Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy v Praze

Autor práce	Petr Kubát
Název práce	Normalizace pojmenovaných entit v českých textech
Rok odevzdání	2014
Studijní program	Informatika
Studijní obor	Obecná informatika

Autor posudku	Mgr. Martin Popel	Oponent
Pracoviště	ÚFAL MFF	

Prosím vyplňte hodnocení křížkem u každého kritéria. Hodnocení *OK* označuje práci, která kritérium vhodným způsobem splňuje. Hodnocení *lepší* a *horší* označují splnění nad a pod rámec obvyklý pro bakalářskou práci, hodnocení *nevyhovuje* označuje práci, která by neměla být obhájena. Hodnocení v případě potřeby doplňte komentářem. Komentář prosím doplňte všude, kde je hodnocení jiné než *OK*.

K celé práci	lepší	OK	horší	nevyhovuje
Obtížnost zadání		X		
Splnění zadání		X		
Rozsah práce *... textová i implementační část, zohlednění náročnosti*	X			
Komentář Náročnost zadání považuji za přiměřenou: na vstupu je text, ve kterém jsou označeny pojmenované entity (obecně to mohou být ale jakékoli jmenné fráze), např. „*Jedeme do <ne>Ústí nad <ne>Labem</ne></ne>*.“ Výstupem je vyplnění normalizovaných názvů pro každou entitu, např. „*Jedeme do <ne normalized_name='Ústí nad Labem'>Ústí nad <ne normalized_name='Labe'>Labem</ne></ne>*.“ Samotné rozpoznání pojmenovaných entit tedy není součástí zadání. V řešení bylo využito nástrojů MorphoDiTa a Treex (byť Treex je použit jen jako externí nástroj, se kterým se komunikuje přes soubor, a normalizace tedy nebyla zabudována do Treexu tak, aby ji mohly využít další komponenty Treexu). Zadání se podařilo splnit, a to včetně přehledného webového rozhraní. Text práce podrobně seznamuje čtenáře s celým procesem normalizace. Nechybí ani dvě úvodní kapitoly o počítačovém zpracování češtiny a o pojmenovaných entitách. Oceňuji průběžné diskuse jednotlivých rozhodnutí.				

Textová část práce	lepší	OK	horší	nevyhovuje
Formální úprava *... jazyková úroveň, typografická úroveň, citace*		X		
Struktura textu *... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*	X			
Analýza		X		
Vývojová dokumentace		X		
Uživatelská dokumentace		X		

Komentář

Práce je psána srozumitelně, téměř bez překlepů a je vhodně strukturována (s využitím příloh, seznamu zkratek atd.). Formální úpravu kazí porušování typografických konvencí (zejména používání spojovníku místo pomlčky, neslabičné předložky na koncích řádků).

Je škoda, že nebylo vyhodnoceno baseline řešení, které by nechalo enity ve tvarech, v jakých se vyskytly v textu. Toto vyhodnocení by pomohlo v interpretaci úspěšnosti naměřené na čtyřech testovacích sadách.

Implementační část práce	lepší	OK	horší	nevyhovuje
Kvalita návrhu *... architektura, struktury a algoritmy, použité technologie*		X		
Kvalita zpracování *... jmenné konvence, formátování, komentáře, testování*		X		
Stabilita implementace		X		

Komentář

Návrh popsaný v kapitolách 3.6 – 3.7 považuji za kvalitní. Zvolené řešení není algoritmicky náročné, což neberu jako nevýhodu, neboť dosahuje uspokojivých výsledků. Z rozboru výsledků navíc vyplývá, že značná část chyb je způsobena spíše použitým taggerem a parserem (které ovšem patří mezi ty nejlepší dostupné pro češtinu), tedy by další vylepšování a komplikování postupu vlastní normalizace nebylo příliš vhodné.

Při testování nástroje jsem se nesetkal s problémy se stabilitou. Drobný nedostatek vidím v tom, že nástroj při chybném použití (nevyplnění druhého parametru se jménem výstupního souboru) ohlásí chybu až na konci zpracování (tj. cca po minutě).

Celkové hodnocení Výborně

Práci navrhuji na zvláštní ocenění Ne

Datum 3. 6. 2014 Podpis