Posudek diplomové práce

Posudek vypracoval: RNDr. Ondřej Bojar; bojar@ufal.mff.cuni.cz
ÚFAL MFF UK
Malostranské náměstí 25
Praha 1
181 00

Dne: 3. 5. 2012

Název diplomové práce: Klasifikátor pro sémantické vzory užívání anglických sloves
Řešitel: Bc. Vincent Kríž
Vedoucí: RNDr. Martin Holub, Ph.D.
ÚFAL MFF UK

Diplomová práce Vincenta Kríže se zabývá automatickým přiřazováním slovesných vzorců (patterns) výskytům sloves v korpusu. Práce tak představuje vhodné empirické ověření metody „corpus pattern analysis“ (CPA).

Práce je přehledně členěna do osmi obsahových kapitol a závěru. Úvodní kapitoly jsou věnovány potřebným rešerším: metodě CPA, metodám strojového učení a konkrétním starším snahám o řešení podobné úlohy. Následuje popis dostupných dat, množiny rysů, které by měly být relevantní podle zkušeností lingvistů a konečně autorova vlastní práce: příprava dat pro nasazení algoritmů strojového učení a série experimentů ve snaze zlepšit úspěšnost metody. Kromě standardních dodatků jako je nadprůměrně rozsáhlý přehled literatury stojí za zmínku bohatý materiál v přílohách, včetně podrobných výsledků provedených experimentů.

Z celé práce je evidentní autorův metodický postup a pečlivost v drobnostech. Rád bych vyzdvihnul podrobně zdokumentované úsilí o volbu vhodných rysů (feature selection).

Následující výčet je spíš seznamem námětů k zamyšlení, než výhrad k práci:

- Má práce ambici potenciálně identifikovat vzorce i pro nová slovesa? Je to vůbec v CPA možné? Jedním z praktických důsledků „syntaktičtějších“ přístupů, jako je valenční teorie FGD, je možnost přisuzovat jeden a tentýž rámec víc různým slovesům. Teoreticky je tak možné klasifikátor záměrně nasadit na výskyty zcela nového slovesa. Autor o toto použití zjevně neusiluje, explicitní zmínku o tom jsem ovšem nenašel. Zejména sekce 7.3 by si ji ovšem zasloužila.
- V souvislosti s předchozí otázkou by mne zajímalo, proč autor volbu rysů provádí pro všechna slovesa současně. Například seznamy uvažovaných předložek ap. mohly být sestaveny vždy pro konkrétní sloveso.
- Jaká je motivace binarizovat všechny rysy? DT a ADA mohly jistě pracovat s kategoriálními hodnotami a nabízelo se tak další možné srovnání.
- Autor zavádí skupiny sloves A, B a C, aby zamezil mimořádnému vlivu jednoho slovesa na celkový výsledek. Není tedy skupina A se třemi slovesy spíš zhoršení situace, když její váha vzrostla na 80 %?
- Tab. 8.1.: Je zajímavé, že skupiny B a C se liší zlepšením. B dosáhlo menšího ERD než C. Spočívá vysvětlení v počtu výskytů sloves v jednotlivých skupinách?
- Z textu práce není zcela zřejmé, zda metoda CPA počítá s tím, že jeden výskyt slovesa může

být současně (exploatací) více vzorců. V takovém případě jsou autorovy klasifikátory jen aproximací a výhledově, nikoli v této práci, je třeba v principu umožnit přiřazení více vzorců jednomu výskytu.

- Proč je žebříček A a B sestaven z klasifikátorů určujících jen ano/ne pro každý vzorec? Tento postup může být prvním krokem k predikci vícero vzorců pro jeden výskyt slovesa, viz předchozí otázka, neodpovídá však zcela současnému způsobu nasazení.
- Pro zajímavost dodávám postřeh, že BestAU je lepší než BestMU o cca 2 procentní body jen u sloves ally, cry, deny, plug, a smell. Zejména skutečnost, že první tři slovesa potřebují změkčení při tvorbě minulého času by mohla naznačit, že výsledky jsou ovlivněny i náhodnými vlivy jako chyby v automatické analýze.

Po formální stránce je předkládaná práce zpracována pečlivě, kladně hodnotím zejména logické uspořádání, stručnost a jasnost vyjádření a množství přehledných ilustrací a tabulek. Mohu-li soudit slovenský text, práce obsahuje jen zcela minimální množství překlepů (avarage, Yaroského, sémenaticky, prepredstavuje). Nepatrné obtíže jsem měl jen s pochopením popisků v některých tabulkách: T4.1 na straně 36 nemá vysvětlené popisky, T5.4 a T5.5 na straně 43 mají chybné popisky (mluví o slovesech místo o skupinách).

Diplomová práce Vincenta Kríže beze zbytku splnila vytyčený úkol navrhnout a vyhodnotit automatický klasifikátor výskytů anglických sloves do slovesných vzorců metody CPA. Práci jednoznačně **doporučuji k přijetí** a navrhuji celkovou známku **výborně**.

V Praze dne 3. 5. 2012,

Ondřej Bojar