

Oponentský posudek diplomové práce

Autor práce: Bc. Jan Václ

Název práce: Tracing Saliency in Texts

Oponent: doc. Ing. Zdeněk Žabokrtský, Ph.D.

Popis práce

Cílem předložené diplomové práce bylo seznámit se s pojmem aktivovanosti objektů v textu, vizualizovat a interpretovat grafy aktivovanosti, použít strojové učení k modelování aktivovanosti a využít aktivovanost v další úloze v rámci NLP.

Práce je strukturovaná následovně. Po kapitole, která vymezuje cíle, následuje druhá kapitola s popisem související literatury. Třetí kapitola popisuje použitá data a nástroje. Čtvrtá kapitola analyzuje tato data. V páté kapitole jsou použity rozhodovací stromy k predikci aktivovanosti. Šestá kapitola experimentuje se využitím aktivovanosti pro shlukovou analýzu dokumentů. Sedmá kapitola je závěrečná.

Práce je psána anglicky, včetně rejstříků a příloh má 69 stran.

Hodnocení

Po formální stránce je práce zpracována kvalitně, je napsaná dobrou angličtinou s minimem jazykových chyb, má standardní strukturu a kvalitní sazbu.

Požadavek na implementaci nástroje pro vizualizaci grafů aktivovanosti byl splněn. Pokud jde o provedené experimenty s predikovaním aktivovanosti, evaluační míry pro regresní i klasifikační formulaci problému jsou patřičně popsány a experiment je vyhodnocen korektně. Druhý hlavní experiment, tj. shluková analýza dokumentů, je vyhodnocen pouze velmi předběžně a subjektivně, nicméně zadání nebylo v tomto bodě nijak specifické. Vráťím se tedy k predikování aktivovanosti. Mohl jsem něco přehlédnout, ale musím říci, že mi poněkud uniká cíl experimentu. Aktivovanost v pojetí, které převzal autor, je číselné ohodnocení výrazu v textu, které je (při dostupné tektogramatické anotaci) spočítáno zcela mechanicky pomocí dříve publikovaného algoritmu, popřípadě s využitím autorovy úpravy nazývané „saliency leap height“. V zásadě jde jen o lineární kombinaci vzdálenosti předchozí totožné reference a kontextové zapojenosti aktuálního výrazu. Nejde tedy o veličinu, kterou by bylo jinak nutné nějak jinak netriviálně „měřit“ (např. ruční anotací), není mi proto jasné, jaký má smysl využívat strojové učení k reprodukci v zásadě velmi jednoduchého výpočtu.

Práci bych dále vytknul příliš stručnou rešeršní kapitolu. Z oblastí mimo zvolený teoretický rámec je jen povrchně zmíněno několik převážně starších prací. Moderní revidování pojmu saliency je přitom bez znalosti novějších psycholingvistických a/nebo aplikačních experimentů obtížně představitelné.

Závěr

Přes uvedené výhrady doporučuji předloženou práci k obhajobě.

V Jahodově, 30. srpna 2015

doc. Ing. Zdeněk Žabokrtský, Ph.D.

Ústav formální a aplikované lingvistiky

MFF UK, Praha