Oponentský posudek doktorské disertační práce Mgr. Martina Čmejrka

Posuzovaná disertační práce pojmenovaná „Using Dependency Tree Structure for Czech – English Machine Translation" se, jak už sám název napovídá, zabývá automatickým strojovým překladem (konkrétně pouze jednosměrným překladem z češtiny do angličtiny), což je téma bezpochyby velmi aktuální a dosud nikoliv uspokojivě vyřešené.

Stěžejní myšlenkou celé práce je možnost využití syntaktické informace obsažené v překládané větě pro zlepšení automatického překladu. Autor v práci představuje dvě metody, které z této základní premisy vycházejí. První z nich kombinuje statistickou syntaktickou analýzu (parsing) české věty s pravidly řízeným lexikálním transferem a následným (opět pravidly řízeným) generováním výsledného anglického překladu. Druhá metoda pak navrhuje přímo statisticky řízený převod mezi závislostními stromy v obou jazycích.

Jelikož oba tyto postupy zahrnují ve větší či menší míře využití metod strojového učení, bylo nutné vytvořit i příslušný datový korpus pro natrénování parametrů navržených modelů. Procesu tvorby těchto trénovacích dat – paralelního česko-anglického závislostního korpusu – je věnována jedna z úvodních kapitol práce.

Další kapitola se zabývá první z autorem použitých metod strojového překladu - statistickou syntaktickou analýzu následovanou pravidly řízeným překladem. Autor podrobně popisuje veškeré implementační detaily a na závěr uvádí kvantitativní porovnání úspěšnosti „svého" systému s volně dostupným čistě statistickým systémem GIZA++, který v současnosti tvoří v oboru jakýsi standard. Předložené výsledky ukazují, že pokud jsou vstupní data pro systém GIZA++ relativně jednoduchým způsobem předzpracována (lematizována), je výsledný překlad mírně lepší než u systému implementovaného autorem předložené práce. Tato skutečnost může být, jak sám autor uvádí, způsobena dvěma hlavními faktory. Za prvé, trénovací data pro čistě statistický překlad jsou strukturálně mnohem jednodušší než data potřebná pro metody využívající syntaktické analýzy (stačí paralelní text místo výše uvedeného paralelního závislostního korpusu) a je jich tudíž k dispozici mnohem větší množství, což přispívá k lepšímu natrénování systému GIZA++. Za druhé, popisovaný systém je první implementací navržené metody a jako takový obsahuje některá úmyslná zjednodušení, která nejsou intuitivně zcela korektní (například použití pouze jednoho nejpravděpodobnějšího překladu při lexikálním transferu). Vzhledem k velmi inovativnímu charakteru implementované metody jsou však tato zjednodušení snadno omluvitelná snahou o dovedení navrženého postupu do podoby funkčního programu v rozumném časovém horizontu.

Následující dvě kapitoly jsou věnovány formálnímu popisu a praktické implementaci metody pro statistický převod mezi závislostními stromy. Je zde představena zcela nová metoda učení pro struktury zvané „Synchronous Tree Substitutions Grammars" (STSG), které dokáží formálně popsat převod mezi různými závislostními stromy. Tato metoda je autorem práce nejdřív neformálně uvedena, poté precizně matematicky popsána a následně také implementována a použita v několika konkrétních aplikacích. Bohužel se už, zřejmě z časových důvodů, autorovi nepodařilo tuto novou metodu použít přímo v systému strojového překladu, což je vzhledem k hlavnímu tématu práce přece jen škoda.

Přesto však lze říci, že předložená disertační práce je velmi přínosná pro rozvoj oboru strojového překladu, neboť rozvíjí originální přístupy k problému založené na rozsáhlém využití lingvistické informace. Za nový vědecký poznatek práce lze považovat zejména prezentovanou metodu pro učení STSG. Předběžné výsledky uvedené v práci ukazují slibný potenciál této metody pro zlepšení automatického překladu.

Celkově je předložená práce na velmi vysoké formální a jazykové úrovni, s minimálním množstvím chyb a překlepů. Publikační činnost autora je dostatečná, s významným podílem příspěvků na prestižních mezinárodních konferencích.

Disertant předloženou disertační prací jednoznačně potvrdil své předpoklady k samostatné tvořivé práci a proto tuto disertaci doporučuji k obhajobě.

V Plzni dne 18.7.2006