

POSUDEK BAKALÁŘSKÉ PRÁCE

Název práce: Hodnocení automatického překladu SMT systémů (Google translate a Bing) z francouzštiny do češtiny: kolokace z oblasti odpadového hospodářství

Diplomantka: Barbora Víšková

Pracoviště: Ústav romanistiky FF JU v Českých Budějovicích

Vedoucí práce: doc. PhDr. Jan Radimský, Ph.D.

Rozsah: 119 stran vč. bibliografie a vlastních příloh

Cílem předkládané bakalářské práce je analyzovat a vyhodnotit správnost překladu 223 víceslovných odborných termínů – kolokací z oblasti „odpadového hospodářství“, který byl pořízen pomocí nástrojů automatického překladu Google translate a Bing.

V úvodní části práce autorka stručně představila principy strojového překladu v historické perspektivě. Dále podrobně shrnula několik předchozích studií o evaluaci úspěšnosti automatického překladu ze zahraniční literatury, tyto evaluace byly provedeny pro následující jazyky: afrikánština a angličtina (s. 20-25), perština a angličtina (s. 25-26), hindština a angličtina (s. 27-29), a konečně angličtina s arabštinou (s. 30-31). Všimla si přitom jednotlivých typologií chyb, porovnávala je a konstatovala, že jsou vždy do jisté míry arbitrární. Testované kolokace pro vlastní analýzu čerpala z BP Lucie Papouškové (2007); kolokace zadávala do překladačů v nekontextualizované podobě a výsledné překlady hodnotila s ohledem na typologii chyb. Při finálním porovnání výsledků obou překladačů došla k závěru, že Bing (jakožto hybridní systém) lépe kontroluje gramatické a syntaktické chyby na výstupu, zatímco Google translate (který v rámci strojového učení pravděpodobně čerpá z obsáhlejší trénovací databáze paralelních textů) dosahuje lepších výsledků v terminologických a lexikálních volbách.

Práce je velmi obsáhlá, autorka čerpala z řady zahraničních studií a samostatně provedla poměrně rozsáhlý materiálový výzkum. Po metodologické stránce lze jistě o řadě věcí vést diskusi, avšak řada problematických otázek je dána spíše náročnou povahou tématu. Principiálně je totiž problematické aplikovat na výstup statistického strojového překladu hodnotící škály, které předpokládají lingvistické (tj. syntaktické, morfologické, sémantické...) parametry, neboť strojový překlad takto nefunguje. Na druhou stranu je však výstup SMT systémů určen lidskému uživateli, který právě tyto parametry nutně (byť neuvědoměle) při interpretaci sleduje a analyzuje. Jiný postup tedy není dost dobře možný, jak ukazuje i sekundární literatura (odhlédneme-li od pokusů, které spočívaly např. v evaluaci formou počítání minimálního množství „změn“, které je nutno s přeloženým textem provést, aby byl přijatelný). V diskusi bych se tedy omezil na několik dílčích otázek:

a) Proč zadávala autorka kolokace nekontextualizované, resp. co by se kontextualizací mohlo změnit?

b) Na s. 24 v oddílu „změna výrazu, chybná syntax“ píše o „elementech textu, které odkazují na jiná slova či věty“. Jedná se o případy anafory, tedy o problém známý z počítačové lingvistiky jako „anaphora resolution“? Mohla by uvést příklady, které v daném sekundárním prameni tento jev exemplifikují?

c) Na s. 49 autorka píše, že „odborné výrazy jsou pak přeloženy buď pomocí opisů, či za použití hyperonym (...) anebo hyponym“. Myslí tím, že Google translate provádí tento typ analýzy?

d) Na s. 58-59 analyzuje autorka kolokaci „déchets lingotés en fer ou en acier“ a tvrdí, že Google translate nenašel překlad pro „lingoté“ a proto daný výraz přeložil, jako by vstupem bylo formálně podobné slovo „ligoté“. Měl tak vzniknout ekvivalent „vázaný odpad ze železa nebo oceli“ (vázaný = *ligoté*). Znamená to, že GT svévolně zaměnil formu ve zdrojovém textu? Když jsem sám zkoušel daný postup opakovat, vyšel mi překlad „přetavený odpad v ingotech ze železa nebo oceli“, kde je klíčové slovo *lingoté* převedeno lexikálně korektně slovem *ingotech*. Jen v našeptávači („měli jste na mysli“) GT sugeruje možnou výchozí formu *déchets ligotés en fer ou en acier*. Samovolná záměna formy ve výchozím textu by byla zásadním parametrem, který ovlivňuje činnost SMT systému.

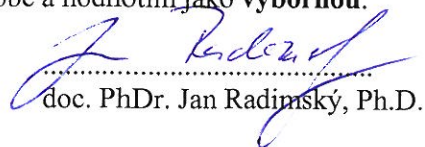
e) Na s. 62 – není překlad *sucrerie* jako *cukroví* (místo náležitého *cukrovarnictví*) spíše problémem dezambiguace než problémem terminologické volby?

Navzdory těmto diskusním připomínkám myslím, že autorka odvedla velký kus poctivé práce a výsledná BP je velmi kvalitní.

Po formální stránce práce odpovídá požadavkům.

Závěr:

Předkládanou bakalářskou práci doporučuji k obhajobě a hodnotím jako **výbornou**.


.....
doc. PhDr. Jan Radimský, Ph.D.

V Českých Budějovicích, dne 11. ledna 2017