Oponentský posudek disertační práce

Kandidát: RNDr. Eduard Bejček, MFF UK
Název práce: Automatické propojování lexikografických zdrojů a korpusových dat
Oponent: doc. Ing. Zdeněk Žabokrtský, Ph.D., MFF UK

Popis práce

Uchazeč se v předložené práci zabývá dvěma tématy: automatickým propojením ručně anotovaných valenčních slovníků a automatickým vyhledáváním víceslovných výrazů v textu.

Práce je členěna následovně. Po dvou úvodních kapitolách následuje kapitola s přehledem korpusů relevantních pro tuto disertaci a kapitola s popisem souvisejících slovníků a analýzou jejich formátů. Obě hlavní témata jsou pak po řadě zpracována v páté a šesté kapitole. Sedmá kapitola je krátkým pojednáním o propojování jednotek uvnitř jednotlivých slovníků, sedmá kapitola je závěrečná. Včetně rejstříku pojmů a zkratek a seznamu literatury má práce 213 stran.

Hodnocení práce

Zaměření předložené práce je velmi aktuální – v posledních dvou desetiletích vznikají značným tempem nejrůznější zdroje lingvistických dat a souběžně s tím je patrný i růst snah o integraci zdrojů do větších, v rámci možností unifikovaných a vnitřně propojených celků. Tento trend je viditelný jak uvnitř jednotlivých jazyků, tak napříč jazyky.

Při prvním čtení práce mě překvapilo, že jsou v ní zpracována dvě poměrně vzdálená témata – propojování dvou valenčních slovníků a rozpoznávání víceslovných výrazů. Autorovo představení jednotícího rámce (str. 3) na mě působí ponekud uměle (ve smyslu předložené kategorizace by bylo s trochou nadsázky možné za propojování jazykových zdrojů označit v zásadě jakoukoli úlohu z NLP). Nicméně hned dodávám, že obě témata jsou pojednána velice detailně.

Po formální stránce je práce zpracována velice pečlivě. V textu jsem nalezl jen nepatrné množství jazykových prohřešků: výraz „gold data" by bylo lepší překládat, zejména je-li skloňován (např. „na gold datech" na str. 6); překlep „ve a-stromě" str. 18; výraz „třeba" ve smyslu „například" (třeba na str. 2) je v SSJČ zatím považován za hovorový; překlep „započítej každý homografy zvlášť" na str. 78; několik málo chybějících čárek.

Pokud jde o obsahovou stránku, je třeba připomenout, že autor se prakticky v celém textu musí potýkat s centrálním problémem lexikografie, a sice s vymezováním jednotlivých významů slova (ve zvoleném pojmosloví delimitace lexií). Přestože tato otázka byla analyzována z různých úhlů předními lexikografy (ve světovém i národním měřítku), všeobecně přijímané řešení se nezdá být na dohled a lexikografové se prozatím zdají být odsouzeni k věčnému rozdělování na „lumpers" a „splitters". Zadání prvního úkolu, tedy propojení rámců mezi slovníky PDT-Vallex a VALLEX tento problém umocňuje. V situaci, kdy má být nalezeno propojení mezi dvěma slovníkovými hesly s dejme tomu deseti a dvanácti rámci, jejichž vymezení bylo navíc do značné míry intuitivní, je opravdu těžké rozhodnout, která možnost z kombinatorického množství $m{:}n$ bipartitních grafů je nejvhodnější. Autor si je ovšem tohoto problému velmi dobře vědom, na několika místech důsledky této situace poctivě popisuje (a vysvětluje, proč nejde o okrajový jev) a snaží se alespoň o statistické přiblížení.

Na práci dále kladně hodnotím následující:

- Vítám podrobné zpracování otázky formátů dat. Na první pohled se to může jevit jako čistě technická záležitost, ovšem srozumitelnost datového formátu nejenže podmiňuje šanci na úspěch datového zdroje u budoucích uživatelů, ale také zpětně ovlivňuje to, jak o daném slovníku sami přemýšlíme.
- Pro účely provázání obou slovníků autor navrhuje formát, který je na úrovni makrostruktury oproti slovníku VALLEX výrazně zjednodušený. Přestože to považuji za určitý ústup z pozic, neboť například zpracování vidových protějšků uvnitř jednoho lexému se mi stále jeví jako lingvisticky adekvátnější než duplikace sdílených rámců (podobně s homografy a pravopisnými variantami), připouštím, že výhody jednoduchosti datové struktury jsme v dřívějších verzích VALLEXu podceňovali.
- Práce je proložena řadou příkladů, které jednak dokládají, že autor má data opravdu „v ruce", a jednak čtenářovi ilustrují některé jevy, kde by nalezení příkladů introspekcí bylo velmi obtížné, např. víceznačný víceslovný výraz „přímá volba" na str. 158 (naopak o některých příkladech by bylo možné polemizovat, např. autorem nepřipouštěný singulár ve výrazu „čistírna odpadní vody" je používán běžně).
- Oceňuji pečlivé vyhodnocení experimentu, například vyhodnocení mezianotátorské shody po jednotlivých množinách sloves v tabulce 5.3.

K předložené práci mám jedinou zásadnější výhradu. Podle mého názoru je metoda zvolená pro propojování rámců VALLEXu a PDT-Vallexu, tedy autorova ruční implementace pravidel založených na regulárních výrazech a ruční přiřazení vah těmto pravidlům, poměrně těžkopádná. V důsledku tak pravděpodobně neumožňuje vytěžit informaci obsaženou v ručně anotovaném párování, které bylo vytvořeno z důvodu značné anotační náročnosti jen pro malý vzorek sloves. Domnívám se, že pružnější řešení, jako například náhrada malého množství pevně implementovaných pravidel řádově větším množstvím rysů (rozgenerovaných z „feature templates"), formulace jasného optimalizačního kritéria a nalezení vah některou standardní metodou strojového učení, by mohla vést k lepším výsledkům. Anebo by alespoň zvýšila míru jistoty, že z dostupných informací vyšší úspěšnost získat nelze. Podobně u rozpoznávání víceslovných entit: proč se nutit ke striktnímu rozhodnutí, kterou rovinu využít, když mohu použít všechny a vyvážení jednotlivých informačních zdrojů nechat na optimalizačním algoritmu?

V textu práce nalézám několik pasáží, které považuji za ne zcela přesné, popřípadě obtížněji srozumitelné:

- Str. 28: „pro aktanty je takový výčet úplný, jiné formy nejsou přípustné" - jednoduchým protipříkladem je pasivní diateze, kdy se přípustnou formou prvního aktantu stane. instrumentál, přestože nebývá uveden ve výčtu.
- Str. 156: „derivovaná reflexiva tantum" - tomuto výrazu nerozumím, pojmy „odvozené reflexivum" a „reflexivum tantum" se obvykle nepřekrývají (viz např. Encyklopedický slovník češtiny).
- Str. 37: „VerbaLex je bohužel v plné verzi jen těžko dostupný." Jde o eufemismistický komentář situace, že VerbaLex nebyl zveřejněn, nebo byl zveřejněn jen v omezené verzi, nebo je získání obtížné nějakým jiným způsobem?
- Str. 1: „informace ze slovníku jsou mnohem použitelnější, jsou-li exemplifikovány" - toto tvrzení by si možná zasloužilo podrobnější exemplifikaci, a to právě proto, že zní příliš samozřejmě. Např. v této práci se korpusová užití svázaná s PDT-Vallexem vzhledem k hlavní úloze (provázat VALLEX a PDT-Vallex) prakticky neuplatnila (resp. byl proveden experiment s využitím korpusových výskytů kontroly a reciprocity, ale nedopadl pozitivně).
- Vyváženost korpusů (str. 13) je notoricky problematický koncept, zde není jasné, jaký typ vyvažování má autor na mysli, ani jakým způsobem se to pojí k tématu práce.

- Na straně 116 a později v závěru na str. 169 je dosažená f-measure (0.77) hodnocena jako úspěch. Hodnota leží v očekávaném intervalu mezi úspěšnosti banální baseline (0.63) a mezianotátorskou shodou (0.91), obávám se ale, že bez srovnání s nějakou alternativní a přiměřeně komplexní jinou metodou je vyjádřené pozitivní hodnocení poněkud předčasné. Téma *word sense disambiguation*, se kterou daná úloha těsně souvisí, je ohromnou oblastí výzkumu a paletu existujících metod není možné vtěsnat do dvou přístupů uvedených na str. 131 (existují také neřízené metody, metody vycházející z paralelních korpusů, atd.).
- K obhajobě užití regulárních výrazů na str. 122: „Kdybychom se rozhodli pro jiné řešení, nevyhnuli bychom se nutnosti zavést nějaký formalismus pro popis pravidel... každý formalismus vykazuje … neobratnost". Tomu nerozumím, i jiné reprezentace informací dostupných ve srovnávaných rámcích přece mohou umožnit využití naší apriorní znalosti o úloze a přitom nejsou o nic méně „teoreticky neutrální" než použité regulární výrazy (např. výše uvedený převod do standardního vektoru rysů).
- Str. 71 – je pravda, že VALLEX nepopisuje vnitřní analytickou strukturu doplnění formálně vzato tak podrobně jako PDT-Vallex, nicméně domnívám se, že v drtivé většině případů (pokud jde o výskyty) je značka obsažené ve VALLEXu na a-strom přímočaře převoditelná.
- Dovoluji si polemizovat také s tvrzením na str. 95, že „automatická procedura propojující dva slovníky se nemůže opřít o nic jiného než o informace obsažené v obou slovnících, a tudíž vzájemně porovnatelné". Užitečnou informací pro párování lexikálních jednotek může být přece i to, kolik lexikálních jednotek u daného lexému rozlišuje nějaký třetí zdroj, aniž by na byl na zkoumané dva slovníky jakkoli navázán.

Otázky k obhajobě:

Rád bych požádal uchazeče o komentář k následujícím otázkám:

- Pokud by (hypoteticky) měly VALLEX a PDT-Vallex v budoucnosti splynout úplně, v čem by uchazeč spatřoval nejtěžší problém?
- K diskusi na straně 92: VALLEX pečlivě rozlišuje homonymní lemmata, PDT-Vallex je nerozlišuje, nicméně nebylo by možné k tomuto rozlišení využít technické přípony lemmat morfologické roviny v PDT?

Závěr

Domnívám se, že předložená práce je po obsahové i formální stránce zpracována kvalitně a že splňuje požadavky pro udělení akademického titulu Ph.D.. Práci doporučuji k obhajobě.

V Jahodově, 20. srpna 2015

doc. Ing. Zdeněk Žabokrtský, Ph.D.
Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta, Univerzita Karlova v Praze
Malostranské náměstí 25
118 00 Praha 1