Oponentský posudek dokorské disertační práce

Ondřej Bojar: Exploiting Linguistic Data in Machine Translation

Struktura práce

Předložená disertační práce se zabývá strojovým překladem s důrazem na překlad z angličtiny do češtiny. Nosným tématem je hodnocení přínosu využití lingvistických zdrojů pro statistický strojový překlad. Po krátké úvodní kapitole o vztahu teorie, aplikací a dat následuje kapitola o valenčních slovnících, kde autor představuje dva valenční slovníky slovesných rámců VALLEX a PDT-VALLEX a navrhuje několik metod automatické extrakce valenčních rámců z korpusu spolu s metrikou hodnotící, kolik práce lexikografa je potřeba na „doladění" automaticky navrhovaného rámce. Třetí kapitola práce představuje systém strojového překladu využívající hloubkovou analýzu vět v rámci teorie funkčního generativního popisu. Transfer mezi jazyky tedy probíhá konverzí struktur na analytické nebo tektogramatické rovině současně s překladem lexikálních jednotek a jim příslušných anotací. Autor provádí množství experimentů s cílem zhodnotit, zda je možné za použití dostupných lingvistických zdrojů dosáhnout zlepšení kvality překladu. Čtvrtá kapitola popisuje aplikaci tzv. frázového strojového překladu, který je v současnosti pro mnoho jazykových párů nejúspěšnější metodou, na překlad mezi angličtinou a češtinou. Autor navrhuje paralelní začlenění modelů s morfologickou anotací a lematizací a ukazuje, že multifaktorový překlad zlepšuje kvalitu překladu. Pátá kapitola shrnuje poznatky, které autor při práci získal, a porovnává je s výsledky dosaženými na jiných pracovištích. Příloha obsahuje ukázky výstupů z porovnávaných překladových systémů včetně výsledků automatické evaluace (metrikou BLEU).

Přínos práce

Práce se zabývá aktuálními tématy, lze konstatovat, že jde o první ucelený experiment použití statistických metod pro strojový překlad ve směru z angličtiny do češtiny; směr z morfologicky chudšího do morfologicky bohatšího jazyka je obecně těžší a u statistiků méně oblíbený.
Autor sestavil (i s využitím komponent a systémů z citovaných zdrojů) a otestoval dva nezávislé systémy strojového překladu. Výsledky prvního systému, založeného na strukturní analýze vět, bohužel ukazují, že využití hlubší analýzy, tedy transfer na analytické, popřípadě tektogramatické rovině, nevede k zlepšení výsledného překladu. Naopak, čím komplexnější systém je, tím jsou výsledky překladu (podle automatického hodnocení pomocí metriky BLEU) horší. Jako hlavní důvody tohoto pozorování jsou uvedeny: kumulace chyb při analýze vstupního jazyka i při generování výstupní povrchové formy (metrika BLEU hodnotí povrchovou, nikoliv hloubkovou strukturu), nedostatečné množství a kvalita trénovacích dat (anglická anotace na analytické a tektogramatické úrovni byla provedena automaticky) a kombinatoricky se rozšiřující prostor pro nalezení optimálního výstupu. Nespornym přínosem je sestavení tohoto komplexního systému s dekodérem, který je schopen v takto širokém prostoru vůbec nějaká řešení nalézt a navíc je schopen brát v úvahu i další modely, včetně n-gramového jazykového modelu a modelu, který vychází z pravděpodobností hran ve stromové struktuře cílového jazyka.
Za hlavní přínos této práce považuji druhý překladový systém, který vznikl rozšířením veřejně přístupného překladového systému Moses vytvořeného (za osobní spoluúčasti autora) během

letního workshopu na Johns Hopkins University v roce 2006. Začleněním modelu s morfologickou anotací a lematizací bylo dosaženo zlepšení kvality překladu z angličtiny do češtiny v rámci celosvětově již tak velice úspěšného systému frázového překladu (angl. phrase-based MT). Je třeba ocenit, že se autor nebál účasti na mezinárodní soutěži ACL WMT v letech 2007 a 2008, kde dosáhl výborných a v určitých kategoriích dokonce nejlepších výsledků.
V práci týkající se valenčních slovníků je přínosem vytvoření korpusu „Golden VALEVAL", který obsahuje anotaci 108 sloves v 7800 větách a který, kromě původního účelu, tedy hodnocení účinnosti navrhovaných metod pro automatickou extrakci valenčních rámců, pomohl odhalit i nemalé množství chybných nebo nekonzistentních anotací v původním VALLEXu. Nesporným přínosem je i podíl na vytvoření nových paralelních česko-anglických datových zdrojů a anotací: dvou verzí paralelního korpusu CzEng, kolekce překladových slovníků a manuální anotace párování slov v části paralelního korpusu.

Poznámky, připomínky a dotazy

Práce je psána strukturovaně, tok informací je logický. Nosné kapitoly (2, 3 a 4) jsou bohatě doplněny citacemi prací v dané oblasti.
V kapitole o valenčních slovnících (kapitola 2) je uvedeno, že nejúčinnější z navrhovaných metod pro automatické vytváření slovesných valenčních rámců by měla podle zavedené metriky ušetřit 67,8 % práce lexikografa, což je však jen o málo lepší než použití jednoduché metody, která pokaždé navrhuje rámec se dvěma aktanty (aktor a patient) s úspěšností 65,3 %. Nebylo by tedy vzhledem k principu posouvání (shifting), který je zmíněn v kapitole 2.2.2 na straně 19, vhodné upravit zavedenou metriku tak, aby přidání uvedených aktantů (aktor a patient) mělo nulovou váhu? Tedy tato dvě vnitřní doplnění v každém novém rámci předvyplňovat? Lze spekulovat o tom, zda ušetření editační práce ulehčí i intelektuální námaze lexikografa při verifikaci automaticky navrženého rámce.
Ve třetí kapitole o systému strojového překladu, který využívá hloubkové analýzy vět, autor navazuje na metody pro konverzi závislostních stromových struktur popsané v předešlých pracích (Eisner, 2003 a Čmejrek, 2006) a navíc zavádí dvě omezující podmínky na strukturu podstromů (dále treeletů). Jedna z podmínek si vynucuje přítomnost alespoň jednoho vnitřního uzlu v treeletu, t.j. uzlu s lexikálním obsazením. Nemůže při trénování, t.j. párování treeletů ve dvou paralelních stromových strukturách nastat situace, že treelet v jednom stromě nelze namapovat na žádný treelet ve stromě paralelním? Jak dopadne párování treeletu se slovy, která jsou podle GIZA++ spárována s tzv. nulovým slovem? Je zřejmé, že v případě překladu (dekódování) tento problém nenastává, resp. řeší ho záchranné (back-off) metody. Pro úplnost by bylo zajímavé uvést statistiku velikostí treeletů v trénovacím korpusu (co do počtu vnitřních a hraničních uzlů).
Zajímavou myšlenkou, kterou autor uvádí v kapitole 4.1.1 jako motivaci pro snahu o zlepšení „morfologické úrovně" přeloženého textu, je vypočítat hodnotu BLEU pro lematizovanou formu výstupu vůči lematizovaným referenčním překladům (resp. jednomu referenčnímu překladu v našem případě). Příloha A obsahuje ukázky výstupu různých překladových systémů, včetně systému Google translate, který byl zveřejněn jen pár týdnů před dokončením této práce. Začlenění ukázek výstupu dává čtenáři dobrý obrázek o tom, jak systémy fungují, a je třeba podotknout, že fungují překvapivě dobře. Bylo by zajímavé pro jednotlivé ukázky dopočítat i lematizované BLEU, to by v určitém smyslu kategorizovalo jednotlivé systémy podle úspěšnosti v morfologii a v samotném překladu.

Závěr

Disertační práce O. Bojara je podle mého názoru kvalitativně na vysoké úrovni. Zkoumaná problematika strojového překladu je v současnosti vysoce aktuálním tématem, autor je schopen svými výsledky konkurovat nejlepším systémům v této oblasti. Danou problematiku zkoumá v širším kontextu a v případech, kdy se mu nedaří dosáhnout výrazných zlepšení, se snaží tato pozorování náležitě zdůvodnit a podložit citacemi výsledků prací jiných autorů. Autor jednoznačně prokazuje schopnost samostatné vědecké práce. Doporučuji tedy, aby předložená práce byla přijata a obhájena jako práce disertační.

V Praze dne 28. července 2008

RNDr. Jan Cuřín, Ph.D.
IBM Česká Republika, spol. s r.o.
IBM Research,Voice Technologies and Systems
V Parku 4, 148 00 Praha 4