Oponentský posudek diplomové práce

Autor a název předložené práce

Martin Kirschner: *Automatické vytváření sémantických sítí*
(druhá verze práce, předložená po loňské neúspěšné obhajobě)

Posudek předložené práce

Předložená práce se zabývá automatickou extrakcí sémantických lexikálních vztahů, a to na základě analýzy rozsáhlých korpusových dat a s pomocí metod strojového učení. Zadáním diplomové práce bylo „navrhnout, implementovat a evaluovat algoritmus, který bude […] budovat celé sémantické sítě“. Práce je implementační a experimentální. Úkolem studenta bylo zpracovat rozsáhlá korpusová data, navrhnout vhodné postupy pro získání sémantických lexikálních vztahů mezi slovy a sestavení sémantické sítě a výsledek vyhodnotit. Pokud je nám známo, jedná se o první práci tohoto druhu pro češtinu.

Student předložil práci, která má vcelku standardní strukturu: obsahuje úvod do problematiky (v 1. kapitole), rešerši známých postupů (ve 2. kapitole), jádro práce pak tvoří popis vytvoření trénovacích a testovacích dat (ve 3. kapitole) a popis použitých metod včetně jejich evaluace a diskuse o výsledcích (ve 4. kapitole). Tyto části předloženého textu mají dohromady 49 stran. V 5. a 6. kapitole je uživatelská a programová dokumentace k vytvořené implementaci (13 stran) a v 7. kapitole je závěr (2 strany). Následuje seznam použité literatury (34 položek) a přílohy (použité zkratky, obsah CD, ukázka 165 úspěšně predikovaných sémantických vztahů).

Na přiloženém CD jsou vedle vlastního textu dostupné jednak zdrojové kódy vytvořených programů středně velkého rozsahu (převážně v C++, dle autora přes 4500 řádků), jednak vypočítaná a anotovaná data. Výpočty byly prováděny na výpočetním clusteru s využitím prostředí Sun Grid Engine.

Za účelem získání lidského hodnocení sémantických vztahů mezi slovy byla provedena ruční anotace 900 párů slov, a to třemi nezávislými anotátory (viz kap. 4.2.1).

Rozsah a myšlenkovou hloubku (celkovou koncepci) celého díla hodnotím jako adekvátní požadavkům na diplomovou práci.

Práce byla předložena již podruhé, po loňské neúspěšné obhajobě. Proto mj. srovnávám s minulou verzí. Rozsah práce se ve srovnání s minulou verzí zvětšil (cca o 25%), a to jednak zvětšením použitého fontu a úpravou formátování (e.g. přidáním nadpisů), jednak přidáním řady ilustrativních obrázků/grafů a mírným rozšířením textu v některých sekcích. Zejména v kapitolách 3 a 4 autor lépe vysvětlil některé detaily použitých metod a také lépe a podrobněji zpracoval diskusi výsledků práce (v kap. 4.4 a 7).

Předložená verze je ve srovnání s minulou verzí celkově podstatně lépe zpracována. Nemá již nedostatky typu nekonsistentních definic, neexistujících odkazů, velkého množství gramatických chyb nebo chybějícího textu. Myšlenkové postupy a použité metody jsou již vcelku uspokojivě vysvětleny. Některé otázky k detailům: Proč byl zvolen "stejný počet" negativních instancí do trénovacích dat (viz kap. 4.1)? Proč autor považuje hodnotu $\beta=0,5$ za "nejvhodnější" (str. 36)?

Stejně jako v předchozí verzi práce mi však stále chybí "budování celé sémantické sítě", což byl cíl výslovně požadovaný v zadání. Např. v Závěru celé práce (kap. 7) není o konstrukci sítě jako takové ani zmínka(!). V kap. 4.4.1 k tomu autor uvádí, že konstrukce sítě nebyla provedena z důvodu nízké přesnosti extrakce sémantických relací. Nebylo však např. možné navržený model natrénovat s větším důrazem na *precision*, aby byla možná konstrukce alespoň "řídké", ale "smysluplné" sémantické sítě? Je možné z modelu *liblinear* popsaného v kap. 4.1.1 získat hodnoty pravděpodobnosti predikce a využít je k vyfiltrování výsledných relací s cílem zvýšit *precision*?

Další zásadní otázkou, která by mě zajímala, je, do jaké míry (zda vůbec?) bylo použití českého WordNetu (CWN) přínosem pro automatický odhad sémantických vztahů? V Závěru autor uvádí své zjištění, že pro popsané metody se CWN "nezdá vhodný". Rovněž v kap. 4.4.1 uvádí, že výsledky automatické metody jsou dokonce lepší než data extrahovaná přímo z CWN. To evokuje myšlenku, že trénování s využitím CWN mohlo věci možná spíše uškodit. Bylo tedy provedeno srovnání, jaká by byla úspěšnost, kdyby CWN vůbec nebyl použit? Tuto námitku proti použité metodě jsem vznesl již při minulé (neúspěšné) obhajobě, ale odpověď na ni v předložené práci nenacházím.

Závěr: Předloženou práci doporučuji k obhajobě. Požadavky na diplomovou práci splňuje. Ve srovnání s prvou verzí, kterou diplomant předložil v srpnu 2011, je předložená práce podstatně lépe zpracována.

V Praze, 23. ledna 2012

RNDr. Martin Holub, Ph.D.