Martin Kirschner: Automatické vytváření sémantických sítí

Posudek vedoucího diplomové práce

Autor se v předkládané diplomové práci zabývá automatickou extrakcí sémanticky asociovaných párů slov, které pak tvoří tzv. sémantickou síť. Použitá metoda je založena na statistické analýze kontextů slov v textovém korpusu a využívá strojové učení k predikci sémantické asociace mezi slovy. Vyhodnocení je provedeno jak automaticky, tak ručně.

Obsah práce

Text diplomové práce je rozdělen do sedmi kapitol. Experimentální část (první čtyři kapitoly) má 50 stran. Zbývajících 25 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 vzorek 165 úspěšně predikovaných asociovaných párů slov. Rozsáhlejší seznam extrahovnaých slovních párů je k dispozici na přiloženém CD. Práce má experimentálně-implementační charakter. Provedeno bylo relativně velké množství experimentů, k jejichž realizaci bylo třeba vytvořit software středně velkého rozsahu.

Hodnocení

Pokud je mi známo, tak tato práce je první, která se zabývá použitím metod strojového učení s učitelem pro predikci sémantické asociace mezi slovy na českých datech. 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 poměrně 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 nadprůměrná a jistě by umožnila převzetí výsledků práce a vytvořeného kódu případným zájemcům.

Práce je obhajována podruhé a je nutné konstatovat, že většina nedostatků z první verze byla odstraněna. Text je mnohem čitelnější a působí přehledněji. Formální nedostatky v číslování definic, kapitol a tabulek, odkazech na neexistující části apod. byly odstraněny zcela. Doplněny byly detaily experimentů a diskuse jejich výsledků. V některých případech nepřesnosti a nejasnosti zůstaly (např. definice hodnoty kontextu na str. 8, popis použití globálního filtrování v druhém odstavci části 3.3.2, chybějící vzorec u metriky Lin na str. 23, případně také chybné vysvětlení TFIDF modelu na str. 29, ilustrace extrakce relací z WordNetu na obr. 3.4, použití termínů slovo vs. lemma vs. lexikální jednotka).

Otázky k obhajobě

1. V kapitole 2 autor srovnává dvě metody extrakce sémantických vztahů: a) ze vzorců větné syntaxe (část 2.2) a b) na základě statistické analýzy kontextů slov (část 2.3). O prvním

přístupu tvrdí, že jeho výsledky jsou ovlivněny jednotlivými výskyty a že je citlivý na nestandardní použití slov (např. metafory) a postrádá tak robustnost a stabilitu. O druhém přístupu naopak tvrdí, že tyto nedostatky nemá a je více robustní. V obou případech jde ale o statistické zpracování korpusových data a problém jejich řídkosti je stejný. Jak by se lišily výsledky obou metod na stejných datech?
2. Jak je možné, že (hladový) algoritmus pro výběr rysů vybírá některé rysy vícekrát (viz tabulka 4.4). Přidání absolutně korelovaného rysu do lineárního modelu by jej přeci nemělo zlepšit.

Závěr

Předkládaná práce je výrazným posunem oproti předchozí verzi a splňuje požadavky kladené na diplomovou práci na MFF UK. Práci doporučuji k obhajobě.

Vypracoval: RNDr. Pavel Pecina, Ph.D. Praha, 21.1. 2012