Posudek *oponenta* diplomové práce

Jméno a příjmení autora posudku: Michal Novák

Jméno a příjmení autora práce: Ján Václ

Název práce: Sledování aktivovanosti objektů v textech

Vlastní text:

Popis práce

Tématem předložené práce byla aktivovanost objektů v českých textech. Základem byl algoritmus pro sledování aktivovanosti (Hajičová, Hladká, Kučová, 2006), který byl již implementovaný včetně vizualizace. Cílem práce bylo tento algoritmus revidovat a reimplementovat, aby ho bylo možné pustit na větších datech, např. PDT 2.0. Výsledkem měly být grafy aktivovanosti, které měl řešitel vhodným způsobem interpretovat. Významným cílem práce bylo aplikovat metody strojového učení a uplatnit znalost aktivovanosti v nějaké další úloze NLP.

Text lze rozdělit na dvě části. První část, která čtenáře uvádí do problematiky, se skládá ze tří kapitol. V první kapitole řešitel ukazuje práce související s úlohou sledování aktivovanosti. Druhá kapitola popisuje klíčové teoretické koncepty, a to koreferenci včetně asociační anafory (bridging), teorii aktuálního členění a samotnou aktivovanost, včetně popisu existujícího algoritmu pro její sledování. Tato kapitola obsahuje i teoretický popis algoritmu strojového učení na modelování témat v textu – Latent Dirichlet Allocation (LDA), včetně popisu metod evaluace. Třetí kapitola seznamuje čtenáře s použitými datovými zdroji a nástroji.

Druhá část textu představuje výsledky autorovy práce. Ve čtvrté kapitole je technicky popsána reimplementace algoritmu sledování aktivovanosti, produkující i vizuální výstup. Kromě základních statistik jsou na základě grafů aktivovanosti navrženy i složitější popisné statistiky, které jsou následně vyhodnoceny a analyzovány. V poslední, páté, kapitole autor použije míru aktivovanosti při řešení dvou úloh – vizuální shlukování dokumentů a modelování témat pomocí LDA a vyhodnotí jejich úspěšnost oproti systémům používajícím četnosti slov.

Práce je psaná anglicky, má 54 stran včetně seznamu literatury, obrázků a tabulek. K práci je přiložen CD-ROM se zdrojovými kódy a daty.

Hodnocení

Na začátek je nutno připomenout, že práce je opravenou verzí původní práce, která nebyla obhájena. Recenzoval jsem rovněž původní verzi, takže se ve svém posudku nemohu vyhnout srovnávání.

Oceňuji, že práce je psaná anglicky s použitím široké odborné slovní zásoby. Závažné formální nedostatky byly odstraněny, některých formálních chyb jsem si však všiml i v opravené verzi. Kromě několika překlepů (napr. „historcal“ na str. 1, „partioning“ na str. 9), to ještě je nekonzistentní odkazování k poznámce pod čárou na konci věty – zatímco na str. 9 a 18 se odkaz nacháazí správně až za tečkou, na str. 8 a 27 je to před tečkou. V první větě v kapitole 2.3.1 chybí sloveso. Atributy v příkladech na str. 33 jsou prohozeny. Navíc číslovaní kapitoly by zřejmě nemělo začínat nulou, i když se jedná „jenom“ o úvod.

Po obsahové stránce je text do jednotlivých kapitol strukturován logicky, kapitoly tvoří samonosné celky a vlastní práce autora je jasně odlišena.

Podobně jako v původní verzi, rovněž v té současné je zpracování kapitoly 2 na kvalitní úrovni. Sekce 2.4.1 popisující evaluaci metody Latent Dirichlet Allocation (LDA) by si však zasloužila více prostoru. Pochopení algoritmu je náročné obzvlášť pro ty, kteří nemají v oblibě pseudokód, ve kterém se navíc používají nikde nevysvětlené konstanty m a R.

V kapitole 3 se nachází krátká sekce věnována PDT 3.0. Vzhledem k tomu, že dat z tohoto korpusu není nikde využíváno, tuto část by bylo lepší vypustit nebo přesunout její zkrácenou verzi do poznámky pod čarou.

V kapitole 4 je dle mého názoru nejsilnější část práce, a to konkrétně v její druhé části obsahující analýzy aktivovanosti v textu na základě navržené statistiky LeapHeight. Naopak, některé obecné statistiky v její první části s chybějícím odkazováním na ně v dalším textu působí trochu samoúčelně. Autor po připomínce doplnil na přiloženém CD grafy aktivovanosti pro 9 ukázkových dokumentů. Z nich je vidět, že zvláště pro dokumenty delší než 30 vět se grafy stávají nepřehlednými. Autor mohl využít reimplementace vizualizace a možnosti vykreslit grafy i pro dlouhé dokumenty a navrhnout přehlednější způsob zobrazení aktivovanosti. Na str. 35 by k pochopení rozdílného použití demonstrativ a osobních zájmen pomohl ilustrační příklad z dat.

Kapitola 5 obsahující experimenty, která by měla být dle mého názoru pro studenta počítačové lingvistiky klíčová, na mně stále působí rozpačitě. Autor přidal jeden experiment s modelem pro LDA kombinujícím frekvence slov s hodnotou průměrné aktivovanosti a zjistil, že kvalita tohoto kombinovaného postupu je sice lepší než s použitím pouhé průměrné aktivovanosti, ale horší než se standardním modelem založeném na četnosti slov. Analýzy, proč je kombinovaný postup stále horší než standardní, se čtenářovi dostává jen v jediném odstavci v podobě úvah nepodložených žádnou další analýzou např. vzniklých tématických shluků nebo mírně modifikovanými experimenty, které by mohly odhalit víc. Rovněž není vysvětleno, proč se na adaptaci průměrné aktivovanosti u jednoho experimentu používá hodnota 100 a u druhého počet vět v dokumentu. Jistě by bylo taky zajímavé reflektovat dotaz, který padl při obhajobě, jestli statistika LeapHeight nekoreluje značně s větnou vzdáleností mezi za sebou

následujícími vyjádřeními konkrétní entity, a tudíž by nestačilo použít tyto hodnoty místo hodnot LeapHeight, které závisí ještě na aktuálním členění.
Z programátorského hlediska řešitel svou úlohu splnil, zručně kombinuje několik programovacích jazyků se skriptovacími nástroji (btred, xsh).

Závěr

Autor reflektoval nedostatky, které byly vytýkány původní verzi jeho diplomové práce. Avšak tato vylepšení přidal jenom v minimální postačující míře, jak mu bylo navrženo. Nezdá se mi, že by do práce zahrnul nějaký svůj další nápad. Přesto tato práce podmínky splňuje a doporučuji ji k obhajobě.

Doporučení k obhajobě:

Z výše uvedených důvodů práci *doporučuji* k obhajobě.

Vynikající práce vhodná pro soutěž studentských prací	ANO ☐

Seznam soutěží studentských prací, viz http://www.mff.cuni.cz/studium/bcmgr/prvzoryace/

Pokud jste výše zaškrtli ANO, zdůvodněte prosím svůj návrh, případně uveďte konkrétní soutěž, pro kterou je práce vhodná (rámeček lze nechat prázdný, pokud za dostatečné zdůvodnění považujete text posudku):

V Praze dne: 2.9.2014

Podpis: