Posudek disertační práce

Mgr. Petra Homola
Syntactic Analysis in Machine Translation

Deklarovaným cílem práce je zkoumání úlohy syntaktické analýzy ve strojovém překladu mezi typologicky blízkými jazyky. Podobnost těchto jazyku přináší možnosti využití metod, které analyzují vstup jen částečně nebo na méně abstraktní rovině.

Po stručném výkladu základních pojmů (včetně pojmů teorie funkčního generativního popisu, kap. 2) autor uvádí zajímavý přehled systémů překladu mezi příbuznými jazyky (kap. 3), pak ukázky jevů, které jsou v baltských a slovanských jazycích natolik podobné, že je lze využít pro zjednodušení postupu překladu (kap. 4), a dále naopak ukázky rozdílu mezi nimi (kap. 5). Těžiště práce je v kapitolách 6 a 7, věnovaných „partial parseru" pro baltské a slovanské jazyky (a nejen pro ně) a modulu transferu a syntézy. Zde je také popsán úspěšný experiment se spojením dvou systému k překladu přes třetí jazyk (z češtiny do slovenštiny přes slovinštinu), přičemž toto spojení se děje na rovině „mezi morfologií a syntaxí". Tuto část doplňuje pasáž o jediném stochastickém modulu v systému, který se na základě trigramů snaží vybrat optimální výsledek, a hodnocení výsledků, které na základě obecně uznávaných kritérií dokládá výrazné zlepšení oproti předchozímu systému bez syntaktické analýzy (ale s morfologickým taggerem pro vstupní text) pro překlad z češtiny do ruštiny a do makedonštiny. Pro překlad do slovenštiny žádné výrazné zlepšení nenastalo („jazyky jsou si příliš blízké") a experiment s překladem do němčiny skončil negativně, což naznačuje meze navržené metody. Práci uzavírá zajímavé a výstižné shrnutí výsledků a poznatku a příloha s gramatickými pravidly.

Práce je po formální stránce bezvadná , po jazykové stránce výrazně nadprůměrná (mohu-li soudit jako nerodilý mluvčí), jednotlivosti uvádím níže (např. nadbytečné užívání členu u termínů: the *morphological underspecification*).

Hlavním a nesporným přínosem je navržený a implementovaný hybridní, převážně pravidlový systém („don't guess if you know"), který jednoznačně prokázal správnost výchozích předpokladu o využitelnosti povrchové syntaktické analýzy pro překlad mezi příbuznými jazyky. Dalším pozitivním výsledem je originální koncepce překladu přes třetí jazyk s využitím vyšší roviny. V obou případech je důležité, že tímto způsobem lze budovat systémy pro dvojice příbuzných jazyku bez zázemí v podobě velkých paralelních korpusu nebo dlouhá léta budovaných gramatik. Práce je přínosná a zajímavá i v částech věnovaných přehledu systému a porovnání jazyku. Lingvistické kapitoly svědčí o jazykové kompetenci i lingvistické kompetenci autora, přesto právě zde je možné vidět slabiny práce.

Lingvistické pasáže jsou totiž poněkud útržkovité a vyvolávají neuspokojená očekávání, že o problémech a jejich případném řešení bude zmínka někde dál. Podobně je tomu s kapitolou 2, která má zejména v teoreticko-lingvistické části nejasnou souvislost se vším, co následuje. Dalším problémem je autorovo užívání pojmu, které nejsou vysvětleny a někdy nejsou ani v souladu se svým obvyklým významem. Někdy vysvětlení přichází až později. Podrobněji níže v jednotlivých bodech.

Dotazy:

str. 14, odst. 1: Jak se FGP uplatní v navrženém systému? Jak se v něm projevuje jednoznačná orientace FGP na závislostní struktury? Pravidla gramatiky dále v textu i v příloze vypadají jako standardní pravidla bezkontextové (složkové) gramatiky. Tento (zdánlivý?) rozpor by bylo dobré vysvětlit.

str. 15, Tektogramatická rovina (hloubková syntax), opět – jak se dále uplatní? Dále se vyskytuje termín *hloubková pravidla*, ale v jiném významu (např. příloha A, str. 104)

str. 37, odst. 1: *syntetické jazyky se analyzují snáze*, ale na str. 91 se o baltských a slovanských jazycích píše, že kvuli *velmi bohaté morfologii a extrémně rohatému slovosledu* je jejich zpracování obtížné, protože je nutné se ve srovnání s většinou germánských a románských jazyků vyrovnat s mnohem vyšší mírou víceznačnosti.

str. 38, 4.4, odst. 2: *Lexikální rozdíly se poměrně snadno řeší v glosářích a obecných slovnících* – to asi neplatí obecně, tzv. lexikální selekce bývá při generování velký problém, a i mezi blízkými jazyky se stává často, že vztah 1:1 mezi lexémem jednoho a druhého jazyka neplatí. Lze toto tvrzení nějak doložit analýzou chyb výsledků překladu?

str. 59, Mediopasivum není totéž, co reflexivní pasivum. Viz např. http://ucjtk.ff.cuni.cz/zdarek/prezentace/2008/19_hudouskova.pdf. Také není pravda, že se vyskytuje pouze ve slovanských jazycích. Prosím o vysvětlení.

str. 65, pozn. 2: *Lesem myslíme množinu složkových stromů, které reprezentují fragmenty analyzované věty a beze zbytku ji pokrývají* („*span it completely*"). Na další stránce, 1. odst.: *Parser nemusí analyzovat celé věty*. Tohle by chtělo lépe vysvětlit. Taky: co když parser daný fragment nedokáže analyzovat a žádný strom ho nereprezentuje? I k pozn. 3: tohle jsou důležité body a měly by být v hlavním textu.

Méně podstatné, upřesňující dotazy:

str. 17, konec předposledního odstavce: *příznakové prvky nebo konstrukce … lze pozorovat ve více jazycích po celém světě*. Ve více jazycích lze pozorovat určitý typ příznakového jevu, nebo příznakové jevy bez ohledu na typ?

str. 39, odst. 1: Předpokládá se, že substantivum řídí předložku, to může být překvapivé, zdůvodnění nebo alespoň odkaz by byl žádoucí. Vlastně existuje, ale až na str. 41, poslední odst. a dále, argumentem je, že předložky se podobají kategorii pádu. Na druhé straně existují i argumenty proti pojetí předložek jako nevázaných afixů: nemusejí být v kontaktní pozici (*po téhle velmi dlouhé větě*), mají pádovou rekci (stejné jako slovesa) a vyskytují se tak spolu s pádem (nikoli jen jako jeho alternativa), většinou mají specifický a stabilní význam, existují i víceslovné předložky, lze je koordinovat (*pod i nad stolem*), lze koordinovat „řídící" substantiva (*za devatero moři a řekami* – nebo tak nějak). Jaké praktické výhody plynou z použitého řešení? Jak se v systému řeší poslední dva jevy?

str. 40, odst. 2: Pro řídící člen se používá výraz *core*, lze jej považovat za synonymní s obvyklým *head*? Pokud ano, proč se nepoužívá obvyklý termín? Pokud ne, bylo by dobré objasnit rozdíl. Opět na str. 57, za 1. nadpisem odstavce.

str. 45, 5.1.3., odst. 2 a dál: Chybí zmínka o tom, že v postpozici bývají rozvitá adjektiva. Řeší se v gramatice nějak příklady se zdvojeným determinátorem (*tvůj dům tvůj nóvójski*) nebo předložkou (*u bratra u staršího*)? Obecně: v této kapitole se uvádí řada pozoruhodných jevů, ale není jasné, které z nich se opravdu řeší nebo alespoň mohou řešit v navrhovaném systému.

Uvedená pravidla by zasluhovala vysvětlení, obecně o pravidlech i konkrétně o těchto dvou.

Ve standardní bezkontextové gramatice by u vstupu typu *Polska Rzeczpospolita Ludowa* nebo *typická kočka domácí* tato pravidla dávala dva výsledky, pravděpodobně nežádoucí. Zde je třeba uvést, že spurious ambiguities se řeší eliminací identických výsledků mimo gramatiku.

Symbol A stojí za lexikální nebo frázovou kategorii? A co znamená symbol N'? Čím se liší od N?

str. 45, odst. 2 od konce: *V jazycích s adjektivní a substantivní pádovou flexí se adjektivum musí se svým řídícím členem shodovat v rodě, čísle a pádu*. Platí tohle obecně? Neexistují jazyky, kde se substantiva a adjektiva skloňují jen podle pádu, ne v rodě a čísle?

str. 47, odst. 1: Substantiva s více než jedním genitivním přívlastkem se považují za příznakové případy, a příklad *královna krásy České republiky* se vysvětluje tak, že *královna krásy* je kolokace, která funguje jako jedna jednotka. Ale: (a) Je správné říct, že substantivní skupina (NP?) *České republiky* nemodifikuje substantivum *královna*, ale jednotku *královna krásy*. To lze ověřit i na významu celého výrazu. Tato jednotka sice může být kolokace, ale i kolokace má svou strukturu, substantivum s genitivním přívlastkem. Můžeme pak zachovat bez výjimek předpoklad, že genitivní přívlastek je jen jeden, ale dostáváme se mimo pole

závislostní gramatiky. (ii) Příklady tohoto typu jsou docela běžné, poukaz na kolokace není vždy obhajitelný: *model letadla našeho Péti*, [ ] [*sbírka známek*] [*nevyčíslitelné ceny*]] [*naší babičky*]].

str. 48, odst. 1, příklad N' → PP N' … v jakém jazyce PP předchází rozvíjené N' ?

str. 55, př. 5.43: vyskytuje se také varianta 4: *korova padnena* ?

str. 58, odst. 2 od konce: Je ve slovenštině běžný genitiv záporový? Není-li, je slovenština také pod vlivem němčiny, nebo zprostředkované přes češtinu? Obecně: opravdu všechny slovanské jazyky bez gen.záp. o něj přišly pod vlivem němčiny?

str. 65, poslední odst.: […] *e-les („e-forest") nereprezentuje strukturu věty jako takovou, ale konkrétní posloupnost aplikace pravidel.* Jaký je rozdíl mezi e-lesem a derivačním stromem, kromě toho, že les může mít víc stromů?

str. 66, 6.6.1, odst. 1: Co je to *chunk parser*? Jak se liší od chart parseru? Autor disertace by u čtenáře neměl předpokládat znalosti speciálních pojmu. … *derivační proces je bezkontextový (ve smyslu Chomského hierarchie), což má pro slovanské jazyky ten klíčový důsledek, že se nedokáže vypořádat s neprojektivními konstrukcemi (alespoň ne přímo).* Proč se tedy takový parser používá?

str. 68–69: *shackles*: Je vždy žádoucí, aby nepoužité hrany na spojité cestě od začátku do konce mizely? Pokud ano, proč se to řeší v gramatice a ne obecněji ve formalismu/parseru?

str. 69–70: likvidace identických výsledku – tohle je naopak případ, který se považuje za takřka žádoucí vlastnost bezkontextových gramatik, za prevenci *spurious ambiguities* odpovídá tvůrce gramatiky, nikoli parseru. Považuje autor toto řešení za přizpůsobení CFG závislostní teorii?

str. 71, bod 3: Neudělá tato procedura tutéž službu jako ony *shackles*?

str. 76, př. 7.1: Uvažoval autor o tom, že by slovník mohl být obousměrný?

str. 78, tabulka 7.1, popis příkazu newChild: *the attribute name of the new feature structure* – není mi jasné, co to je, nemá to být typ struktury? Atribut relorder má ve druhém příkladu na následující straně hodnotu –9, co to znamená?

str. 79, první příklad: transfer *u-l* na *pri* vypadá jako lexikální transfer, který se má provádět až v dalším kroku

str. 79, druhý příklad: v češtině bude slovo přičesti ve 3. osobě označeno patrně jako finitní, ve slovinštině už to platit nebude. Obrácené v dalším příkladu.

str. 80–82, zvl. obr. 7.5: Není mi úplně jasné, jak vypadají data na výstupu z prvního a vstupu do druhého systému. Jsou to stromy nebo řetězce označkovaných lemat?

str. 85, odst. 1–2: Mělo by se uvést, že takový výsledek dávají povrchová pravidla, hloubková by to zvládla, je to tak?

Připomínky:

str. 18, odst. 1: *kombinatorická varianta* – popisovaný příklad se v kontextu překladu mezi dvěma jazyky obvykle označuje termínem překladová homonymie (v češtině) a synonymie (v litevštině)

str. 18 a dále: *podspecifikace* – Opět případ použití pojmu spíš z informatiky v čistě lingvistickém kontextu, může to být záměr, přesto by vysvětlení vztahu k běžnějším termínům homonymie a polysémie nebylo na škodu. Uvedené příklady jsou navíc ukázkou spíše disjunkce než podspecifikace, která se obvykle chápe jako nedisjunktní vyjádření, např. pomocí nadtypů.

str. 25, poslední odstavec: *ranker* – myslím, že se objevilo poprvé, bylo by vhodné vysvětlit, o co jde

str. 33, 4.1 Typologická podobnost: má jistě aspekty syntaktické, morfologické i lexikální, není proto zřejmé, proč je třeba ji vydělovat zvlášť, když se další sekce zabývají právě těmito jednotlivými aspekty

str. 37, odst. 1: *se* je zde spíše částice než zájmeno (*rozhodl se* je reflexivum tantum, *inherent reflexive*)

str. 40, odst. 2: Chybí zmínka o tom, že NP může obsahovat vztažnou větu, a tedy řadu problémů spojených se větnou syntaxí.

str. 42, 5.1.1, odst. 1: Přidělování pádu v jazycích bez pádových koncovek u substantiv, jako je angličtina a francouzština, a v ostatních jazycích se staví proti sobě. Ve skutečnosti i v těch prvních musí být pád alespoň někdy vyjádřen (zájmena), jinak by kategorie morfologického pádu neměla smysl (čínština). Angličtina a čeština se výrazně liší v morfologii, „přidělování pádu" je věcí syntaxe a může být velmi podobné. Překlep: *morpheme*.

str. 43, odst. 1: V komentáři k litevskému příkladu s anglickou glosou *The moon is visible tonight* se uvádí, že ve slovanských jazycích by se užil akuzativ. Překladem anglické glosy by však mohlo být *Dnes večer je vidět luna/lunu*.

str. 44–45, příklady 5.14–5.18: Bylo by dobré opatřit příklady glosami i s kategoriemi určitosti a specifičnosti.

str. 45, odst. 3 od konce: Bylo by dobré uvést litevská adjektiva *mažas* a *mažasis* v kontextu, jako ostatní příklady.

str. 49, odst. 1: Čtenář již přivykl pravidlům, zde chybějí!

str. 61, odst. 1: Pokud anglický lékař vyšetřuje pacienta, pak ho *examinates*, nikoli *investigates*. A když ho ošetřuje, tak ho *treats*.

str. 62 a dále, možná i dříve: S rostoucí složitostí příkladů čtenář stále více touží po jejich řádném oglosování. Touha vrcholí v odstavci o „parazitních infinitivních komplementech".

str. 66, odst. 2 od konce: *by means of sole shallow rules*, snad lépe: *by means of shallow rules only*.

str. 73, odst. 2 od konce: V původních systémech Q nebyly typy ani atributy, bylo by dobře objasnit vztah mezi typovanými FS a stromem.

str. 73, pozn. 6: zbytečná, totéž se říká na str. 75

str. 84, odst. 1: *stochastic~~hal~~*

str. 86, pozn. 2: *Bojar* místo *Boran*, ještě že ne *Borat* ☺

str. 87, odst. 2 od konce, posl. věta: *appears*

str. 101 a dále: Příloha A: jednotlivá pravidla by zasluhovala komentář

Závěr

Uvedené nedostatky nezpochybňují celkový dojem, že disertace prokazuje předpoklady k samostatné tvořivé práci a že její autor by měl rozhodně dostat příležitost k její úspěšné obhajobě a získání titulu Ph.D.

8. srpna 2009