Oponentský posudek disertační práce

Název:	*Automatické propojování lexikografických zdrojů a korpusových dat*
Autor:	Eduard Bejček
Obor:	matematická lingvistika, MFF UK
Rozsah:	197 stran textu

Předložená práce podrobně představuje postupy prací při propojování několika vybraných komplexních jazykových zdrojů (specializovaných lexikonů, slovníků, korpusů, ...). Autor člení práci na dvě hlavní části – v první polovině textu popisuje (více či méně podrobně) obsah a formáty jazykových zdrojů, se kterými dále pracuje nebo se na ně odkazuje v druhé části. Druhá část práce (kap. 5, 6 a 7) rozebírají konkrétní postupy při jednotlivých projektech propojování vybraných slovníků mezi sebou, případně vyhledávání výskytů slovníkových hesel v korpusech.

Text práce je psán česky, je v něm vidět velký důraz na kompletnost a preciznost. V textu se téměř nevyskytují překlepy ani gramatické chyby, většina otázek, na které čtenář při čtení narazí je následně alespoň stručně zodpovězena. Styl práce je čtivý, místy ovšem není vhodně členěn podle úrovní důležitosti informace – formální definice se volně propojují se spíše diskusními analytickými rozbory, důležité informace se občas nalézají v příkladech apod. Místy jsou také použity nevhodně přejaté slangové termíny např. skloňované *treebankům*, či *moduly parsují větu (tzv. analýza)*.

Přednosti a přínos práce:

a) podrobné představení a upřesnění vývoje a vztahů mezi zúčastněnými jazykovými zdroji ÚFAL,

b) precizní rozbor a analýza postupů při implementaci praktických automatických postupů pro inteligentní propojování daných jazykových zdrojů ("inteligentní" ve smyslu maximálního využití všech údajů dostupných pro automatické zpracování),

c) kvalitní a podrobný rozbor výstupů popisovaných automatických metod a jejich vyhodnocení na manuálně anotovaných vzorcích.

Nedostatky a nejasnosti v práci:

a) Práce je většinou až příliš prakticky orientována, autor několikrát přímo zamítá možnosti zobecnění daných postupů jako nedůležité pro daný účel.

b) Práce ne vždy důsledně cituje příslušné zdroje na očekávaných místech v textu, kde se daný projekt nebo jazykový zdroj představuje (např. Penn Treebank, VerbaLex, SemLink, anotace BBN entit od Josefa Tomana, ...).

c) Popis jednotlivých projektů v práci není vyvážený, výrazně převažují podrobnosti jediného projektu – propojení lexikonů VALLEX a PDT-Vallex (cca 40 stran). Ostatní tři projekty jsou popsány celkem na cca 5 stranách, přitom např. propojení strukturně výrazněji odlišných lexikonů VALLEX a FrameNet je zajímavé právě z pohledu univerzálnosti navržených postupů.

d) V práci zabývající se propojováním slovníků je zarážející absence podrobnější diskuze navrhovaných formátů ve vztahu k aktuálním tématům (otevřených) propojených dat (*Linked Open Data*). Ta se zaměřují nejen na úzké propojování konkrétních zdrojů na úrovni jednotlivých položek, ale právě na univerzální propojování (odkazování) mezi jednotlivými zdroji bez nutnosti přesné kompatibility formátů.

Otázky k obhajobě:

a) U schématu na str. 7 není jasný význam dvou dlouhých odkazů vedoucích do levého horního rohu.

b) Jaká je výsledná sada pravidel pro porovnání shody valenčních rámců? Z textu se dozvíme pouze příklady pravidla č. 1 a 5. Jaký je celkový počet pravidel? Do jaké míry jsou daná pravidla univerzální? Lze je (všechna nebo vybraná) použít obecně pro srovnání valenčních rámců i jiných formalismu, např. VerbNet nebo VerbaLex?

c) Pravidlo č. 1 na str. 101 zřejmě nebude standardně fungovat, jak je popisováno. První položka (`$formchunk*`) bude odpovídat všem zbývajícím variantám kvůli hladovému zpracování regulárních výrazů. Pro popisovanou funkcionalitu by místo operátoru * měl být použit operátor ?. Nebo je uvedený výraz zpracován odlišně?

d) V sekci 6.1 o vyhledávání víceslovných výrazů z PDT v textech ČNK je několikrát uvedeno, že se při vyhledávání nehledají pojmenované entity, i když "zdrojová" data informace o nich obsahují. Z textu ale není zřejmé, proč tomu tak je?

e) V sekci 6.1.5, zejména u hypotézy B, jsou navrhována pravidla pro slučování variantních výskytu víceslovných výrazů do společných významových jednotek. Není ovšem zřejmé, kdy a jak by se uvedená pravidla měla uplatnit? Má to být pouze podklad pro lidské anotátory? Nebo se má jednat o obecně platná automaticky aplikovaná pravidla?

Vzhledem k tomu, že se vždy jedná o případy *užití ve stejných kontextech*, byly by zde použitelné např. automatické metody distribučních sémantických modelů (word2vec, GloVe)?

Závěrečné hodnocení:

Práce se zabývá výrazně aktuálními tématy z oblasti automatického inteligentního zpracování jazykových zdrojů. Výsledky práce vidím zejména v precizní analýze, implementaci a vyhodnocení nástrojů na propojení vybraných komplexních jazykových zdrojů, což vede ke zpřesnění a aktualizaci dosavadních teoretických postupů. Pro širší využití výsledků práce v praxi by ovšem ve většině případů bylo nutné výzkum ještě dále generalizovat. Celkově konstatuji, že předložená dizertační práce **prokazuje** předpoklady autora k samostatné tvořivé práci a představuje vhodný podklad pro udělení titulu Ph.D.

V Brně dne 23. července 2015

doc. RNDr. Aleš Horák, Ph.D.
Fakulta informatiky
Masarykova univerzita, Brno