Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Gabriela Závodská
Název práce	Analyzing differences in alternative translations
Rok odevzdání	2024
Studijní program	Informatika
Specializace	Obecná informatika
Autor posudku	Mgr. Martin Popel, Ph.D. Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání		X		
Splnění zadání		X		
Rozsah práce *... textová i implementační část, zohlednění náročnosti*		X	X	

Cílem práce bylo vyvinout nástroj pro analýzu rozdílů v alternativních překladech téhož textu a použít ho pro porovnání tzv. *optimálních* a *standardních* referenčních (tedy lidských) překladů (případně i strojových překladů) z angličtiny do češtiny. Tento cíl byl splněn. Nástroj počítá metriky několik typů, jak bylo požadováno v zadání.

Implementačně je nástroj spíše jednodušší: Chybí interaktivní grafické uživatelské rozhraní, které ale nebylo v zadání požadováno. Většina metrik je založena na jednoduchém počítání anotací získaných nástrojem UDPipe; jednou z mála výjimek je metrika detekující nominalizace dle databáze DeriNet. Jediná metrika využívající anglického originálu je Untranslated počítající počet nepřeložených slov.

Nižší náročnost implementační části je vyvážena rozsáhlejším rozborem rozdílů optimálních a standardních referenčních překladů popsaným v kapitole 4, která zabírá více než polovinu stran vlastního textu práce. Podařilo se najít mnoho rozdílů mezi oběma překlady; některé potvrdily pozorování popsaná v literatuře (Kloudová et al., 2021; Kloudová et al., 2023b), některé byly nové. Oceňuji snahu o lingvistickou interpretaci pozorovaných rozdílů, dokládanou na příkladových větách (byť jen u některých jevů; třeba rozdíly v TTR vysvětleny nejsou). Tato interpretace by z hlediska translatologického často neobstála,[1] ale na informatickou bakalářskou práci mi přijde více než dostatečná a vhodně ukazující potenciál popsaného nástroje pro budoucí výzkum (ideálně již v týmu s translatologickou odborností).

[1] Například mi přijde nepřesvědčivá hypotéza, že N1 má mírně nižší TTR než N2 kvůli vyšší podobnosti s anglickým originálem, který má též nižší TTR. Angličtina má ale TTR nižší ze zcela jiných důvodů (zejména kvůli členům a předložkám nahrazujících např. český 7. pád), které se nepřenášejí ani do doslovných překladů.

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	X	
Analýza		X		
Vývojová dokumentace		X		
Uživatelská dokumentace		X		

Z práce je vidět, že většina částí byla psána na poslední chvíli, což se projevuje zejména na formální úpravě, jazykové úrovni, množství překlepů a drobných nepřesností, které ale naštěstí většinou nebrání porozumění.[2]

Sekce 3.1 sice zmiňuje, že u features v tabulce 4.5 se dělí počtem slov s danou feature, (na rozdíl od word-based metrics v tabulce 4.3, kde se dělí počtem všech slov), ale chybí vysvětlení motivace pro toto rozhodnutí.

Oceňuji naopak vlastní iniciativu autorky, když navrhla počítání korelací prezentovaných v tabulce 4.4. Pozitivní korelace jsou většinou nepřekvapivé, ale negativní korelace mohou odhalit, že nějaký jev v jednom překladu (např. vyšší četnost plnovýznamových sloves) je nahrazen jiným jevem v druhém překladu (např. vyšší četnost pomocných sloves a spon).

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		

Celkové hodnocení Velmi dobře, spíš lepší

Práci navrhuji na zvláštní ocenění Ne

Datum 26. 8. 2024 Podpis

[2] Výjimkou, která porozumění může komplikovat, je tvrzení „None of the metrics take into consideration the English column“, přičemž metrika Untranslated (jako jediná) anglický originál v úvahu bere.
Zapeklitým případem je tvrzení „words *uveden**ý*** and *released* are participles“, obsahující překlep, protože původní věta obsahuje krátký/jmenný tvar *uveden*. Oba tvary jsou sice v UD (na rozdíl od PDT) anotovány jako přídavná jména (ADJ), ale liší se deprel slova *bude*: `cop` vs. `aux:pass`.