Agentúra na podporu výskumu a vývoja
Mýtna 23, P.O.BOX 839 04, 839 04 Bratislava 32

Formulár ZK

Záverečná karta projektu

Názov projektu

Typológia chýb strojového prekladu do slovenčiny ako flektívneho typu jazyka

Evidenčné číslo projektu **APVV-14-0336**

Zodpovedný riešiteľ **doc., RNDr. Daša Munková, PhD.**

Príjemca **Univerzita Konštantína Filozofa v Nitre**

Názov pracoviska, na ktorom bol projekt riešený

Katedra translatológie, Univerzita Konštantína Filozofa v Nitre
Katedra anglistiky a amerikanistiky, Univerzita Konštantína Filozofa v Nitre
Katedra germanistiky, Univerzita Konštantína Filozofa v Nitre
Katedra slovenského jazyka a literatúry, Univerzita Konštantína Filozofa v Nitre
Katedra informatiky, Univerzita Konštantína Filozofa v Nitre

Názov a štát zahraničného pracoviska, ktoré spolupracovalo pri riešení

Ústav systémového inženýrství a informatiky, Univerzita Pardubice, Česká republika
Generálne riaditeľstvo pre preklad (DGT) Európskej komisie v Luxemburgu, Belgicko

Udelené patenty/podané patentové prihlášky, vynálezy alebo úžitkové vzory, ktoré sú výsledkami projektu

-

Najvýznamnejšie publikácie (knihy, články, prednášky, správy a pod.) zhrňujúce výsledky projektu – uveďte aj publikácie prijaté do tlače

1. Najvýznamnejšie publikácie
A. Karentované a impaktované zahraničné časopisy:
1) Munk, Michal; Drlik, Martin; Benko, L'ubomir; Reichel, Jaroslav, (2017): Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques. IEEE ACCESS, 5, pp. 8989-9004.
[WoS, Scopus]
2017IF: 3,557; Q1; Current Contents Connect
2) Munk, Michal; Benko, L'ubomir, (2018): Using Entropy in Web Usage Data Preprocessing. ENTROPY: Special Issue on Entropy-based Data Mining, 20(1), pp. 1-15.
[WoS]
2017IF: 2,305; Q2; Current Contents Connect
3) Munk, Michal; Munkova, Dasa, (2018): Detecting errors in machine translation using residuals and metrics of automatic evaluation. JOURNAL OF INTELLIGENT & FUZZY SYSTEMS: Special Issue on Language & Knowledge Engineering, 34(5), pp. 3211-3223.
[WoS]
2017IF: 1,426; Q3; Current Contents Connect
4) Munk, Michal; Munkova, Dasa; Benko, L'ubomir, (2018): Towards the Use of Entropy as a Measure for the Reliability of Automatic MT Evaluation Metrics. JOURNAL OF INTELLIGENT & FUZZY SYSTEMS: Special Issue on Language & Knowledge Engineering,

34(5), pp. 3225-3233.
[WoS]
2017IF: 1,426; Q3; Current Contents Connect
5) Munkova, Dasa; Wrede, Olga; Absolon, Jakub, (2018): Vergleich der menschlichen, maschinellen und Post-Editing-Übersetzung aus dem Slowakischen ins Deutsche mittels Maße automatischer Evaluation. ZEITSCHRIFT FÜR SLAWISTIK, 64(1), pp. 1-31. (in press)
[WoS]
Current Contents Connect
B. Vedecké monografie vydané v zahraničí:
1) Absolon, Jakub; Munkova, Dasa; Welnitzová, Katarina, (2018): Machine Translation: Translation of the Future? Machine Translation in the Context of the Slovak Language. Praha: VERBUM, 2018, 78 pp. ISBN 978-80-87800-45-4
2) Bánik, Tomaš; Benko, Ľubomír; Machová, Renata; Munk, Michal; Munková, Dasa, (2018): Wie irrt die Maschine? Probleme maschinellen Übersetzens. Verlag: Dr. Kovač, 2018, 203 pp. ISBN 978-3-339-10626-1
C. Vedecké monografie vydané v domácom vydavateľstve:
1) Munková, Daša; Munk, Michal, (2016): Evalvácia strojového prekladu. Nitra : UKF, 2016. 173 s. ISBN 978-80-558-1116-1
2) Munková, Daša; Vaňko, Juraj; Absolon, Jakub; Bánik, Tomáš; Benko, Ľubomír; Glovňa, Juraj; Machová, Renáta; Munk, Michal; Petráš, Patrik; Welnitzová, Katarína, (2017): Mýliť sa je ľudské (aj strojové). Analýza chýb strojového prekladu do slovenčiny. Nitra: UKF, s. 260. ISBN 978-80-558-1255-7
3) Absolon, Jakub, (2018): The Need for Competency-based Selection and Training of Post-editors. Nitra : ASAP-translation.com, 2018, 26 pp. ISBN 978-80-89970-00-1
D. Domáce a zahraničné časopisy evidované v databáze Scopus alebo WoS:
1) Vaňko, Juraj, (2017): Morfologické a syntaktické kategórie - väzby a súvislosti. In: SLOVENSKÁ REČ, 82(1), s. 52-64.
[Scopus]
2) Petráš, Patrik, (2017): Plurálové podoby s cudzou gramatickou koncovkou -s v slovenčine. In: SLOVENSKÁ REČ, 82(3), s. 282-295.
[Scopus]
E. Zborníky a série vydané vydavateľstvom Elsevier/Springer/Wolters Kluwer/IEEE a evidované v databáze WoS alebo Scopus:
1) Munk, Michal; Munková, Daša; Benko, Ľubomír, (2016): Identification of Relevant and Redundant Automatic Metrics for MT Evaluation. In: Lecture Notes in Artificial Intelligence. Springer, 10053, 2016., pp. 141-152.
[WoS, Scopus, SpringerLink]
2) Munková, Daša; Munk, Michal, (2016): Automatic Metrics for Machine Translation Evaluation and Minority Languages. In: Lecture Notes in Electrical Engineering. Springer, 381, 2016. pp. 631-636.
[WoS, Scopus, SpringerLink]
3) Benko, Ľubomír; Munková, Daša, (2016): Application of POS Tagging in Machine Translation Evaluation. In DIVAI 2016. Wolters Kluwer, 2016, pp. 471-479.
[WoS]
4) Kasáš, Karol; Munková, Daša, (2016): Automatic Evaluation of Machine Translation Output for Slovak Language. In DIVAI 2016. Wolters Kluwer, 2016., pp. 533-540.
[WoS]
5) Munková, Daša; Kapusta, Jozef; Drlík, Martin, (2016): System for Post-Editing and Automatic Error Classification of Machine Translation. In DIVAI 2016. Wolters Kluwer, 2016, pp. 571-579.
[WoS]
6) Kapusta, Jozef; Benko, Ľubomír, (2019): Recommender System for Post-editing of Machine Translation. In: Lecture Notes in Electrical Engineering. Springer, 2019, 489, pp. 170-175. (in press)
[Scopus, SpringerLink]
7) Munkova, Dasa; Kapusta, Jozef; Munk, Michal; Reichel, Jaroslav, (2016): Evaluation of Machine Translation Output in Context of Inflectional Languages. Book Series: International Conference on Application of Information and Communication Technologies. IEEE, 2016,

pp. 85-89.
[WoS, Scopus, IEEEXplore]
2. Pozvaná prednáška "MT@Work"
Pozvaná prednáška na 4th annual MT@Work 2015 Conference on Machine Translation in Translation Practice, organizovanou European Commission's Directorate-General for Translation v Bruseli, 04.12. 2015.
(Daša Munková, Jakub Absolon)
3. Pozvaný hosť diskusnej relácie "Hosť pyramídy"
Pozvaným hosťom diskusnej relácie "Hosť pyramídy" v Slovenskom rozhlase a televízia (RTVS), na tému Strojový preklad, 11.09.2015.
(Daša Munková)

Uplatnenie výsledkov projektu

Jazyk je zrejme oveľa zložitejší systém, ako by sa zdalo nám ľuďom, ktorí ho prirodzene používame. Podstatou zložitosti jazyka sú neustále zmeny, ktorým podlieha. Gramatické pravidlá mŕtveho jazyka by bolo veľmi jednoduché osvojiť si, živý jazyk však prináša stále nové aktuálne spojenia. Preto je asi najlepšou cestou, ako naučiť počítač jazyk, zásobovať jeho „pamäť“ tými najnovšími jazykovými produktmi. Podstatou úspechu je, aby sa prekladateľ či prekladač učil rýchlejšie, ako sa mení spoločnosť a jazyk. Samozrejme, práve nové texty, ktoré prinášajú zmenu, je potrebné najčastejšie prekladať. Sú oblasti, v ktorých je internet (a teda pamäť počítača) nenahraditeľnou oporou a zas sú oblasti, ktorým stále najlepšie rozumie človek (štýl, modalita). Preto sa nám ako najprínosnejšie javí spolupráca stroja a človeka. Táto spolupráca spočíva nielen v oprave chýb, ktoré urobil stroj, ale aj v úprave vstupných textov (preeditácia). Jazyk ponúka vždy viacero možností, ako formulovať určitú myšlienku, môžeme ju sformulovať zložitejšie alebo jednoduchšie, môže byť sformulovaná podobne, ako by ju sformuloval hovoriaci v inom jazyku, alebo odlišne. Tu však opäť hrá úlohu cit človeka, ktorý ovláda systém obidvoch jazykov.
Vychádzajúc z vysokej závažnosti chýb zo súvetnej syntaxe konštatujeme, že ďalším faktorom úspechu pri preklade je pohľad na vetu ako celok. Prekladač musí dešifrovať celú vetu s jej gramatikou aj lexikou a potom vyprodukovať preklad. Neprekladať lineárne - slovo po slove. Mať prehľad v prekladanom texte znamená v reči počítačov porovnávať ju s existujúcimi vetami a zmeniť len plnovýznamové slová. Funkcie predložiek a spojok je možné počítač pomerne ľahko naučiť. Dôležité je pokročiť aj k zložitejším súvetiam.
Nevyrovnaná kvalita prekladov aj štatistická analýza chýb ukazuje, že veľkým kameňom úrazu je lexika, najmä špeciálna lexika (terminológia). Pretože vety obsahujúce bežné frekventované výrazy sú prekladané správne, zatiaľ čo neprítomnosť lexikálnych jednotiek (najmä napríklad nemeckých kompozít) v databáze spôsobuje vážne chyby v preklade aj v gramatickej rovine. Preto zlepšovaním paralelných databáz je možné výrazne zlepšiť výsledky strojového prekladu.
Vhodným výberom jednoduchších a jednoznačných vetných konštrukcií, ktoré sú ekvivalentné v oboch jazykoch. Tento postup smeruje k niečomu ako „kontrolovaný jazyk“, ktorý sa už využíva pri prekladoch z angličtiny. Najťažšou úlohou je dodržiavanie kongruenčných kategórií a rekcie v cieľovom texte, pretože stroj neovláda jazykový systém v podobe nemenných zákonitostí, ale učí sa postupne na základe masy údajov a správnych kolokácií, ktoré má v databáze. Toto sa dá vyriešiť len zlepšením korpusu a lepším systémom prepojenia v rámci korpusov.
Vytúženou métou je to, aby strojový prekladač zvládal aj komunikatívne funkcie a modálne odtienky výpovede. Nie je to však nedosiahnuteľné. Vidíme, že modalita nie je na poprednom mieste v rámci chýb. Dokonca chyby v rámci komunikatívnych funkcií sa ukazujú ako zanedbateľné. Možno povedať, že modálnu stránku výpovede prekladač zvláda dobre, pretože je zväčša reprezentovaná jasnými formálnymi znakmi. A platí to pre strojový preklad všeobecne: To čo je v jazyku reprezentované jasne, je dobre zrozumiteľné aj pre prekladač, podobne ako pre človeka, ktorý príde do cudzej krajiny a učí sa jej jazyk.
Jedným zo zámerov nášho výskumného projektu bolo naznačiť možnosti zdokonaľovania systémov a programov SP, ktoré by sa mali dostať do centra pozornosti vývojárov a programátorov. Preto z pozície lingvistov na záver formulujeme niekoľko námetov ako výziev pre programátorov, ale aj jazykovedcov na vylepšenie alebo zdokonalenie systémov a programov SP:
1. (aj v kooperácii s jazykovedcami) vytvárať databázu dvoj-, resp. viacjazyčných termínov z

jednotlivých odborov, vrátane administratívno-právnej sféry, resp. čerpať z jestvujúcich terminologických slovníkov;
2. vytvárať databázu dvojjazyčných najfrekventovanejších ustálených slovných spojení;
3. využívať jestvujúce, resp. vytvárať nové dvojjazyčné valenčné slovníky;
4. pokúsiť sa aplikovať do systému také metódy, ktoré by umožnili správnu identifikáciu predikátu a jeho transfer do slovenčiny v korektnej osobe, čísle, čase a mennom rode (v prípade minulého času a podmieňovacieho spôsobu);
5. implementovať postupy na správnu identifikáciu analytických slovesných tvarov ako vyjadrujúcich jeden gramatický význam typu angl. was suffering (´trpel´, nie ´bol trpia´); had shouted (´vykrikoval´/´kričal´, nie ´mal kričal´); neprekladať ich po jednotlivých slovách, ale ako celok.
Analýza chýb pri strojovom preklade nemeckých a anglických publicistických, administratívno-právnych textov a textov manuálov do slovenčiny potvrdzuje, že pre skvalitnenie a zdokonalenie strojového prekladu môže byť osožná, vzájomne obohacujúca, ba dokonca nevyhnutná spolupráca jazykovedcov a počítačových expertov, najmä tvorcov programov strojových prekladačov.

Súhrn výsledkov riešenia projektu a naplnenia cieľov projektu v slovenskom jazyku (max. 20 riadkov)

V projekte sme sa zamerali na hodnotenie kvality strojového prekladu (SP) z anglického/nemeckého jazyka do slovenského metódami manuálnej evalvácie, vrátane identifikácie a klasifikácie chýb strojového prekladu. Na sledovanie dimenzií (vecnej) presnosti a (jazykovej) korektnosti bol na evalváciu textov SP vytvorený kategoriálny model. Predkladaný kategoriálny model/rámec je napasovaný na štvorčlenný model chýb MQM: 1. Language („jazyk“), t.j. gramatika, v našom ponímaní morfológia a syntax, resp. morfosyntax. V predkladanom modeli s ňou korešpondujú kategórie predikatívnosti, syntakticko-sémantickej korelatívnosti (usúvzťažnenosti), kategoriálny rámec súvetia a čiastočne aj modálnosti (napr. negácia); 2. Accuracy („presnosť“) – ide najmä o nekorektný význam v texte cieľového jazyka, vynechanie lexémy a pod. V našom modeli sú tieto prvky nekorektného transferu zastúpené v bloku s názvom „lexika“; 3. Terminology („terminológia“) – tu ide o neadekvátny transfer termínu z pôvodného jazyka; v našom modeli je problematika transferu termínov prítomná ako subkategória v rámci lexiky. Pri poslednej kategórii s názvom Style („štýl“) ide o nesúlad medzi štýlom originálneho textu a štýlom cieľového textu. Analýza chýb vzniknutých pri SP vybraných textov nám umožňuje formulovať tieto závery: 1) Vo vzťahu ku kritériu zrozumiteľnosti textu sa vyskytujú dvojaké chyby: tie, ktoré nespôsobujú nezrozumiteľnosť, a tie, ktoré spôsobujú nezrozumiteľnosť textovej jednotky a celého textu. 2) Pre všetky kategórie chýb bola identifikovaná priamoúmerná závislosť. Veľké rozdiely sú však v miere závislosti. 3) Kvantita nie je určujúcim faktorom hodnotenia kvality SP. 4) Na znížení plynulosti slovenčiny sa podieľajú chyby v nominálnej morfosyntaxi.

Súhrn výsledkov riešenia projektu a naplnenia cieľov projektu v anglickom jazyku (max. 20 riadkov)

The project has focused on evaluation methods of machine translation (MT) quality from English/German into Slovak by using methods of manual evaluation, including the identification and classification of MT errors. In order to monitor both the accuracy of meaning and language, a categorical framework covering various linguistic spheres and categories has been designed. The model reflects the four-dimensional MQM error model observing the following issues: 1. language, i.e. grammar (morphology and syntax, or morpho-syntax); 2. accuracy, particularly incorrect/inadequate transfer into the target language; 3. terminology – it represents inadequate transfer of the term from the source language; and 4. style observes discrepancies between the style of the source text and the style of the target text.
The results of error analysis of machine translation output of the selected texts allow us to formulate the following conclusions: 1) In relation to text comprehensibility, double errors occur: those that do not make the text incomprehensible and those that make the text unit and the whole text incomprehensible. 2) For all error categories, a direct dependence was identified. However, there are large differences in the degree of dependency. 3) Quantity is not a determinative factor in MT quality assessment. A higher number of some errors does

not automatically mean a lower quality of MT output when the errors do not affect the comprehensibility of the text. On the other hand, a lower number of errors in predictability can cause a less successful transfer. The most serious errors are errors in the category of predictiveness, and in the incorrect identification of subject and verb relation which causes a lesser degree of comprehensibility.4) The errors in nominal morphosyntax determine the fluency of the Slovak language.