Oponentský posudek diplomové práce

Miroslav Týnovský: Využití lingvistických informací při EBMT

Obsah práce

Předložená práce je zaměřena na jeden z přístupů ke strojovému překladu, konkrétně na tzv. strojový překlad založený na příkladech (Example-Based Machine Translation, EBMT), a to ve směru čeština-angličtina a čeština-němčina. Práce je členěna takto: po úvodní kapitole následuje kapitola s vymezením EBMT, třetí kapitola popisuje použitá česko-anglická a česko-německá paralelní trénovací data a jejich předzpracování, čtvrtá kapitola představuje autorem implementovaný překladový systém, pátá kapitola zmiňuje lingvisticky motivované možnosti vylepšení tohoto systému a související experimenty, šestá kapitola je závěrečná. Dále práce obsahuje přílohu s ukázkami automaticky vygenerovaných syntaktických stromů pro češtinu, angličtinu a němčinu, přílohu s ukázkami automatických překladů, přílohu se stručným návodem pro použití překladového systému a přílohu s popisem adresářové struktury přiloženého CD. Celý text včetně poděkování, obsahu, abstraktu a seznamu použité literatury má 59 stran.

Hodnocení

Na prvním místě je třeba uvést, že diplomant si zvolil velmi náročné téma, a to především co do implementační pracnosti. Z vlastního jádra zadání - tedy vylepšení metody EBMT pomocí lingvistických znalostí - vyplynula i potřeba nejdříve vytvořit samotný překladový systém, na kterém bude vůbec možné případná vylepšení testovat. Bylo nutné vyřešit řadu dalších dílčích úloh, jako například shromáždit a předzpracovat potřebná data pro oba jazykové páry (tj. provést konverzi formátů, aplikovat nástroje pro automatickou morfologickou a syntaktickou anotaci), dále šlo o implementaci extrakce slovníků z paralelních korpusů, implementaci tří základních kroků metody EBMT, hledání optimálních parametrů vzniklého překladového systému, aplikace nástrojů pro měření shody automatického překladu s referenčními překlady atd.

Není proto překvapivé, že pro vlastní lingvistická vylepšení zbylo méně prostoru. Z šesti navržených možností, které autor stručně popisuje v páté kapitole a které kombinují rovinu jazykového popisu s některou z fází EBMT, byly realizovány pouze dvě. Nicméně experimenty s využitím morfologických a syntaktických příznaků ve fázi matchingu diplomant pečlivě vyhodnotil, a to nejen standardními metrikami, ale i pomocí dvou lidských anotátorů.

Kladně hodnotím také to, že diplomant přistoupil k zadanému tématu kriticky. Ve druhé kapitole se nejprve snaží vymezit pojem EBMT oproti pravidlově a statisticky založeným přístupům, ale po pečlivém rozboru charakteristických vlastností jednotlivých přístupů dochází na stranách 15 a 16 ke zjištění, že rozdíl mezi EBMT a současnými frázovými statistickými metodami v podstatě není dobře zřejmý. Zde by ale bylo možné jít ještě dál: z dnešního pohledu lze EBMT považovat pouze za speciální případ frázového překladu - základní myšlenka je zde totožná, ovšem potenciál EBMT pro využití statistické bohatosti trénovacího korpusu je ve srovnání s moderními frázovými metodami dramaticky degradovaný. Tím se zcela vysvětluje, proč kvalita překladu diplomantova systému zdaleka nepřekonala aktuální výsledky na poli strojového překladu. S omezením na klasickou definici EBMT to ani nebylo možné.

Diplomová práce je přehledně strukturovaná a psaná velmi srozumitelně. I při pozorném čtení lze narazit jen na malé množství chyb a nedostatků, například na určitou formulační neobratnost v zapojování anglických termínů do české věty (např. "anglickou sadu evaluation dat", str. 40), "můj" místo "svůj" (str. 23), chybějící interpunkci (str. 26), překlepy ve jménech (jeden na str. 18, další v seznamu literatury), nevhodně vnořené závorky (str. 22), drobnou nejednotnost ve formátu položek v seznamu literatury nebo popiskách tabulek. Věcně bych si pak dovolil vytknout především absenci

konkrétnějších měření týkajících se rychlosti systému, která byla pro provádění experimentů zjevně nejvíce omezujícím faktorem. Po formální stránce by také bylo vhodnější na několika místech doplnit referenci, například u blíže neurčené zmínky o výzkumu DLT skupiny v Utrechtu (str. 12), o Carollově úhlu podobnosti (str. 18) nebo o McDonaldově parseru (str. 26). Co se týká obsahu přiloženého CD, nepraktickou nevýhodou je nutnost uživatelských úprav zdrojového kódu jednotlivých nástrojů (na místě by zde bylo spíše řešení s přepínači na příkazové řádce nebo systémovými proměnnými) a omezení na platformu vyplývající z použití Perl Storable.

Doplňující otázky

Rád bych diplomanta požádal o vyjádření ke dvěma otázkám.

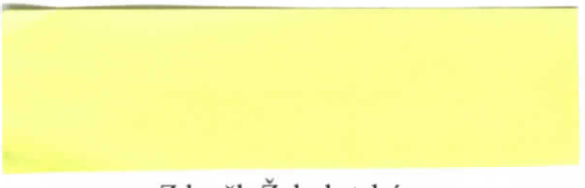
(1) Na str. 20 je uvedeno, že výhoda systémů EBMT oproti systémům statistického strojového překladu spočívá v transparentnosti, konkrétně v přímočarých opravách chyb pomocí jednoduchého rozšíření trénovacích příkladů. Potvrdilo se toto tvrzení při vývoji prezentovaného překladovém systému?

(2) Předložený systém vykazuje značnou nestabilitu ohledně časových nároků na překlad věty. Bylo by možné tyto nároky nějak radikálně snížit?

Závěr

Diplomant v předložené práci potvrdil, že dovede samostatně pracovat na náročném tématu. Prokázal rozhled v oblasti strojového překladu i značnou programátorskou zdatnost. Práce odpovídá zadání a prezentace dosažených výsledků je po jazykové i formální stránce velmi kvalitní. Práci doporučuji přijmout k obhajobě.

V Jahodově, 30. srpna 2007



Zdeněk Žabokrtský
Ústav formální a aplikované lingvistiky
MFF UK, Praha