Posudek bakalářské práce

předložené na Matematicko-fyzikální fakultě
Univerzity Karlovy v Praze

posudek oponenta

Autor: Jan Hajič

Název práce: Popularita osob automaticky

Studijní program a obor: Informatika,

Rok odevzdání: 2011

Jméno a tituly vedoucího/oponenta: Mgr. Martin Popel

Pracoviště: ÚFAL MFF UK

	excelentní	**odpovídající**	**slabší**	**nevyhovující**
Náročnost zadaného tématu	X			
Míra splnění zadání	X	X		
Rozsah práce	X	X		
Struktura textové části práce	X			
Analýza	X			
Vývojová dokumentace		X		
Uživatelská dokumentace		X		
Jazyková a typografická úroveň	X	X		
Návrh a design implementace		X	X	
Kvalita zpracování softwarové části		X	X	
Stabilita aplikace		X		

Nejvýznamnější klady:

- vytvoření korpusu s ručně označkovanými polárními výrazy
- podrobná analýza anotace a anotačních neshod
- provedení a vyhodnocení experimentů (dva korpusy, mnoho porovnávaných metod)
- interpretace výsledků a navržené možnosti zlepšení
- precizní zavedení používaných lingvistických termínů i matematických vzorců
- přehledná struktura textu a typografická úprava

Nejzávažnější nedostatky:

Práce má charakter pilotní studie s podrobnou analýzou problematiky. V této oblasti žádné závažné nedostatky neshledávám, některé připomínky uvádím v dalších poznámkách. Nedostatky ve zpracování softwarové části nejsou závažné vzhledem k charakteru práce, kdy softwarová část není určena pro uživatele v klasickém smyslu, ale spíše slouží k vyhodnocení experimentů, přesto alespoň dva uvedu:

- Softwarovou část vystihl sám autor v úvodu technické dokumentace. Nepřenositelnost na jiný systém (způsobená `#!/bin/perl` a `use lib '/home...'`) je sice snadno opravitelná (jak se uvádí v části 2.2.1 uživatelské dokumentace), ale zbytečná při dodržení základních postupů běžných pro jazyk Perl (packages a `#!/usr/bin/env perl` případně `perl script.pl` místo `./script.pl`).
- Zdrojový kód je okomentovaný (a popsaný v technické dokumentaci), ale chybí POD dokumentace a např. `perl classify_articles.pl --help` vypíše chybovou hlášku Illegal division by zero at artlib.pl line 497.

- Nepoužití metody LSA (která byla zmíněna v zadání) je v práci odůvodněno malým rozsahem dostupných dat a lexikonu, ale není uvedeno, zda a jaké experimenty s LSA byly provedeny (byť s neuspokojivými výsledky).
- Práce je psána kvalitní češtinou, ale s občasnými překlepy.

Další poznámky:

- Autor sám v textu reflektuje moji hlavní koncepční připomínku ke zřejmě nevhodné volbě „perspektivy čtenáře“ (což je záměr příliš ambiciózní, těžko realizovatelný a možná i špatně definovatelný). Naopak se mi zdá, že identifikování mluvčích jednotlivých polárních stavů (bez ohledu na to, jak to působí na čtenáře) by mohlo být užitečnou aplikací s širším využitím.
- Doporučoval bych anotovat celé polární stavy (tedy trojice mluvčí, polární prvek, cíl), nikoli jen cíle. To by zvýšilo další využitelnost takto anotovaných dat.
- Je pro anotaci jedné věty potřeba i kontext okolních vět? Ztrácí se tento kontext zvoleným dělením na train a eval data? Pokud ano, tak je dané rozdělení nevhodné pro pozdější porovnání s metodami, které by uměly okolní kontext využít.
- Není jasné, co je jednotkou anotace. Na str. 8 se např. uvádí „větě přisuzovali polaritu“, ale jedna věta může obsahovat více polárních stavů (dokonce různě orientovaných). Zjednodušení „věta = 1 polární stav“ je popsána v úvodu kapitoly 4, ale bez prozkoumání dat na CD není zcela zřejmé, zda bylo toto zjednodušení uplatněno i při anotaci.
- Při prvním čtení není zcela zřejmé, jaký je rozdíl mezi anotováním všech cílů a anotováním všech polárních prvků.

	výborně	**velmi dobře**	**dobře**	**neprospěl/a**
Návrh známky	**X**			

Datum: 25. 8. 2011

Podpis: