Posudek diplomové práce Jana Ptáčka „Generování vět z tektogramatických stromů Pražského závislostního korpusu“

Obsah práce

Autor se v předložené práci zabýval úlohou syntézy českých vět na základě jejich abstraktního zápisu, jak je definován na tektogramatické rovině Pražského závislostního korpusu. Hlavním cílem bylo navrhnout a implementovat automatický softwarový systém, jehož vstupem je tektogramatická stromová struktura a výstupem řetězec věty, která svým významem vstupní struktuře odpovídá.

Práce je členěna následovně. V první kapitole autor blíže popíše zadání a motivaci úlohy a zmíní základní vlastnosti systému rovin Pražského závislostního korpusu. V druhé kapitole, která je jádrem práce, pečlivě dekomponuje úlohu na řadu podkroků (ať už lingvisticky relevantních, nebo spíše technických). Zabývá se mj. zpracováním slovesných diatezí, syntaktickými i lexikálními derivacemi, různými typy shody, složenými slovesnými tvary, způsobem výpočtu morfologických značek, doplňováním interpunkce apod. Ve třetí kapitole autor uvede některé technické aspekty implementace, realizované v programovacím jazyku Perl, a vyhodnotí generované věty pomocí standardní metriky BLEU. Po čtvrté, závěrečné kapitole následuje seznam použité literatury a příloha se vzorkem vygenerovaných vět.

Hodnocení

Je nepochybné, že stanoveného cíle bylo v práci dosaženo. Autorovi se podařilo překonat vzdálenost abstraktních vstupních struktur od vět v přirozeném jazyce, vygenerované věty jsou ve většině případů dobře srozumitelné a gramaticky správné, jak je na první pohled patrné z přílohy práce. Kvalitu vygenerovaných vět potvrzuje i vysoké objektivní hodnocení pomocí metriky BLEU. Je třeba podotknout, že jde o nový výsledek, v oblasti počítačového zpracování češtiny dosud žádný srovnatelný systém neexistoval. Systém se už prakticky osvědčil v prototypu strojového překladu z angličtiny do češtiny, kde byla kromě jeho funkčnosti potvrzena i značná robustnost.

Autor při řešení úlohy prokázal schopnost analýzy komplexního problému a velkou implementační zdatnost. Kromě toho si iniciativně doplňoval znalosti v oblasti lingvistiky. Z hlediska Pražského závislostního korpusu jde mimo autorský tým podle mého názoru o prvního člověka, který dosáhl takového vhledu do tektogramatické reprezentace přirozeného jazyka.

Nad rámec požadavků kladených na diplomovou práci autor představil svou práci na pondělním semináři katedry ÚFAL. Svými výsledky rovněž vzbudil pozitivní ohlas na setkání členů mezinárodního projektu PIRE.

Připomínky

Domnívám se, že značné úsilí, které bylo vloženo do implementace systému, by si zasloužilo méně stručnosti a větší formulační pečlivost ve vlastním textu diplomové práce. Rukopisnou verzi jsem četl a autor k většině mých námětů v konečné verzi přihlédl, nicméně přesto v textu kromě nejednotného odkazování na literaturu zůstalo několik překlepů (např. str. 12 – jsem konzultovali) a rušivě znějících výrazů (např. str. 5 - problémy jsme průběžně reportovali, str. 9 - pád, který chybu zamaskuje, str. 35 - počáteční písmeno u toplevel sentence).

Závěr

Posuzovaná diplomová práce představuje významný praktický výsledek v oblasti počítačového zpracování češtiny. Doporučuji ji k obhajobě.

V Praze 30. ledna 2006

Ing. Zdeněk Žabokrtský, Ph.D.
Ústav formální a aplikované lingvistiky MFF UK