Oponentský posudek doktorské disertační práce

Pavel Straňák: Annotation of Multiword Expressions in the Prague Dependency Treebank

Doktorská disertační práce P. Straňáka se zabývá anotováním víceslovných výrazů (VSV) v pražském závislostním korpusu (PDT2) a nástroji vytvořenými pro tento účel.
Práce je tvořena osmi kapitolami: první je úvodní a motivační. Ve druhé kapitole najdeme charakteristiku VSV a také zdrojů dat mimo PDT2. Ve třetí kapitole autor popisuje situaci kolem VSV ve vztahu k PDT2, ve čtvrté kapitole mluví o pražském značkovacím jazyce (PML) použitém pro anotování VSV, návrhu schématu s-dat, jejich vizualizaci a nástrojích - editorech SemAnn a TrEd, v páté a šesté kapitole se oba nástroje probírají podrobněji. V sedmé kapitole je popsán způsob vlastní anotace VSV a hodnocení mezianotátorské shody, osmá kapitola je závěrečná.
Práce je psána anglicky, výklad je vcelku přehledný, ale místy poněkud fragmentární, což způsobuje, že srozumitelnost textu není vždy optimální. V textu se najdou překlepy a gramatické chyby, např. *byt* v poznámce na s. 11, *redifined* na s. 25, shoda - *creates* na s. 21, druhý odst., slovosled v titulku na s. 15 – *How are things in PDT2?*. Dále je autor nekonzistentní v psaní interpunkce (*however* na s. 9, 12, 17, 21, 41, 45, 46, 48, 52, 53, 56, 58, 61, 71).
Na s. 16 se v prvním odstavci mluví o nějakém článku (…,*which we report on in this paper*'…), co přesně má autor na mysli? Kde je vysvětleno, co znamená zkratka *SVG* na s. 24-5? Použití korektoru by pomohlo některé chyby odstranit. Po typografické stránce je text práce standardní.

Přínos práce:

a) Problematika víceslovných výrazů a jejich anotování zkoumaná v práci je v oblasti NLP velmi aktuální. Přínos práce vidím v následujících výsledcích:
   - vytvoření editoru SemAnn, v němž se editují anotace získaných VSV,
   - seznam VSV uložených v souboru s názvem SemLex a čítající 30 506 VSV získaných z jiných zdrojů než PDT (s. 48-49) a 8 816 VSV z PDT (s. 55), celkem tedy SemLex obsahuje, pokud jsem správně pochopil, 39 322 VSV. Jde o nový výsledek.,
   - extenze nástroje TrEd, k němu ale viz otázku níže (bod c),
   - popis způsobu anotování českých VSV a návrh struktury hesla v SemLexu v návaznosti na anotaci jednotlivých rovin v PDT2 – tento výsledek je originální a použitelný v dalších projektech,
   - hodnocení mezianotátorské shody – navržená metodika hodnocení se jeví jako uplatnitelná i v dalším výzkumu.

Problematické body a otázky:

a) zdroje VSV a výsledná podoba SemLexu – autor nechal stranou existující českou databázi VSV (Pala, Svoboda, Šmerk, Czech MWE Database, publikováno v Proceedings of LREC 2008), která obsahuje rozsáhlý seznam českých VSV čítající cca 160 000 položek. Pokud by autor ke zmíněné databázi přihlédl, nepochybně by to viditelně ovlivnilo výsledky práce (rozsah SemLexu?).
b) práce je krátká, místy fragmentární, do textu na s. 29-39 jsou jako vycpávka (?) vloženy kusy kódu, které by, když už, snad měly být uvedeny v příslušné příloze? Kromě toho z textu nevyplývá dost jasně, jaký je tu skutečný autorův podíl (viz s. 29, Listing 4.2, je

zde poznámka, written by P. Pajas'?).

c) podobně u nástroje TrEd, který není řádně citován (i když pochází z dílny ÚFALu, v uvedené literatuře jsem nenašel korektní odkaz), není dostatečně zřejmé, co je vlastní autorův výsledek a co přebírá (viz předchozí bod). Byl bych proto rád, kdyby P. Straňák při obhajobě uvedl podíl svého autorství zejména ve vztahu k extenzi TrEdu.

d) na s. 10 práce se říká, že anotace VSV je fakticky totéž co jejich identifikace. V této souvislosti bych čekal rozsáhlejší zmínku o automatickém rozpoznávání VSV, které je v oblasti NLP velmi aktuálním tématem. Tato problematika by do práce zapadla velmi dobře a nebylo by pak potřeba vyplňovat text práce kusy kódu (viz výše), navíc bez jakékoli kvalitativní interpretace.

e) proč se uvádí klasifikace VSV jen u NE, a nikoli u ostatních VSV? Anotátoři se o ni jistě mohli pokusit, když už měli jednotlivé VSV v ruce?

f) na s. 46 a 71 autor zmiňuje anotační logy jako jedinečný rys práce: jinými slovy, jde o žurnálování, což je ovšem v lexikografických aplikacích běžná technika. Obávám se, že tu autor poněkud objevuje Ameriku. Kromě toho předposlední odstavec na s. 46 mi není srozumitelný.

g) proč není k disertační práci přiloženo CD s plným textem práce a příslušnými přílohami a celým SemLexem nebo aspoň jeho ořezanou demo verzí? Pokud jde o přístup k nástrojům i SemLexu, na adrese http://ufal.mff.cuni.cz/lexemann/mwe/data.html se mi objevila hláška (6. 8. 2010): ‚Tady zatim nic neni'. Tuto informaci musím pokládat za neseriózní a vzbuzující fakticky i jisté pochybnosti o kvalitách disertační práce jako takové. Pokud autor nemůže poskytnout slíbené informace, neměl by adresu nabízet. Kromě toho jsou na uvedené stránce uvedeni autoři dva – P. Straňák a A. Bejček, takže se opět se musím zeptat, jaký je tu podíl autorství P. Straňáka?

h) Na s. 50-54 se popisuje struktura hesla v SemLexu, jehož součástí jsou ‚glosses' (definice?) – zde se nabízela možnost podívat se podrobněji na jejich typy a jejich strukturu, případně, v jakém jsou vztahu ke standardním slovníkovým definicím, např. v SSJČ nebo SSČ (nebo v jiných zdrojích)? V souboru SemLex.yml, který mi autor mezitím poslal, jsem na příklady definic nenarazil.

Hodnocení:

Autor si v práci vytyčil celkem ambiciózně některé metodologické a teoretické cíle – podle mého názoru jich dosáhl jen zčásti, což je patrné i z rozsahu práce. Práce byla zjevně psána ve spěchu a najdou se v ní některé relevantní metodologické nedostatky, zejména (viz výše) jde o problémy s (ne)citováním literatury. Výsledky týkající se nástrojů, jako je SemAnn, hodnotím jako průměrné. Na druhé straně rád konstatuji, že práce přináší některé nové a originální poznatky.

Závěr:

Přes uvedené výhrady dospívám k závěru, že P. Straňák dovede samostatně řešit problémy v oblasti počítačového zpracování přirozeného jazyka. Předloženou disertační práci lze akceptovat jako podklad pro získání stupně Ph. D., bude však potřeba přihlédnout k průběhu obhajoby.

V Brně, 10. 8. 2007

Doc. PhDr. K. Pala, CSc.
Katedra informačních technologií FI MU