Oponentský posudek doktorské disertační práce

RNDr. Jan Cuřín

Statistical Methods in Czech-English Machine Translation

Disertace se zabývá možnostmi využití statistických metod při strojovém překladu z češtiny do angličtiny. Je to obtížné téma – automatizace překladu přirozeného jazyka patří mezi nejnáročnější úkoly počítačové lingvistiky. Navrhnout a realizovat adekvátní řešení tradičními prostředky, tedy lingvistickými pravidly, možná ani nejde. Taková pravidla by totiž měla explicitně popsat znalosti a postupy, které jsou nutné k analýze vstupního textu, vyjádření výsledků této analýzy v pojmech cílového jazyka a vytvoření gramaticky i stylisticky přijatelného textu v cílovém jazyce, zachovávajícího význam originálu. Ne vždy jsou k tomu nutné všechny zdroje, které má k dispozici překladatel – člověk, například schopnost představit si skutečnosti v textu popisované. Záleží na typologické blízkosti jazyků, žánru, tématu, jazykových prostředcích. Přesto lze strojový překlad považovat za disciplínu, která s sebou potenciálně nese všechny problémy teoretické lingvisticky a počítačového zpracování jazyka.

Podobně jako u některých jiných úkolů počítačového zpracování jazyka se od konce minulého století začínají prosazovat řešení založená na využití statistických metod. Místo snahy odhalovat zákonitosti a způsoby užívání jazykového systému a pokusů o jejich explicitní a implementovatelnou formulaci nastupuje simulace vědomého poznání empirické skutečnosti v podobě velkého objemu statisticky zpracovaných dat, která jsou jako model jazyka a jeho užívání přímo použita při řešení například strojového překladu. Přestože má při takové metodě výsledek empirického zkoumání a jeho zobecnění v lepším případě povahu surových statistických dat a v horším případě neprůhledné černé skříňky, mohou být statistické metody v lingvistice z hlediska praktického využití úspěšnější. V jejich prospěch svědčí kromě inherentních výhod (relativní nezávislost na jazyce, absence pracného vývoje zdrojů pro jednotlivé jazyky) a dlouhodobých trendů (postupně ztrácejí na váze hardwarová omezení limitující kapacitně náročné metod) i předložená práce, která konstatuje, že takové metody jsou srovnatelné s komerčně dostupnými systémy vyvinutými tradičními způsoby a – alespoň při porovnání dvou provedených experimentů – překonávají systém s převahou tradičních metod.

Poznámky a výhrady uvádím v kontextu příslušných pasáží, pro přehlednost *kurzívou*, přání a dotazy ***tučně.***

Na první pohled je zřejmé, že autor věnoval tématu mimořádné úsilí. Pasáže věnované popisu přípravných prací (kap. 3), vytváření slovníků (kap. 4), vlastních experimentů (kap. 5) a jejich vyhodnocení (kap. 6) svědčí spolu s třemi dodatky o obrovském rozsahu práce s daty. Je samozřejmé, že zejména pořízení velkých a syntakticky označkovaných korpusů je zcela mimo možnosti jednotlivce a předpokládá týmovou práci. Odkazy pečlivě dokumentují podíl kolegů, zejména na přípravě korpusů, ale i na výsledcích prezentovaných v kap. 4 a 5. *Přesto by asi bylo vhodné explicitně uvést, kdy jde o týmovou práci a kdy o vlastní výsledky. Tím by se také poněkud korigoval občasný dojem, že disertace kvůli širokému tématickému záběru působí v některých pasážích (kap. 4) spíše jako sborník než monografie.*

To je však jen poznámka na okraj. Autor nás uvádí do tématu fundovaně, výklad je srozumitelný, postupuje logicky a směřuje k pointě celé práce: dvěma experimentům se strojovým překladem z češtiny do angličtiny, které různým způsobem kombinují statistické a „lingvistické" (dále „symbolické") metody. Po stručném objasnění rozdílu mezi symbolickými a statistickými

metodami a vylíčení předchozích experimentů se strojovým překladem na MFF UK následuje kap. 2, věnovaná možnostem využití statistického modelování jazyka ve strojovém překladu. Ty vycházejí ze systému Candide z konce 80. let, založeného na metodách označovaných jako model IBM 1 až 3, k nimž později přibyly modely 4 a 5. Experimenty se statistickým překladem dostaly mocný impuls v roce 1999 v podobě volně dostupných nástrojů GIZA (o rok později GIZA++), které slouží vytvoření těchto modelů na základě paralelních textů a dalších textů v cílovém jazyce. Tyto nástroje využil také autor disertace. Popis jednotlivých metod je důsledně ilustrován příklady a v dodatku A.1 je uveden trénovací algoritmus. Kapitolu uzavírá popis novější metody „alignment templates", která odstraňuje omezení původního překladového modelu na korespondence mezi jednotlivými slovy, a popis vlastního procesu překladu na základě vytvořených modelů, tedy tzv. dekódování.

Čtenář by možná získal ještě lepší představu o vývoji oboru a úloze statistického překladu, kdyby v úvodu padla zmínka také o hybridních koncepcích strojového překladu, které s paralelními texty pracují převážně symbolickými, nikoli statistickými metodami (example-based MT, translation memories). Navíc lze u některých takových koncepcí najít spojitost s předloženou prací v podobě syntakticky analyzovaných paralelních textů (viz např. Sato a Nagao 1990 [1] *a další práce citované v Somers 1999* [2]*). Uvádí-li autor přehled systémů vyvinutých na MFF UK*[3]*, bylo by vhodné uvést i geograficky širší kontext, třeba jen v podobě odkazů na reprezentativní, nejběžnější nebo jinak relevantní systémy nebo koncepce.*

Kapitola 3 se nezabývá paralelními korpusy obecně, jak by se mohlo z jejího názvu usuzovat, ale třemi anglicko-českými paralelními korpusy, které autor použil ve své práci. (V závěru kapitoly se jako „další zdroj" uvádí část Českého národního korpusu – asi 39 mil. slov a téměř 2,3 mil. vět, další anglické korpusy byly použity v obou experimentech k vytvoření jazykových modelů.) Dva z nich byly vytvořeny standardním způsobem z existujících paralelních textů: Reader's Digest Corpus (články z časopisu, v každém jazyce asi 70 tis. vět a 900 tis. slov) a Computer Oriented Corpus (hlášení a dokumentace operačního systému IBM, v každém jazyce asi 120 tis. vět a 1 mil. slov, *v závěru kapitoly je Reader's Digest Corpus uveden zbytečně znovu mezi „dalšími zdroji"*.) Autor uvádí popis klasické metody Gale-Church pro zarovnávání (alignment) vět v paralelních textech, která byla použita u obou korpusů. Vzhledem k neuspokojivým výsledkům u korpusu Reader's Digest, kde je český překlad volnější, byl tento korpus nakonec zarovnán jinak (na metodu SIMR/GSA je uveden pouze odkaz). Třetí korpus – Prague Czech-English Dependency Treebank (PCEDT, v každém jazyce 21 tis. vět; do tohoto korpusu patří navíc i tři slovníky, kterým se věnuje kap. 4) – je na rozdíl od předchozích dvou syntakticky označkovaný (povrchově i hloubkově, tedy analyticky a tektogramaticky) a vznikl (nikoli strojovým) překladem více než poloviny anglického korpusu Penn Treebank speciálně pro účely výzkumu strojového překladu. Český text se v tomto korpusu úmyslně drží originálu více, než je v překladatelské praxi obvyklé. Autor uvádí, že ze dvou možností – syntakticky označkovat již existující paralelní korpus, nebo přeložit již existující syntakticky označkovaný korpus – byla zvolena ta druhá proto, že značkování by bylo finančně i časově náročnější (str. 33).

Takový důvod může čtenáře zarazit, protože český překlad se beztak značkuje automaticky a nezávisle na značkování originálu (s.38), značkování originálu je třeba netriviálními postupy konvertovat (str. 40-45) a nástroje na automatické značkování anglického originálu by jistě byly po ruce. Další možná otázka se vznáší nad snahou přiblížit překlad trénovacích dat originálu, která je zřejmá i z krátké ukázky korpusu na str. 34. Důvodem byl nepochybně záměr přizpůsobit data metodě, která přináší lepší výsledky, pokud se překlad (zejména slovosledně) nevzdaluje příliš

---

[1] Sato, S. a M. Nagao:1990, Toward Memory-Based Translation, *COLING-90*, sv. 3, s. 247–252.
[2] Somers, H.: Review Article: Example-based Machine Translation, *Machine Translation* 14:113–157,1999.
[3] *Výčet členů týmu projektu RUSLAN asi není vyčerpávající, myslím, že tam ještě patří alespoň Vladislav Kuboň.*

originálu. Výsledek ovšem nezávisí jen na paralelním korpusu, který se používá k trénování překladového modelu, ale i na textech v cílovém jazyce, které slouží jako trénovací data pro model odpovědný za konečnou podobu výsledku. Přesto zůstává pochybnost, zda tato ne zcela přirozená data nebudou dříve či později omezovat možnosti této či jiné metody, která by byla s to zpracovat a využívat paralelní korpusy přirozenějších, volněji překládaných textů. Jistě by bylo vůči statistickým metodám ve strojovém překladu lichotivější předpokládat, že taková situace spolu s potřebou reálných textů nastane. ***Prosím autora o vyjádření.***

Po popisu postupu vytvoření tří paralelních korpusů pokračuje kap. 3 pojednáním o systému morfologického, analytického (povrchově syntaktického) a tektogramatického (hloubkově syntaktického) značkování, který byl převzat z Pražského závislostního korpusu a použit k označkování české i anglické části PCEDT. Závěr kapitoly je věnován vlastnímu postupu značkování.

Těžiště práce tvoří kap. 4 a 5. V úvodu kap. 4 autor zmiňuje systémy podpory překladu využívající databáze překladů, tzv. překladové paměti, kde optimální překlad vychází z existujících překladů celých vět (s. 50). *Statistický překlad má údajně stejný cíl, což ale není zcela zřejmé, neboť uváděné statistické metody vycházejí primárně z překladu jednotlivých slov. Také je trochu zavádějící tvrzení, že výsledky statistických metod mohou mít podobu nesrozumitelných vět, zatímco symbolické metody produkují zdánlivě přijatelné, ale zavádějící výsledky. Ve skutečnosti není žádná z obou metod vůči takovým výsledkům imunní.*[4]

Dále v kap. 4 autor zkoumá možnosti vytvoření nebo obohacení překladových slovníků na základě korpusových dat. Takové slovníky mohou být užitečné pro lidského uživatele i pro systém strojového překladu symbolické koncepce. Výsledkem jsou čtyři různé česko-anglické slovníky, z nichž první dva byly použity v „lingvističtějším" z obou experimentů popsaných v kap. 5:

a) „Velký" slovník pro transfer (109 tis. hesel, 159 tis. překladů s váhami, slovní druhy). Jako zdroj posloužily tři volně dostupné elektronické česko-anglické slovníky. Průměr 1,46 překladu na heslo byl dosažen několikastupňovým filtrováním na základě výskytu a frekvence v anglickém korpusu (365 mil. slov), váhy překladu podle zdroje, korespondence slovních druhů a porovnáním s korpusem PCEDT, které obstaral systém GIZA++ .

b) „Pravděpodobnostní slovník" pro PCEDT (46 tis. dvojic heslo-překlad, slovní druh), jen slova, která se vyskytují v ČNK (455 mil. slov – *patrně část ČNK zvaná „banka", která není veřejně přístupná*), jinak podobný postup jako výše.

c) „Slovník slovních tvarů" pro PCEDT (497 tis. dvojic heslo-překlad, „pro experimenty se strojovým překladem na čistých textech"), ukázka je zařazena jako dodatek B.1. *Obr. 4.5 na str. 56 uvádí příklad tvarů lemmatu „bankéř" s překlady „banker" a „bankers" a určením pádu a čísla, mezi tvary chybí lokál plurálu a mezi pády instrumentál plurálu* – ***je to proto, že se tyto možnosti v textech nevyskytly?***

d) „Terminologický slovník" získaný z (veřejnosti nedostupného) počítačového korpusu IBM, též anglicko-český; 6 tis. hesel v každém směru; zdroj: počítačový korpus IBM; obsahuje také víceslovné termíny označené v přípravné fázi regulární gramatikou. *Bylo by vhodné uvést,* ***proč nebyla použita některá ze standardních statistických metod k získávání kolokací z korpusu (mutual information, t-score, log-likelyhood)***. Výsledkem trénování jsou pravděpodobnosti všech

[4] Pokud lze za představitele symbolické metody považovat DBMT (s. 75–80), pak je ukázka nesrozumitelné věty po ruce: *The of he base have in Munich*, příklad z dodatku C.5 na s.114.

dvojic výrazů (jednoslovných i víceslovných) z odpovídajících vět a hlavním úkolem je najít kombinaci kritérií pro výběr relevantních dvojic. Úspěšnost optimálního postupu výběru dvojic je 85 %, u substantivních skupin 87–91 %. Ukázka 45 hesel „manage" až „maximum" výsledného slovníku na str. 61 tomu sice neodpovídá (obsahuje celkem 96 překladů, z toho jen 62 správných), ale to může být dáno snahou autora předvést ukázat typické chyby. Další ukázka v dodatku B.2 obsahuje chyb výrazně méně. *Každopádně je zřejmé, že takový slovník sice nelze rovnou vydat, ale může být velmi užitečný pro lexikografy nebo jako pomůcka pro překladatele. Autor uvádí na str. 59, že by měl pomoci překladatelům držet se zavedené terminologie. Možnost automatické tvorby terminologických slovníků nebo glosářů z existujících paralelních textů by překladatelé jistě uvítali, což ale předpokládá uživatelsky přítulnější implementaci.* ***Jsou slovníky veřejně dostupné?***

V kap. 4 a 5 se dostáváme k jádru práce, k popisu dvou experimentů s česko-anglickým překladem, které se liší úlohou a mírou využití statistických metod. Ten první (SMT) vychází ze systému Candide v implementaci GIZA, *překladový model byl trénován na PCEDT (aspoň takový dojem čtenář získá, explicitně uvedenou jsem tuto informaci nenašel),* jazykový model pro cílový jazyk na korpusu o 100 mil. slovech. Autor se nespokojil s aplikací na holé texty, ale zkusil výsledky vylepšit dvěma kroky. V prvním český text lemmatizoval, ve druhém lemmata doplnil značkami, které zhruba odpovídají repertoáru morfologických značek v anglickém korpusu (slovní druhy, číslo u substantiv, čas a 3. os. singuláru u slovesa). Kromě toho český text přiblížil anglickému doplněním nevyjádřeného podmětu[5] a také předložek a členů tam, kde angličtina tato pomocná slova oproti češtině používá. Jako vstup předzpracování byl využit výsledek povrchově syntaktické analýzy (analytická rovina, *víckrát použitý termín* analytical analysis *asi není příliš šťastný*), tedy poměrně náročný postup. Pravidla předzpracování jsou uvedena v dodatku A.3.[6] Jak lemmatizace, tak předzpracování vedly ke zlepšení (podle BLEU z 3,291 na 3,487 – str. 83, u profesionálních překladatelů uvádí autor hodnotu 5,560). ***Nebylo by výsledek možné ještě zlepšit využitím dalších informací ze syntaktické analýzy, případně i předzpracováním anglického textu?*** *To by znamenalo vyšší roli lingvistických metod a jistý posun ke koncepci druhého experimentu.*

Zajímavá je pasáž o výsledcích staršího experimentu s korpusem Reader's Digest, které byly porovnány s výsledky dvou dostupných komerčních systémů. V nejsofistikovanější variantě (s předzpracováním českého textu a alignment templates) byl výsledek téměř stejně dobrý jako u lepšího komerčního systému. *Bylo by zajímavé učinit podobné srovnání znovu metodou BLEU s novými systémy.*

Experimentovat může i čtenář: v dodatku A.2 je zdokumentována sada skriptů, které první z experimentů provedou. Vše lze získat včetně ukázkových dat z webových stránek autora. Oponentovi zatím bohužel nezbyl na vyzkoušení čas.

Druhý experiment (DBMT) vychází z klasického schématu analýza – transfer – syntéza s transferem na tektogramatické rovině. Analýza se opírá o nástroje použité při morfologickém a syntaktickém značkování české části PCEDT, včetně automatického odvození tektogramatického značkování. Využívá tedy např. Charniakův statistický parser. Při transferu se používají buď velký slovník pro transfer, nebo menší pravděpodobnostní slovník, popsané v kap. 3. Při syntéze se až na poslední fázi výběru z více vygenerovaných možností na základě trigramů (trénováno na korpusu o 52 milionech slov) uplatňují pravidla, nikoli statistika. Některé aspekty slovosledu a rozhodování o užití členu se dějí na základě údajů o aktuálním členění věty, jinak se provádějí běžné operace

---

[5] jako zájmena s gramatickými kategoriemi podle tvaru slovesa, *výraz „pro-drop languages" se užívá jako synonymní s „null-subject languages", jde tedy jen o zájmena v podmětové funkci, nikoli obecně o zájmena, jak by mohlo vyplývat z poznámky na s. 65 nebo z popisu pravidla V2 na s. 71*

[6] *Na str. 69 jsou dvě potenciálně matoucí pasáže: pravidlo N2 se zřejmě neuplatňuje v případě, kdy na substantivu závisí* přivlastňovací *zájmeno. A ve větě* The artificial article may not be inserted *jde patrně o význam „cannot be inserted", nikoli „it is possible that the article is not inserted".*

vkládající pomocná slova, upravující slovosled a generující slovní tvary. *Popis syntézy budí dojem, že jde o poměrně jednoduchý úkol, tím však jistě není. Bylo by vhodné uvést alespoň odkaz na podrobnější popis.*

Výsledky druhého experimentu jsou výrazně horší. S velkým slovníkem pro transfer bylo dosaženo nejlepšího skóre BLEU 0,1705, což je polovina hodnoty z předchozího experimentu. Skóre výrazně ovlivňuje rozsah slovníku a preference překladových variant, naopak minimální vliv má výběr možností podle trigramů. *Bohužel se autor nepokusil o rozbor příčin tak výrazně horších výsledků, které patrně nelze zdůvodnit tím, že statistické metody jsou obecně lepší – u předchozího experimentu symbolické metody pomohly a komerční nestatistické systémy dopadly u staršího hodnocení srovnatelně.* ***Prosím, aby tak učinil v rámci obhajoby.***

V dodatku C jsou uvedeny 4 verze strojového překladu úryvku z PCEDT o 10 větách (z češtiny do angličtiny): jednou je to výstup z DBMT (***druhý experiment, zde chybí věta 10, proč?***) a třikrát z SMT (první experiment), a to na holých textech, s lemmatizací a s předzpracováním. Srovnání s původním anglickým zněním z korpusu Penn Tree Bank, českým „originálem" a dvěma verzemi lidského překladu do angličtiny dává dobrou představu o možnostech systému.

Shrnutí

Kombinace statistických a symbolických metod při zpracování přirozeného jazyka je velmi zajímavé a aktuální téma, které slibuje pozoruhodné praktické i teoretické výsledky. Disertace představuje významný příspěvek k úsilí v tomto směru, navíc v jedné z nejobtížnějších disciplín. Autor provedl dva velmi rozsáhlé experimenty, které statistické a symbolické metody kombinují originálně, pokaždé od základu odlišným způsobem. Výsledky těchto experimentů, vyhodnocené uznávanou standardní metodou, vykazují díky navrženým řešením významná zlepšení. Navíc autor z existujících volně dostupných elektronických slovníků a paralelních korpusů vytvořil několik překladových slovníků s údaji o statistické významnosti ekvivalentů. Takový slovník je velmi užitečný pro MT i člověka, a navíc jde patrně o první exemplář svého druhu. Práce má nejlepší šance stát se východiskem pro další výzkum i praktické aplikace.

Dotazy, výhrady a připomínky se týkají méně podstatných aspektů. Po formální stránce je text přijatelný, autor se vyjadřuje srozumitelně, stručně a výstižně. Občasné prohřešky proti anglické gramatice srozumitelnosti nepřekážejí, důkladnější jazyková korektura by však byla na místě – objevují se i chyby v podmětové shodě (např. „if you translates" na str. 49). Po úpravách by byla žádoucí jeho publikace.

Disertační práce jednoznačně prokazuje předpoklady autora k samostatné tvořivé práci. Doporučuji, aby byla přijata k obhajobě.

15. srpna 2006

[redacted]
Ústav teoretické a [redacted] FF UK
Celetná 13
110 00 Praha 5