Posudek bakalářské práce

předložené na Matematicko-fyzikální fakultě
Univerzity Karlovy v Praze

posudek vedoucího ✔ posudek oponenta

Autor/ka: Adam Liška

Název práce: Čištění paralelních dat pro strojový překlad

Studijní program a obor: Informatika, obecná informatika

Rok odevzdání: 2010

Jméno a tituly vedoucího/oponenta: RNDr. Daniel Zeman, Ph.D.

Pracoviště: ÚFAL MFF UK

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

Posudek:
Tato bakalářská práce je převážně experimentálního charakteru a studuje možnosti automatického filtrování paralelních trénovacích dat pro statistický strojový překlad. Jako trénovací data slouží rozsáhlé kolekce dvojjazyčných textů, které však pocházejí z různě spolehlivých zdrojů (včetně webu) a ne vždy je jeden zcela věrným překladem druhého. Stává se proto, že data obsahují větné páry, které by bylo lepší vyřadit, vzhledem k rozsahu dat však není možné je zkontrolovat ručně. To je hlavní motivací této práce, která navrhuje a vyhodnocuje automatické filtry pro vyhledání a vyřazení nevhodných větných párů.

Autor pracuje s česko-anglickým korpusem CzEng, všechny navržené filtry až na jednu výjimku jsou však jazykově nezávislé. Práce shrnuje počáteční stav včetně některých již existujících filtrů, které byly s korpusem CzEng použity dříve. Těžištěm práce je potom návrh nových filtrů a vyhodnocení nových i starých filtrů.

Práce má i svou implementační část (zejména implementace filtrů).

Práci hodnotím kladně po experimentální i implementační stránce. Jazykově je práce také v pořádku, až na několik drobných překlepů.

Nejvýznamnější klady:
Práce popisuje několik možností, jak automaticky odhadnout, že daný větný pár není vhodný do paralelního korpusu, protože cílová věta není dost dobrým překladem zdrojové. V rámci práce byl také vzorek dat ručně anotován co do (ne)vhodnosti větných párů a navržené filtry byly na těchto datech otestovány. Díky jazykové neutralitě většiny filtrů lze výsledky práce využít i při přípravě paralelních dat pro jiné páry jazyků v budoucnosti.

Nejzávažnější nedostatky:
Žádné závažné, ale viz též další poznámky:

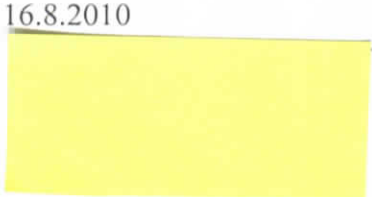
Další poznámky:

- Str. 5: „překladu pracujícího s jedním či dvěma jazyky" – jak pracuje překlad s jedním jazykem?
- Str. 11: Číselný filtr: Existují korpusy (např. Emille), ve kterých se čísla systematicky liší, např. jazyk 1: „podrobnosti najdete na straně 11" – jazyk 2: „podrobnosti najdete na straně 22".
- Str. 18: Text by byl přehlednější, kdyby obecný výklad, jak filtr pracuje, byl oddělen od technických detailů o tom, jak ho pouštět.
- Str. 20: „Jejich případné odstranění ovšem opět strojový překlad neovlivní, jelikož většina systémů neznámá slova na vstupu použije na výstupu v nezměněné podobě." – To je sice pravda, ale může to ovlivnit zarovnání těchto slov v delších větách.
- Str. 21: Objevuje se termín „ROC křivky", aniž by byl podrobněji vysvětlen.
- Str. 22: Obrázek 4.1: Popis obrázku by neškodil trochu explicitnější. Na první pohled není zrovna jasné, že ona „křivka" je tučná část toho obdélníku, který vypadá jako okraj grafu.
- Vyhodnocení a závěr: Neškodilo by upřesnit pravidla, podle kterých anotátor rozhodoval, jak se pozná, že pár je špatný. Částečná korespondence už je špatně?
- Je možné vyhodnotit dopad na strojový překlad pomocí BLEU skóre nebo jiné metriky?

Závěr:
Práce je vhodně zpracovaná, přináší užitečné výsledky a považuji ji za odpovídající požadované úrovni bakalářských prací. Doporučuji ji ke schválení.

	výborně	velmi dobře	dobře	neprospěl/a
Návrh známky	×			

Datum: 16.8.2010

Podpis: 

Poučení k formuláři pro hodnocení informatických bakalářských prací

Tento formulář je určen pro hodnocení vedoucího i oponenta bakalářské práce, která má formu softwarového projektu. Bakalářské projekty jiných typů (teoretická práce, srovnávací studie apod.) budou hodnoceny pomocí standardních textových posudků.

Jednotlivá políčka vyplňte nejlépe elektronicky (lze případně i ručně), je možné zaškrtnout i dvě sousední políčka (např. pro hodnocení typu 'něco mezi odpovídající a slabší'), a to i u návrhu výsledné známky. Pokud některá položka nemá vzhledem k práci smysl (např. stabilita aplikace u práce bez vlastní implementace), položku nevyplňujte. Výsledná navrhovaná známka nemusí být žádným 'průměrem' hodnocení jednotlivých kritérií. Pokud některé položky hodnotíte jako slabší nebo nevyhovující, v sekci Nejzávažnější nedostatky popište důvody vašeho hodnocení a zjištěné nedostatky.

Výklad stupňů hodnocení:

- excelentní znatelně lepší/rozsáhlejší/dokonalejší než je pro Bc práci požadováno
- odpovídající přiměřené Bc práci, student splnil to, co měl
- slabší výhrady ke kvalitě, rozsahu, hloubce nebo zpracování
- nevyhovující neodpovídá požadavkům na Bc práci, práce nemá být obhájena

Vyplněné a ručně podepsané (i v případě elektronického vyplňování) hodnocení odevzdejte na sekretariát KSI, elektronickou verzi pošlete na sekretariat@ksi.ms.mff.cuni.cz. Pokud máte emailový kontakt na autora práce, pošlete posudek i jemu.