Oponentský posudek na diplomovou práci

Karel Vandas: Automatic Identification of Semantic Preferences for Valency Complementations of Verbs

Cílem diplomové práce je automatické určování sémantických preferencí pro slovesná doplnění.

Diplomová práce obsahuje 8 kapitol, dvě přílohy a CD-ROM. V první kapitole je představena motivace k tématu a struktura práce. Druhá kapitola je věnována valenci sloves obecně a jejímu místu v předložené práci. Neřízená metoda strojového učení shlukování (clustering) je popsána ve třetí kapitole. Čtvrtá kapitola je přehledem použitých datových zdrojů. Experimenty, jejich přehled a nastavení jsou popsány v páté kapitole. Jejich evaluace je předmětem kapitoly šesté. Závěrečná diskuse a shrnutí uvádějí kapitoly 7 a 8. Příloha A je uživatelskou příručkou výstupní aplikace a podrobnější implementační záležitosti shrnuje příloha B. Text práce doplňuje CD. Práce je psána anglicky.

Souhrnné hodnocení

Text práce je obsáhlý a je doplněn webovým rozhraním pro sledování výstupů experimentů a pro správu jejich nastavení. Hlavní výtka směřuje k samotnému textu, ze kterého je obtížné rekonstruovat jednotlivé kroky řešení. V textu jsou jistě uvedeny i popsány všechny komponenty řešení, ovšem hutnost textu a vysoká četnost nejasných pasáží velmi ztěžuje pochopení cílů práce a posouzení míry jejich naplnění.

Otázky a komentáře

1. V motivační části 1.1 je uveden termín tectogrammatical layer bez popisu, pouze s odkazem na literaturu. Bylo by vhodné termín alespoň stručně popsat. V závěru části se píše: „*Also valency lexicons and ontology for the Czech language are available.*“ O jaké ontologii se zde mluví?
2. V části 1.2 je popsána struktura práce. Formulace typu „*In chapter **Verb Valency** on page 3*“ jsou nezvyklé. Pokud autor chce uvádět stránkový rozsah kapitol, ať tak činí systematickým uvedením rozsahu stránek, a ne pouze odkazem na první stránku kapitoly.
3. V úvodním odstavci kapitoly 2 se píše: *It has been developing for generations with different types of influence on it.* Jak mám této větě rozumět?
4. Poznámka č. 1 na str. 3 by měla být doplněna odkazem do literatury.
5. V části 2.2 uvozovky působí nepatřičně.
6. V části 2.3 by bylo vhodné doplnit výčet valenčních slovníků ukázkami ne jako odkaz do přílohy, ale přímo v hlavním textu.
7. V práci se spojení *complementation abstraction* objevuje poprvé v nadpisu části 2.4. Jedná se o ustálený termín, nebo o autorovo pojmenování? V každém případě chybí jeho popis či definice. Existuje nějaká spojitost mezi výstupy shlukování a abstrakcí?
8. V části 2.5 je vhodné doložit tvrzení v druhém bodě odkazem do literatury.
9. V části 2.8.1 je zmíněna „nějaká“ ontologie. Jaká ontologie?
10. V části 2.8.3 je u citace disertace Jiřího Semeckého chybně uvedeno, že se v ní aplikují neřízené metody strojového učení pro desambiguaci slovesných rámců. V citované práci je na str. 20 uvedeno následující: *In this work, we will be using only supervised methods, and unless stated differently, the term machine learning methods will refer to supervised machine learning methods.*
11. Části 3.1 – 3.2 jsou základním popisem neřízené metody shlukování. Na několika řádcích jsou poznámky k vlastnímu autorovu řešení – je třeba oddělit obecný popis metody od popisu implementace. Část o evaluaci by měla být podrobnější.
12. Kapitola 4 seznamuje s daty, se kterými se provádějí experimenty. Formáty dat jsou bodem do uživatelské a programátorské dokumentace. Proč si autor vybral k experimentování korpus CzEng? Často je v textu použito spojení *CzEng output* – co to

prosím je? Tento korpus je zpracován řadou nástrojů – autor toto zpracování označuje jako *CzEng workflow*. O jaké nástroje se jedná?

13. Jakkoli se autor v části 5.1 snaží představit osnovu experimentů, není jasné, co jsou jejich vstupy a výstupy. Z jakých statistik a frekvenčních tabulek výběr sloves vycházel? Pokud čtu pozorně, tak až na str. 26 se rozlišují dvou typy rámců, a sice *mixed* a *separated*. Jedná se o zavedené pojmy, nebo je to autorovo pojmenování? Je vhodné doplnit jejich popis s příklady.
14. Rozumím tomu správně, že CzechWordNet chápe autor jako onu ontologii, kterou uvádí v předcházejícím textu?
15. Co znamená v části 5.2.4 věta … *might cluster together documents that share similar morphological structure...?*
16. V klíčové kapitole 6 je těžké se zorientovat. Proč je do ní začleněna část *Evaluation of CzEng Output for Distinguished Examples of Valeval* ? Jak mám prosím rozumět tabulkám 6.2 a 6.3?
17. Příloha 1 je uživatelská dokumentace aplikace pro sledování výstupů experimentů a pro správu jejich nastavení. Bylo by vhodné na jejím začátku uvést stručnou anotaci aplikace.
18. Práce v seznamu literatury nejsou abecedně seřazeny.
19. Instalace aplikace (bez jména?) proběhla dle instrukcí v dokumentaci přímočaře. Její ovládání je relativně intuitivní, i když celkový dojem kopíruje styl samotného textu práce.

I přes uvedené nedostatky doporučuji diplomovou práci Karla Vandase k obhajobě. Autorovi práce bych doporučila, aby se při obhajobě zaměřil na jasný výčet cílů, jasný popis jejich řešení a na jasnou prezentaci evaluace. Nechť doplní i seznam vlastních přínosů.

V Praze 30. srpna 2012

Barbora Vidová Hladká
ÚFAL MFF UK