Posudek vedoucího diplomové práce

Martin Kirschner: Automatické vytváření sémantických sítí

V předkládané diplomové práci se Martin Kirschner zabývá automatickou extrakcí sémanticky asociovaných párů slov, které pak tvoří tzv. sémantickou síť, strukturu podobnou např. WordNetu. Použitý postup vychází ze statistické analýzy kontextů slov v textovém korpusu a používá metod strojového učení k predikci míry sémantické asociace mezi slovy. Vyhodnocení je provedeno jak automaticky, tak ručně.

Obsah práce

Text diplomové práce je relativně krátký, rozdělen do 7 kapitol. Experimentální část (první čtyři kapitoly) má 35 stran. Zbývajících 30 stran textu obsahuje dokumentaci, seznam literatury a přílohy. První kapitola obsahuje úvod a motivaci celé práce. Druhá kapitola je teoretická a zabývá se různými metodami konstrukce sémantických sítí. Ve třetí kapitole autor popisuje vytváření trénovacích a testovacích dat, která byla použita v experimentech. Čtvrtá kapitola popisuje vlastní experimenty a jejich výsledky. Pátá kapitola obsahuje uživatelskou a šestá kapitola programátorskou dokumentaci. Výsledky práce jsou shrnuty v závěrečné sedmé kapitole. Přílohy tvoří seznam použitých zkratek, obsah přiloženého CD a seznam 165 úspěšně predikovaných asociovaných párů slov. Práce má experimentálně-implementační charakter. K provedení experimentů bylo třeba vytvořit software středně velkého rozsahu.

Přínos a přednosti práce

Pokud je mi známo, tak tato práce je první, která se zabývá použitím *supervised* metod strojového učení pro predikci sémantické asociace mezi slovy. Doposud byly kontextové asociační míry používány nezávisle a tato práce se snaží využít možnosti jejich kombinace v lineárním modelu, který je trénovaný pomocí SVM na datech (párech slov v různých sémantických relacích) získaných z WordNetu.

Předností práce je relativně důkladná ruční evaluace, která nejen vyhodnocuje úspěšnost použité metody, ale také ověřuje některé hypotézy a předpoklady použité v práci. Za zmínku stojí také uživatelská a programátorská dokumentace, která je spíše nadprůměrná a jistě by usnadnila převzetí výsledků práce a vytvořeného kódu případným zájemcům.

Nedostatky práce

Hlavní nedostatek práce je v samotném textu. Je zřejmé, že byl dokončován v rychlosti, což se projevilo jednak v celkové neuhlazenosti textu, občasných překlepech a nesprávných

referencích (číslování definic, kapitol a tabulek, odkazy na neexistující části apod.), ale především v chybějících detailech popisovaných postupů, experimentů a jejich výsledků. Text není zcela samovysvětlující, místy je potřeba uplatnit jisté externí znalosti a občas je nutné si některé souvislosti domýšlet.

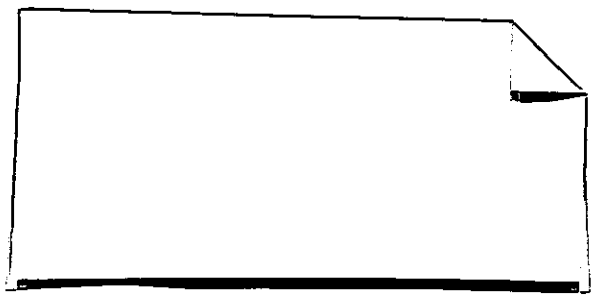
Autor se snaží o celkovou preciznost, což dokazuje množství explicitních definic v textu a užívání matematické terminologie, ale mnohdy je text nepřesný a použití termínů nekonzistentní. Problematická je např. definice pojmu „koncept" na str. 7 a jeho další použití v textu, kde (zřejmě) dochází k záměně s pojmem „lemma". V některých případech vymezení pojmu chybí a dochází k jeho použití v různých významech (např. „slovo" a „výskyt slova", apod.). Nejasná je také např. definice „sémantické sítě" na str. 8, zvláště tvrzení, že lze tento termín zaměňovat s pojmem „sémantický slovník". Problematické jsou i další definice, např. je kontext opravdu *množina* slov? V textu se objevují pojmy „similarity" a „relatedness". jak se mají k sobě navzájem a k pojmu „sémantický vztah", který je pro práci stěžejní (viz část 2.4.2 a 4.3.1), a co přesně znamenají sloupce „Similar" a „Related" v tabulce 4.8? Co měl autor na mysli výrazem „množina kontextu daného konceptu" (str. 19)? Co přesně znamenají jednotlivé položky na obrázku 3.3?

V textu chybí vysvětlení některých (pro experimenty důležitých) detailů, např.: Při přípravě dat byla použita tzv. boolenizace – jak byl nastaven její parametr a proč? Autor uvádí, že z Českého WordNetu extrahoval celkem 79 363 párů slov v nějaké sémantické relaci (tabulka 4.1) – kolik z nich (a kterých) bylo použito pro trénování a testování (krosvalidaci)? Jak přesně bylo provedeno trénování SVM? V práci se uvádí, že SVM bylo použito pro regresi a nikoliv pro klasifikaci (str. 27), a to z důvodu nutnosti odhadu pravděpodobnosti jednotlivých tříd – jak přesně byla tato pravděpodobnost získána? Jak bylo potom prováděno přiřazení tříd. A jak tomu bylo v případě predikce do tří tříd? V samém závěru práce (str. 51) jsou potom uvedeny dva body (zjištění), na základě jakých pozorování k nim autor došel?

Autor se soustředil na návrh a evaluaci algoritmu pro extrakci sémantických párů slov, ale navržený postup již nijak neaplikoval za účelem automatického vytvoření skutečné (a rozsáhlé) sémantické sítě a její případné analýze.

Závěr

I přes uvedené nedostatky lze diplomovou práci Martina Kirschnera doporučit k obhajobě. Předpokládám ovšem, že výše položené otázky budou při obhajobě dostatečně zodpovězeny.



22.8. 2011, Dublin, Irsko