Posudek vedoucího diplomové práce

Michal Richter: Pokročilý korektor češtiny

Michal Richter se ve své práci zabývá automatickou detekcí chyb a překlepů, které není možno opravit pouze podle slovníku, v psaném textu.

Cíl práce

Autor implementuje korektoru češtiny, který si klade za cíl usnadnit uživatelům práci zlepšením klíčových oblastí práce korektoru: schopnost nalézt překlepy či některé gramatické chyby, které korektor pracující jen porovnáním slov se slovníkem nalézt nemůže, a optimalizace pořadí nabízených náhrad podle jejich pravděpodobnosti.

Obsah práce

Práce je rozdělena do sedmi kapitol. Postupuje od zadání přes motivační úvod do problematiky a stručný popis použitých statistických metod. V kapitole 4 pak následuje popis vlastních použitých statistických modelů a dat použitých k jejich sestavení. Pátá kapitola se věnuje podrobnostem implementace. Následuje podrobná evaluace standardními statistickými metodami a závěr.

Hodnocení

Práce je napsána dobrou, srozumitelnou angličtinou. Přesto se místy kromě překlepů jako „markov" s malým „m" vyskytují i chyby ztěžující porozumění textu. Např na straně 14: "...a scoring function assigning the scores to syntactically or semantically malformed sentences and high scores to well-formed sentences." Na místě "the scores" má patrně být "low scores". Tuto semantickou chybu by s vysokou pravděpodobností neodhalil ani korektor, jenž je předmětem práce, některé snazší by ale odhalit mohl: "...is filled from left from to right ..." na str. 24, a další podobné.

Ojediněle se vyskytují také drobné nepřesnosti věcné. Např. na téže straně: "The probability [of all the sentences of a language] cannot be modelled directly, because the set of all possible sentences is tremendously large."

Z práce je zřejmé, že autor dostatečně rozumí použitým statistickým metodám, provedl pečlivé experimenty a formuloval dobré hypotézy i pro další práci (např. diskuse vhodných metod vyhlazování jazykových modelů na slovních formách a na morfologických značkách na str. 28 a 29).

Implementace je v zásadě plně funkční a díky pečlivé optimalizaci i rychlá a paměťově úsporná. Jde tedy o korektor, který je z hlediska uživatelského pohodlí v podstatě bezproblémový a může být běžně používán, jak jsem si v praxi ověřil.

V evaluaci poněkud postrádám srovnání s jinými dostupnými řešeními a to jak pro korekci pravopisu, tak pro doplnění diakritiky. Alespoň samotný korektor Aspell, jehož slovník autor použil pro základ svého slovníku, by bylo zajímavé zahrnout. Evaluace je

jinak velmi pečlivá a závěry z ní vyvozené jsou podle mě věcně správné. Na místa, kde nelze na základě použitých dat a provedených experimentu vyvodit závěry s dostatečnou jistotou, autor správně upozorňuje (na str. 60).

Neúplné dodržení zadání pokud jde o šíři funkcí korektoru proběhlo po konzultaci s vedoucím. Nepovažuji je za významné, protože jde hlavně o důsledek velkého množství kvalitní vědecké práce na klíčové funkcionalitě korektoru. Ten je ve výsledku velmi kvalitní a v některých ohledech přinejmenším pro češtinu jedinečný.

Závěr

Předložená práce podle mě splňuje požadavky na diplomovou práci na MFF UK a doporučuji ji k obhajobě.

V Praze 3. září 2010

Pavel Straňák