Vyjádření vedoucího doktorské disertační práce

Martin Čmejrek: Using Dependency Tree Structure for Czech-English Machine Translation

Předložená práce se zabývá zcela novou metodou převodu („transformací" nebo „substitucí") obecných stromových struktur pro účely strojového překladu (ale nejen jeho), a zároveň implementuje první obecný systém překladu z češtiny do angličtiny, který bude sloužit jako základní systém pro porovnání se systémy vytvořenými v budoucnu. Autor v práci dále popisuje první anotovaný paralelní česko-anglický (závislostní) treebank (PCEDT 1.0), na kterém se podstatnou měrou podílel jako na nezbytném zdroji dat pro svoje experimenty. Je třeba rovněž poznamenat, že autor se zúčastnil letního workshopu na Johns Hopkins University v roce 2002, kde byl daný systém za jeho účasti navržen a poprvé částečně implementován.

Práce je rozdělena do šesti kapitol a dvou příloh. Úvod je relativně podrobný, a autor v něm popisuje dosavadní systémy statistického strojového překladu i metody a postupy strojového překladu obecně (včetně přesných matematických formulí a formálních algoritmických postupů). Ve druhé kapitole pak popisuje vytvořené a použité datové zdroje (The Prague Czech-English Dependency Treebank, vydaný v roce 2004 v LDC, Philadelphia, PA, USA). Třetí kapitola popisuje základní překladový systém z češtiny do angličtiny a další kapitola systém tzv. pravděpodobnostních synchronních stromových substitučních (překladových) gramatik, poprvé tak vůbec precizně matematicky a algoritmicky popsaný ve zveřejněné publikaci (kap. 4). Pátá kapitola popisuje implementaci takových pravděpodobnostních substitučních gramatik včetně jejich aplikace na nalezení nejvhodnějšího párování (alignmentu) pro tři různé problémy vznikající při strukturálním strojovém překladu, a naznačuje se zde i směr možného dalšího vývoje. V šesté kapitole autor shrnuje svůj hlavní přínos k dané problematice překladu popsaný v této disertaci. Přílohy obsahují příklady alignmentu pro vybrané tři problémy (jejichž řešení je popsáno v kap. 5), a implementační detaily pro použití vyvinutých programových nástrojů.

Vlastní samostatný autorův přínos je popsán zejména v kapitole 5 (přičemž formulačně je jeho vlastní prací i kap. 4), a dále autor podstatně přispěl i k systému překladu popsanému v kap. 3 a k Pražskému česko-anglickému závislostnímu korpusu (kap. 2).

Hodnocení:

Celkově: výsledky autorovy práce jsou i z celosvětového hlediska originální a v některých bodech poprvé publikované (kap. 4 a 5). Je rovněž třeba konstatovat, že autorovy výsledky byly již 7x publikovány na mezinárodních konferencích nejvyšší třídy (ACL, EAMT, ...) a vyšly i 4 články v časopisech (PBML, LNAI, IJCL); tyto výsledky autor shrnuje nebo nově formuluje v ostatních částech práce. Přesto, že nově vyvinuté metody pravděpodobnostních substitučních gramatik nebyly „zasazeny" do systému strojového překladu a evaluovány v jeho kontextu, je vůbec fakt, že se podařilo zreprodukovat a ověřit prakticky nepopsanou práci J. Eisnera z Workshopu 2002 z JHU alespoň do fáze alignmentu (tj. bylo provedeno plné natrénování systému), velmi pozitivní.

Věcné (drobné) poznámky a výhrady: bylo by vhodné vysvětlit podrobněji, jak autor došel k prahům (thresholds: 0.10, „higher than the most probable translation so far") uvedeným v kap. 3.1.3, a věcně je zdůvodnit. Práci (nebo spíše nějaké další následné publikaci) by prospělo podrobnější zveřejnění pravidel pro přiřazení anglických analytických funkcí (popisované poněkud stručně v kap. 2). Rovněž by bylo vhodné ozřejmit pojem „renaming of frontier nodes", který není nikde definován a vzhledem k absenci literatury na dané téma jej čtenář může chápat pouze intuitivně.

Jazykové problémy: práce je psaná anglicky, a to srozumitelně a logicky, bylo však třeba ji pozorněji přečíst před závěrečnou publikací. Občas vázne shoda podmětu s přísudkem v čísle, někdy je nevhodně užitý člen (chybějící „the", ale častěji „the" místo správného „a/an"). Další drobnosti zahrnují např. nadbytečné pomlčky v neadjektivních termínech (parse-tree, sentence-pair, tree-pairs; naopak adjektivní užití packed-tree v kap. 3.2, plain-text v kap. 3.3 jsou samozřejmě správně). Při případné další publikaci na

konferenci či v časopise by bylo vhodné, aby text ještě jednou přečetl rodilý mluvčí a odstranil další drobné nedostatky (nejde přitom o logiku věci, ale o čistě jazykové detaily).

Formální chyby a překlepy: V popisu obrázku na str. 27 ma být (c) a (d) místo (3) a (4). V obr. 2.1. nejsou „slibované" identifikátory uzlů. Obr. 3.4. (který bohužel obsahuje spíše text než grafiku) vyšel ve výsledném stránkování nevhodně doprostřed číslovaného seznamu. Popisek k obr. 3.6. – chybí číslo sekce, ke které text na konci odkazuje. Na str. 75, „As in usual" má být „As is usual". V první větě kap. 5.9 vypadlo „evaluation of", což poněkud nevhodně pozměnilo smysl této věty.

Chybějící reference: experimenty se strojovým překladem založené na korespondenci mezi „analytickou" (povrchovou) rovinou (kap. 5.8., A.3) prováděl Yuan Ding (mj. ve své disertaci); přesto, že jím používané metody jsou zcela jiné, bylo by vhodné jeho práce uvést v seznamu literatury.

Závěr: celkově práci považuji za velmi dobrý příspěvek k problematice strojového překladu (vytvoření základního systému k budoucímu porovnávání) a zejména k metodě tzv. stromových transformací obecně. Přes některé výše uvedené výhrady (spíše však formálního nebo jazykového charakteru) doporučuji, aby byla přijata a obhájena jako práce disertační; autor dostatečně prokázal schopnost k samostatné tvořivé práci (jak mj. ukazuje i neobvykle velké množství jeho dalších mezinárodně publikovaných prací).