Oponentský posudek diplomové práce

Autor a název předložené práce:
Martin Kirschner: Automatické vytváření sémantických sítí

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 35 stran. V 5. a 6. kapitole je uživatelská a programová dokumentace k vytvořené implementaci (12 stran) a v 7. kapitole je závěr (1 strana). Následuje seznam použité literatury (32 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 kapitola 4.3.1).

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

Text práce však není zpracován příliš dobře – pro ilustraci uvádím několik nejkřiklavějších nedostatků:

- Zmatené, nekonsistentní, nejasné nebo nepřesné (pokusy o) definice (e.g. „definice" na str. 6-7, dále „definice" cílové třídy (str. 25) nebo confusion matrix (str. 25-26)).
- Značné množství gramatických chyb (interpunkční, překlepy, neúplné věty).
- Odkazy na kapitoly v části 1.2 neodpovídají skutečným číslům kapitol.
- Zcela chybí popis některých zásadních detailů použitého postupu (e.g. booleanizace (kap. 3.3.3) nebo použití regrese při SVM učení (kap. 4.2.1)).
- Popisek pod tabulkou 4.8 odkazuje na konfidenční intervaly, které však nikde nejsou uvedeny.
- Chybějící text na konci kap. 4.4.3.

Za věcně nejslabší část považuji klíčovou kapitolu 4, kde dle mého názoru měl student nejvíce prokázat své odborné schopnosti. Automatický klasifikátor/detektor

sémantických vztahů je trénován pomocí dat získaných ze sémantické sítě Czech WordNet. Metoda strojového učení je však aplikována na velmi primitivní úrovni. Autor např. přiznává, že parametry SVM modelu nijak neladil (kap. 4.6). Volba lineárního SVM modelu není nijak zdůvodněna (kap. 4.2.1), využití hodnot regresní funkce není vůbec ukázáno. Zvolený model (resp. jeho výsledky) není srovnán s žádnou jinou metodou, přestože autor připouští, že by to mohlo přinést signifikantní zlepšení (kap. 4.6).

Velmi nejasná je zejména evaluace klasifikátoru v sekci „Extrakce nových relací" (kap. 4.3 a 4.4). Obsah prezentovaných tabulek (zásadní budou asi zejména tabulky 4.5 a 4.8) mi není příliš jasný a budu požadovat vysvětlení u obhajoby.

Celkový dojem, který předložená práce vyvolává, lze shrnout tak, že student odvedl dobrou práci při analýze problému a předzpracování dat, ale v (hlavní) části, kde jde o vlastní postup extrakce sémantických vztahů, jeho evaluaci a následné budování sémantické sítě, je předložená práce rudimentární. Prezentovaný postup pro extrakci sémanticky vztažených párů slov již nebyl dále nijak aplikován. Zadání práce požadující implementaci algoritmu pro budování sémantických sítí v tomto smyslu nebylo splněno.

Závěr: Předloženou práci doporučuji k obhajobě. Zda splňuje požadavky na diplomovou práci, považuji však za velmi diskutabilní.

V Praze, 30. srpna 2011