílišnou triviálnost metody, pomocí které "baseline" dosáhl.

Celá práce je psána anglicky, relativně srozumitelně a doplněna seznamem literatury.

Přínos práce

Nepochybně významným úspěchem je nasazení klasických metod z oboru strojového učení na tak „velmi lingvistický" problém jako je určování významu slovesa. Statistické metody si už sice našly svoji cestu do morfologické analýzy, parsingu nebo i strojového překladu, ale autorova precizní analýza opět o trochu více sbližuje svět statistických metod a „klasické lingvistiky". Speciálně oceňuji použití metody SVM, která by si jistě zajistila v komputační lingvistice více prostoru.

Mezi přednosti práce je třeba uvést také to, že autor své postupy hojně ilustruje na příkladech a podkládá kvantitativními údaji.

Dotazy k obhajobě

Je pro mě překvapující, že pro korpus VALEVAL a pro korpus PDT vyšla pokaždé jiná optimální metoda. Má autor nějaké vysvětlení?
Pro korpus PDT, bylo optimálního výsledku dosaženo metodou SVM za použití charakteristik S+I. Přidávání dalších charakteristik již pouze škodilo. Má autor nějaké vysvětlení?
Je možné, že další konkrétní dotazy vyplynou ještě z průběhu ústní prezentace a diskuse.

Závěr

Autor v předložené práci prokázal, že dovede samostatně vědecky pracovat a řešit složité problémy v oblasti matematické lingvistiky. Doporučuji, aby práce byla přijata jako disertační a aby Matematicko-fyzikální fakulta Univerzity Karlovy v Praze udělila po ukončení disertačního řízení Jiřímu Semeckému titul Ph.D.

Praha, 12.8. 2007
RNDr. Pavel Krbec, Ph.D.
NetCentrum, s. r. o.
Drtinova 10, 150 00 Praha 5