# Oponentský posudek diplomové práce

**Jan Rouš: Probabilistic Translation Dictionary**

Diplomant Jan Rouš se v předkládané práci zabývá konstrukcí pravděpodobnostního překladového slovníku. Práce je to implementačně-experimentální, psaná anglicky. Text je rozdělen do 9 kapitol na celkem 59 stranách. Dále je práce opatřena 5 přílohami, včetně CD obsahující vytvořený slovník a kód systému TectoMT.

Poměrně rozsáhlý úvod autor věnuje problematice strojového překladu a popisu systému TectoMT, pro který je vytvářený slovník určen.

V druhé kapitole se autor snaží o formální popis použitých postupů a také slovníku, jehož vytvoření je cílem práce. Formálně definuje několik pojmů, zabývá se maximálně věrohodnými odhady apod. Část kapitoly je věnována popisu tzv. současného slovníku TectoMT, který má být nahrazen novým.

Kapitola třetí čtenáře seznamuje s daty použitými v této práci a úrovní jejich lingvistického zpracování. Je to především paralelní česko-anglický korpus CzEng 0.7 a jednojazyčné korpusy BNC (anglický) a SYN2006PUB a SYN2005 (oba české). Poslední část je věnována tzv. manuálním slovníkům z nichž je zmíněn jeden, a to GNU/FDL anglicko-český slovník.

V kapitole čtvrté diplomant popisuje přípravné práce, které bylo dle autora nutné provést před hlavními experimenty. V první části autor pojednává o konverzi a shlukování tagsetu z Penntreebanku, v druhé pak o automatické anotaci manuálních slovníků.

Pátá kapitola je věnována vytváření samotného slovníku a dle autora tvoří jádro práce. V její úvodní části autor popisuje použití slovníku v překladovém systému TectoMT. V další části se autor zabývá konkrétní strukturou slovníku, resp. jeho položek. Dále autor popisuje vlastní konstrukci slovníku, tedy extrakci překladových párů a jejich následné filtrování (tzv. pruning). Tento krok je popsán velice detailně. Ve zbytku kapitoly se autor zabývá několika technikami, které souvisí s možností slovníky zkombinovat (lineární interpolace, backoff, perceptron a bucketed smoothing).

V kapitole šesté se autor zabývá několika nedostatky slovníku (danými už jeho návrhem) a navrhuje rozšíření slovníku, která by je řešila. Jedná se o problém s nepřenášenou negací a překlad složenin.

Sedmá kapitola je nazvaná „Evaluace", ale je po většinou teoretická a ke skutečnému vyhodnocení se autor nedostává. Autor se nejdříve pokouší navrhnout nové evaluační míry, které by mohly nahradit standardně používané míry BLEU/NIST (z důvodu výpočetní náročnosti celého překladu ). Dále stručně popisuje různé evaluační postupy, které by bylo možné pro evaluaci jeho slovníku využít.

Kapitola osmá obsahuje některé implementační detaily provedených experimentů.

Kapitola devátá je závěrečná, obsahuje jednak shrnutí některých zjištění, ke kterým autor dospěl v průběhu práce, ale také zcela nové výsledky (ovšem bez dalších detailů), konkrétně BLUE/NIST skóre překladového systému, který používá nový slovník. Lze konstatovat, že ve srovnání se systémem využívajícím původní slovník, dochází jak u BLEU a NIST ke zlepšení.

## Hodnocení

Je zřejmé, že diplomant se důkladně seznámil s problematikou strojového překladu, především se systémem TectoMT a provedl řadu složitých experimentů. Zadání práce splnil – zkonstruoval požadovaný pravděpodobnostní slovník, který navíc vedl k celkovém zlepšení překladu pomocí TectoMT. Navíc přispěl návrhem nových evaluačních metrik a dalšími experimenty.

Text práce ale působí dojmem jisté rozpracovanosti a nedokončenosti, navíc obsahuje poněkud rozporuplné a nikam nevedoucí pasáže. Zdá se, jako by autor měl původně v plánu úlohu řešit poněkud komplexněji a ve větší šíři, ale k „dotažení" se nedostal.

Jak již bylo uvedeno, zamýšlený slovník se autorovi vytvořit podařilo, na přiloženém CD lze nalézt dva. I přesto, že se jedná o hlavní výsledek práce, autor slovníky v práci nijak blíže nepopisuje (co obsahují za informace, jak interpretovat číselné údaje, kterými jsou opatřeny překladové páry, apod.) ani nesrovnává se slovníkem původním (kvalitativně ani kvantitativně). Je také nutné poznamenat, že vytvoření těchto slovníků je poměrně triviální. Z dat, které měl diplomant k dispozici, to lze provést jednoduchým programem v Perlu o několika řádcích.

Autor se v práci snaží kolem celé problematiky konstrukce pravděpodobnostního překladového slovníku vybudovat rozsáhlý matematicko-formální aparát, což je nejenom zbytečné (pro úlohu tohoto rozsahu), ale ještě se při tom dopouští chyb a nepřesností. Už v první definici (str. 19), ve které autor definuje tzv. model, není jasné, co přesně označuje symboly $S$ a $T$. Jsou to množiny odpovídající slovům (typům) v jazyce, a nebo spíše posloupnosti (či. multimnožiny) jednotlivých slov (tokenů) z dat. V definici 2 níže pak autor definuje

tzv. redukci modelu, která se ovšem jeví jako zcela zbytečná: už v definici je uvedeno, že funkce $\phi$ a $\psi$ jsou většinou identity, takže zřejmě k žádné „redukci" ve skutečnosti nedochází a navíc v jediném místě, kde se zdá, že by redukované modely mohly být použity (definice lineární interpolace modelů, str. 20 dole), autor uvádí, že tato vlastnost není u interpolovaných modelů nutná (což je správně).

Podobných formálních nepřesností a nekorektností je v práci více. Hned na stejné straně (str. 20, třetí odstavec) autor uvádí, že pro každý model $M = (A, B, C_M)$, existují odhady podmíněných pravděpodobnostních rozdělení na prostoru $A \times B$, což není pravda z důvodu, že nelze dělit nulou. Podobně problematická je definice 4 na straně 56, kde autor poměrně neobvykle definuje tzv. konzistenci metrik (opravdu ji autor takto zamýšlel?) a především následující tvrzení, které opět neplatí (paradoxní je, že toto tvrzení autor nikde nepoužívá a zřejmě je tedy zbytečné).

Práce je v toto směru značně nevyvážená. Autor se na jednu stranu snaží být velice formální (co lze, definuje matematicky – což samo o sobě není na škodu, ale často chybí popis význam použitých symbolů a čtenář je nucen jen hádat, viz např. část 7.1), na druhou stranu vůbec nedefinuje jiné, pro práci stěžejní pojmy, které pak běžně používá i v matematických zápisech (např. alignment). Obtížné je porozumět výrazům se symboly $t$ a $s$, které se v práci často vyskytují a není jasné, co přesně označují (zřejmě se vztahují ke zdrojové a cílové straně překladu, vztahují se ale ke slovům, tecto-uzlům, nebo celým větám?).

Jako poměrně problematická se jeví celá kapitola 4. Není totiž zřejmé, jaký význam mají jednotlivé popsané kroky a experimenty. Aniž by je definoval, autor se zde zabývá tzv. correspondence models. Porovnává odhady pravděpodobnostního rozdělení značek a slovních druhů na zdrojové a cílové straně získané z několika vzorků paralelních dat a konstatuje, že jsou podobné. Není ovšem jasné, co z tohoto pozorování pro slovník vyplývá. Na konci kapitoly se sice dozvídáme, že correspondence model byl použit při značkování manuálního slovníku (není ovšem uvedeno jakého), ale dále v práci takto označkovaný slovník není použit. Problém je také v samotném značkování slovníku. Ve vzorci 4.5 (str. 30) figuruje ještě jeden model $(P(x|L_s))$ o kterém není opět jasné, kde se vzal. V této kapitole se autor také velice podrobně zabývá shlukováním anglických značek, ale opět není zřejmé proč. Dále v práci již tento krok není zmíněn a zřejmě ani využit.

Výtku si zaslouží také část kapitoly 5, která se týká kombinace modelů. Autor popisuje několik možných postupů. Zaměřuje se na použití perceptronu a metody bucketed smoothing, ale záhy tuto volby označuje za nevhodnou nicneříkajícím a nijak nevysvětlujícím tvrzením „However, tests have shown that the reranker is not suitable for this task" (str. 43). Je opravdu škoda, že autor neuvádí, jaké testy provedl a proč jej výsledky vedly k tomuto závěru. Autor dále volí lineární interpolaci, jako vhodný způsob kombinace dvou pravděpodobnostních slovníků i přesto, že na str. 20 (správně!) uvádí, že parametry lineární interpolace lze optimalizovat pomocí EM algoritmu, tak se uchyluje k nevhodné „ruční" optimalizaci (str. 43) navíc bez uvedení cílové funkce! Podobně nevěrohodný postup je použit v části 5.4.7. V obou případech je trénování (optimalizace) prováděna na testovacích datech, což snižuje validitu výsledků.

V kapitole sedmé čtenář očekává evaluaci vytvořeného slovníku (jake je uvedeno v části 1.5, str. 16), ale tato kapitola žádnou takovou evaluaci neobsahuje. Autor nejdříve navrhuje a implementuje evaluační míry TTM a F-comp, které mají (částečně) nahradit míry BLEU/NIST při ladění parametrů celého systému. Autor sice tvrdí, že je použití těchto nových metrik na místo BLEU/NIST možné (str. 57), ale výsledky experimentů, které to mají potvrdit, tak jednoznačné nejsou. Ve zbytku kapitoly autor stručně teoreticky popisuje různé možnosti evaluace slovníku, ale praktické experiemnty a výsledky chybí.

Jediné znaky evaluace jsou v kapitole poslední (Závěr), kde by je čtenář ovšem nečekal, a tvoří je čtveřice BLEU/NIST skór, která ukazuje, že použití nového slovníku vede k jistému zlepšení překladu pomocí TectoMT. Autor ale opomíjí zmínit, jak se těchto výsledků dobral, či jaká testovací data použil. Z celé práce dokonce ani není zřejmé, v jakém směru autor překládal! Z češtiny do angličtiny, nebo naopak?

Na přiloženém CD je aktuální verze systému TectoMT, není ale jasné, kterých částí je diplomant autorem a které se přímo týkají jeho diplomové práce, a tak není možné určit, kolik kódu bylo nutné pro získání výsledků napsat. Práce není příliš samovysvětlující, bez předchozí znalosti systému TectoMT je pochopení mnoha pasáží velice problematické. Porozumění stěžují také časté gramatické chyby. Systematickým nedostatkem práce jsou pak chybějící reference (ať již na publikovanou literaturu, nebo jiné části práce, na jejichž základě autor konstruuje různá tvrzení), nedostatečný popis obrázků a tabulek.

Závěr

Práce sice obsahuje řadu nedostatků, ale pokud se k nim autor při obhajobě dostatečně vyjádří, doporučuji, aby byla obhájena.

Vypracoval: Pavel Pecina, ÚFAL MFF UK

Praha, 7.9.2009