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 dat 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 45 stran včetně seznamu literatury, obrázků a tabulek. K práci je přiložen CD-ROM se zdrojovými kódy a daty.

Hodnocení

Oceňuji, že práce je psaná anglicky s použitím široké odborné slovní zásoby. Mnoho jazykových obratů je však příliš komplikovaných, což stěžuje pochopení textu. Neměl by být na str. 4 místo „arguing with Hajičová" použit spíše výraz „agreeing"?

V některých případech porozumění komplikuje i nejednotná terminologie. Např. koreferenční řetězce jsou na str. 5 zavedené termínem „coreference chain", zatímco na str. 26 autor používá výraz „coreference string", což je při prvním čtení i trochu zavádějící.

Věta „Natural languages use various surface means..." na str. 7-8, která je dlouhá 62 slov, je doslovně zkopírovaná z (Hajičová, 1993)[1], dokonce i s chybou způsobenou systémem na optické rozpoznávaní znaků. Zdroj sice uveden je, ale táto věta není uzavřená v uvozovkách. Autor tak sice přiznává, že myšlenka není jeho, ale slovní formulaci už za svou vydává. Pravděpodobně ne úmyslně, ale autor se tak dopustil plagiátorství.

Mnoho odkazů na sekce, obrázky a tabulky je špatných (str. 21, 29-30), v seznamu literatury často chybějí názvy sborníků a vydavatelů. Práci by určitě pomohla dodatečná pečlivá kontrola napsaného textu.

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

Kapitola 1 s nejasným názvem „Research" obsahuje sekci „Related Work", která je však jen těžko pochopitelná bez podrobnějšího teoretického úvodu do úlohy, kterou autor řeší (v textu označované jako „the analysis of a discourse structure with its dynamic development in relation to a sentence structure"). Této sekci chybí i pohled na předchozí práce v kontrastu s přístupem popsaným v diplomové práci.

Kapitola 2 popisující teoretické koncepty je zpracována pěkně, až na sekci popisující metodu „Latent Dirichlet Allocation", kde se objevuje několik chyb a nejasností v matematických formulích (např. není jasné, co vyjadřuje proměnná *n,* přes kterou je zřejmě definován produkt ve formuli 2.1 na str. 12).

Vlastní práci jsou věnované kapitoly 4 a 5. Kapitola 4 je zpracována nejlépe. Zavádí se v ní nápaditá statistika LeapHeight, která je následně vykreslená v grafech a jejich pečlivá analýza přináší zajímavé poznatky a potvrzuje některé lingvistické intuice.

Kapitola 5 měla být podle mého názoru klíčovou částí diplomové práce. Míra aktivovanosti objektů je tady použitá při úlohách vizuálního shlukování dokumentů a modelovaní témat pomocí LDA. V případě první úlohy (str. 33) není jasné, jak je definován překryv párů řetězců. Je to podobně jako v druhé úloze, tedy když jsou lemmata prvních výrazů z řetězců stejná? Výsledkem shlukování dokumentů jsou grafy shluků, které však bohužel vůbec nejsou blíže popsány ani analyzovány. Výsledkem druhé úlohy, modelovaní témat, je výrazně nižší úspěšnost testovaného systému oproti baseline systému, což nepovažuji za problém. Problémem ale je, že tento výsledek je vysvětlen strohým tvrzením o zásadní chybě v nastavení experimentu, které však není ničím podpořené ani rozvinuté. Cílem

1 http://www.coli.uni-saarland.de/~korbay/Courses/esslli04/Scans/hajicova93.pdf

jakékoliv vědecké práce by mělo být zkoumaný jev nejen popsat, ale pokusit se i o fakty podložené vysvětlení. Jeho absence při jediných dvou experimentech považuji za největší nedostatek této práce.

Z programátorského hlediska řešitel svou úlohu splnil, zručně kombinuje několik programovacích jazyků se skriptovacími nástroji (btred, xsh). I když ze zdrojových kódů a dat obsažených na CD-ROM-u je možné si grafy aktivovanosti vygenerovat, bylo by praktičtější, kdyby byly alespoň pro několik ukázkových dokumentů přiloženy.

Závěr

Část cílů řešitel splnil. Ten z mého pohledu nejzajímavější a nejpodstatnější – použití v experimentech a jejich analýza – však splněn nebyl. Navíc práce obsahuje mnoho formálních nedostatků a jeví známky, že byla šita horkou jehlou.

Proto diplomovou práci v této podobě nemohu doporučit k obhajobě. Věřím, že řešitel dokáže zpracovat stanovené téma lépe.

Doporučení k obhajobě:

Z výše uvedených důvodů práci *nedoporučuji* 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: 19. 5. 2014

Podpis: