Oponentský posudek doktorské disertační práce

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

Doktorská disertační práce M. Čmejrka se zabývá využitím závislostních stromů při strojovém překladu (SP) mezi češtinou a angličtinou.
Práce má dvě části a je tvořena šesti kapitolami: první část obsahuje úvod, dále v ní autor věnuje pozornost paralelním syntakticky značkovaným korpusům (konkrétně PCEDT). Ve druhé části autor popisuje hybridní statistický a pravidlový systém strojového překladu Cz-E a stromový převodník (ve čtvrté kapitole) a jeho implementaci (v páté kapitole). Šestá kapitola je závěrečná a shrnující. Následují přílohy A a B. K práci je přiložen CD ROM s textem práce a veškerými přílohami včetně implementačních detailů a testovacích příkladů.

a) Práce je psána anglicky, výklad je přehledný a srozumitelný.
b) formální chyby: konstatuji, že překlad na s. 42 (Fig. 2.8 a 2.9) zdaleka není jediný možný (*Such loans remain classified... – obdobné úvěry jsou nadále klasifikovány*, proč ne doslovněji a přesněji *Takové úvěry zůstávají klasifikovány*)? Nakolik jsou pak dále vyvozované závěry spolehlivé?

Přínos práce:

c) Problematika zkoumaná v práci je aktuální, získané výsledky jsou nové.
d) Vlastní přínos práce spočívá ve významném podílu autora na vytvoření paralelního česko-anglického syntakticky značkovaného korpusu PCEDT (uvítal bych však přesnější specifikaci toho, co dělal autor sám a co jeho „bezejmenní" spolupracovníci).
e) autor v práci předkládá hybridní systému strojového překladu Cz-E kombinujícího statistický strojový překlad s pravidlovým.
f) originálním výsledkem je převodník stromů využitý v předkládaném systému SP.

Problematické body a otázky:

a) přehled historie MT (s.14-19) je značně fragmentární, některé výsledky, např. ruské (Kulagina, Meľčuk. 1955-6) a evropské, třeba projekt Eurotra, nejsou zmíněny. Autor cituje jen zprávu ALPAC, která se primárně vztahuje k situaci SP v USA, ostatní pozdější pozitivněji orientované zprávy evropské jsou ponechány stranou.
b) v práci se nabízela možnost podrobnější teoretické analýzy vztahů mezi SP využívajícím převodního jazyka a tektogramatickou rovinou popisu jazyku. Tyto vztahy jsou v práci jen naznačeny. Čekal bych, že si autor položí otázku, zda a nakolik může tektogramatická rovina, resp. notace pro ni v práci zmíněná a použitá, sloužit jako převodní jazyk na rozdíl od převodních jazyků založených na jazykově nezávislých logických formalismech, např. autorem necitovaný systém Rosetta využívající Montagueovy intenzionální logiky?
c) mohl by se autor pokusit objasnit (při obhajobě), zda a nakolik použití tektogramatické roviny zlepšuje kvalitu překladu?
d) výsledky překladu lze najít na přiloženém CD ROM, jsou však pro ne bezprostředně zasvěceného čtenáře obtížně čitelné (viz např. soubor n8801g.javatransfer.log.auto). Mohl by autor při obhajobě nabídnout výsledky v přístupnější podobě (např jako dostatečně velký seznam překladových dvojic představujících výstup z autorova systému, aby bylo

možno u jednotlivých dvojic přímo porovnávat kvalitu překladu (kromě použitých metrik)? Na s. 53-60 se probírají výsledky překladu (fakticky na jednom příkladě) a jejich vyhodnocení, přiznám se však, že se mi z nich nepodařilo jednoduše vyčíst, kolik vět bylo přeloženo, kolik správně a kolik chybně a s jakými konkrétními chybami. K Tab. 3.2 na s. 57 chybí potřebná kvalitativní interpretace. Vzhledem k povaze práce bych očekával, že tyto údaje najdu souhrnně i v závěrečné kapitole.

e) v práci jsou zmíněny některé lingvistické problémy spadající do oblasti konfrontačního studia češtiny a angličtiny, např. určování členů v anglickém překladu, výběr vhodných časů ve vztahu k českým vidům, úvahy o gerundiích a přechodnících (s. 41, 54-56). Rozhodování o kladení členů v anglickém překladu a výběru příslušných časů v závislosti na videch v českém originálu je pro kvalitu česko-anglického překladu relevantní, naproti tomu problematika přechodníků je v češtině naprosto okrajová, to lze snadno doložit korpusovými údaji, takže asi není potřeba se jí příliš zabývat. Kromě toho mám jisté pochybnosti o tom, že statistický přístup ke SP se s těmito problémy dokáže adekvátně vyrovnat (budu rád, když autor bude schopen mé pochybnosti vyvrátit).

f) je dobře vidět, že na stránkách 41-45 (viz též Fig. 2.12 a 2.13) se evidentně objevují potíže s problematickými složkovými strukturami pocházejícími z Penn Treebanku: zatímco Fig. 2.13 se jeví jako akceptovatelný, struktura v 2.12 je patrně deskriptivně neadekvátní. Není jasně řečeno, zda jde o jednotlivé případy (kolik jich je?) nebo mají systematickou povahu – odpověď na tuto otázku musí ovšem významně ovlivnit kvalitu výsledků překladu.

Hodnocení:

Práce přináší nové výsledky zahrnující tvorbu paralelního česko-anglického syntakticky značkovaného korpusu (PCEDT) a hybridní systém SP mezi češtinou a angličtinou. Práce však trpí jistou nevyvážeností, na jedné straně předkládá nová originální řešení technická, na druhé straně mi však v práci chybí odpovídající kvalitativní empirické interpretace. Nemám nic proti inženýrskému přístupu, ale jde tu přece o počítačovou lingvistiku, nejen o softwarové inženýrství?

Závěr:

Přes uvedené výhrady autor v práci prokázal, že se dovede samostatně vyrovnat se složitými problémy v oblasti počítačového zpracování přirozeného jazyka. Předloženou disertační práci pokládám za **vhodný podklad** pro získání stupně Ph. D.

V Brně, 23. 8. 2006