

Oponentský posudek doktorské disertační práce

Petr Homola: Syntactic Analysis in Machine Translation

Disertační práce Petra Homoly se zabývá problematikou strojového překladu na základě lingvistického rozboru jazykových jevů na lexikální, morfologické a syntaktické úrovni. Inspirací pro tuto práci jsou existující moduly pro strojový překlad mezi relativně blízkými jazyky, např. Česílo pro překlad mezi češtinou a slovenštinou, GUAT pro překlad mezi slovinštinou a srbštinou apod. vč. řadu experimentů pro překlad mezi skandinávskými jazyky. Táto práce prezentuje strojový překlad s částečným transferem založený na jazykových zvláštностech výchozího a cílového jazyka; zobecnění lze následně využívat na základě analýze podobnosti jazyků v rámci jednotlivých jazykových množin a pro jednotlivé strukturální vrstvy - autor pracuje s lexikální, morfologickou a syntaktickou vrstvou. V práci je experimentálně ověřeno, že vytčený směr má svoji perspektivu jak z hlediska standardně používaných evaluací jednotlivých překladů tak i z hlediska eliminace běžných faux pax ryze statistických metod.

Této práci má základní přínos v posílení následujících tvrzení:

1. Role taggeru ve strojovém překladu (založeném na pravidlech) je důležitá i pro překlad blízkých jazyků
2. Částečná gramatika pro jazyky s bohatou flexí je užitečným nástrojem pro překonání morfo-syntaktických rozdílů a pro redukci morfologických víceznačností; formalizovaná pravidla odlišností jazykových párů napomáhají transferu na lexikální, morfologickou a syntaktickou úroveň
3. Využití dvou strojových překladů za účelem získání jednotného kvalitnějšího překladového výstupů poskytuje validní experimentální prostor - tato práce předkládá experiment tohoto typu

Práce obdobného charakteru by neměly přínosy této práce ignorovat a měly by je buď nadále rozvíjet nebo vymezovat.

Navržený postup má svoji perspektivu v rámci aplikací pro překlad příbuzných jazyků. Ponechává však prostor pro bližší rozbor experimentů: výsledky prezentované v tabulce 8.2 a v tabulce 8.3 poukazují, že využití gramatických znalostí obou jazyků přispívá zejména u příbuzných avšak typologicky odlišných párů a jejich přínos je menší pro páry blízké a typologicky shodné, např. čeština a slovenština. Elaborace důvodů těchto rozdílů na úrovni prezentovaných přesností zůstává nedostatečně zodpovězenou otázkou této práce.

Z hlediska způsobu prezentace textu tato práce má v úvodních popisech lingvistických rozdílů velké cíle, jejichž shrnutí neshledávám dostatečně formální a v částech i méně přehledné. Co do šíře záběru autor vynaložil velké úsilí o nevynechání důležitých rozdílů - tyto jsou však stručně sepsány a chybí u nich provázanost mezi teoretickým příkladem a výsledkem překladu navrhovaných pravidel. Doporučuji proto v této práci pokračovat a ve vymezených částech jít do větší formální hloubky. Nedostatečnou formalizaci shledávám i u některých ne-lingvistických pojmů, zejména např. pro *shakle* či pro *c-forest*.

Dále shrnuji některé nedostatky:

- Příklad 5.9: forma slovesa *видам* v presentu neexistuje, je třeba ji nahradit formou *зледам*
- Příklad 5.10: forma slovesa *рекам* neexistuje vůbec, je třeba ji nahradit formou *велем*
- Příklad 5.11: i když příklad je uvádí správně, je důležité v kontextu poukázat na důležitost pořadí krátkých zájmenných forem v makedonštině
- Podkapitola 5.1.2: Bulharština má i neurčitý člen (... *Bulgarian and Macedonian which have a definite article.*)
- Podkapitola 5.1.2: Text "*In Baltic and Slavic languages, the definiteness is reflected morphologically at the adjective. The so called short (nominal) forms are indefinite whereas the long (pronominal) forms express definiteness.*" je příliš zobecněn a nelze ho v tomto znění aplikovat na současné slovanské jazyky vyjma například chorvatštiny a srbštiny.
- Podkapitola 5.1.4: "*Macedonian uses only prepositional phrases to express possessivity*" - tento způsob je nejčastější avšak není jediný a existují případy kdy není obvyklé vyjádřit posesivitu předložkou nýbrž přivlastňovacím zájmenem: *ке го прашиам татко ми, такто ми рече дека* apod.
- Příklad 5.21: *президентот* jako forma neexistuje, je třeba ji nahradit formou *претседателот*
- Příklad 5.25: *npjдеи* je v současné makedonštině mimo použití, existuje forma *npjдеи*, avšak v tomto příkladě je nutné použít formu *доjдеи*
- Podkapitola 5.2.1., část Simple Past Tense, první odstavec: "*These have disappeared except in ...*" je zapotřebí zařadit mezi výjimky i srbštinu
- Podkapitola 5.2.1., část The Category of Aspect - poznámka: budoucí čas je v makedonštině vždy perifrastický
- Str. 80: odstavec nad podkapitolou 7.2.2, makedonské příklady mají znít následovně *напишав/имаам напишано ... напишал сум ... сум напишал*

Žádná z těchto připomínek nemá zásadní význam pro výsledek práce. Práce jasně vymezuje svoje přínosy a buduje základy pro další studii, která má perspektivu aplikační a syntetickou (vazby existující teorie a praxe).

Disertační práce prokazuje předpoklady autora k samostatné tvořivé práci. Práci doporučuji k obhajobě.

Praha, 2. srpna 2006

RNDr. Kiril Ribarov