Vyjádření vedoucího doktorské disertační práce

Jan Cuřín: Statistical Methods in Czech-English Statistical Translation

Předložená práce se zabývá aplikací statistických metod na automatický strojový překlad z češtiny do angličtiny. Jedná se o jednu z prvních dvou disertací v Česku, které se touto problematikou zabývají; v této disertaci je kladen důraz na lingvistické zpracování textů v obou jazycích tak, aby vlastní statistický překladový systém pracoval co nejlépe (na rozdíl od druhé disertace (Mgr. Martin Čmejrek), která se zabývá novými metodami vlastního jádra statistického překladu). Autor při práci na této disertaci vycházel i ze své zkušenosti člena týmu statistického strojového překladu, do kterého byl pozván v létě 1999 na Johns Hopkins University v Baltimore, USA (WS'99). Autor touto disertací navázal na předběžné výsledky a první pokusy, které prováděl jako diplomant ÚFAL.

Práce je členěna do sedmi kapitol a tří příloh (které obsahují – zejména příloha A - podstatné součásti autorova vlastního přínosu k disertaci). Po úvodní kapitole následuje skutečný „úvod do statistického strojového překladu (klasickými metodami, původně vyvinutými IBM)". Ve třetí kapitole autor popisuje způsoby přípravy dat, nezbytné (a klíčové) „přísady" každého statistického strojového překladu, ve čtvrté kapitole pak způsob získání překladového slovníku, další nezbytné součásti takových systémů. V páté kapitole popisuje experimenty se statistickým strojovým překladem, které sám prováděl a ve kterých využíval různé metody a stupně lingvistického (před)zpracování dat (paralelních česko-anglických korpusů). V předposlední, šesté kapitole popisuje použitou vyhodnocovací metriku a dosažené výsledky různých evaluací zaměřených na kvalitu překladu (vč. lidsky-subjektivní). V poslední kapitole pak shrnuje dosažené výsledky práce (včetně výsledků ve formě programových nástrojů a datových zdrojů). Práce je doplněna řadou velmi dobře připravených a relevantních obrázků a tabulek. Autorovým vlastním přínosem jsou zejména kapitoly 3, 4, 5 a přílohy A.2, A.3 a B. Je třeba ocenit to, že autor uvedl (na rozdíl od řady i renomovaných článků z poslední doby) i příklady výstupu ze systému překladu (příloha C).

Autor dosáhl velmi dobrých výsledků zejména ve spojení nástrojů vyvinutých na WS'99 a vlastního (lingvistického) předzpracování textu (BLEU skóre 0.365), a prokázal tak, že vyšší stupeň lingvistického předzpracování při zapojení do standardního „state-of-the-art" statistického systému pomáhá takto měřené výsledky zlepšit. Zároveň prokázal, že těmito metodami lze vytvořit lepší překlady než komerční systémy, a to při subjektivní evaluaci lidmi (tj. bez zkreslení jistou nevyvážeností automatické evaluační metriky).

Hodnocení: Za nejpodstatnější přínos(y) autora k problematice strojového překladu považuji

- vytvoření fungujícího („end-to-end") systému strojového překladu z češtiny do angličtiny s velmi dobrou kvalitou, který bude sloužit jako základ pro další zkvalitňování překladu pro případné pokračovatele;
- vytvoření překladových slovníků, které je možno vzít za základ a používat i v jiných než čistě statistických systémech překladu;
- významný podíl na vytvoření prvního paralelního česko-anglického korpusu (navíc bohatě lingvisticky zpracovaného) jako zdroje dat pro další statistické experimenty (ve spolupráci s řadou dalších);
- podíl na vypracování prvního volně dostupného softwarového prostředí (souboru nástrojů) pro statistický strojový překlad (jako součást WS'99)
- soubor pravidel (uvedený v Příloze A.3) pro lingvistické předzpracování textu (prokazatelně zvyšující kvalitu překladu), který bude sloužit jako inspirace i pro další podobné pokusy se strojovým překladem;
- a v neposlední řadě vysokou pedagogickou hodnotu vlastní disertační práce, která bude nepochybně sloužit pro další zájemce o uvedenou problematiku jako velmi srozumitelný a názorný text.

O kvalitě autorovy práce a zájmu o ni svědčí i již šest recenzovaných publikací (vesměs se spoluautory) na nejvýznamnějších světových konferencích, čtyři články v časopisech (PBML, IJCL) a samozřejmě i podíl

na CDROM s Pražským česko-anglickým závislostním korpusem, vydaným v Linguistic Data Consortium v USA.

Práce je psána velmi pěknou angličtinou, tok informací v práci je velmi logický a zejména v úvodních kapitolách i učebnicově didaktický (ve spojení s grafikou a tabulkami). Ani po formální stránce (citace, literatura, titulní strana apod.) k ní nemám výhrady.

Závěr: celkově práci považuji za vynikající příspěvek k uplatnění kombinace lingvistických a statistických metod ve strojovém překladu a doporučuji tedy, aby byla přijata a obhájena jako práce disertační.