Posudek diplomové práce

Posudek vypracoval: RNDr. Ondřej Bojar; bojar@ufal.mff.cuni.cz
ÚFAL MFF UK, Malostranské náměstí 25, Praha 1, 118 00

Název diplomové práce: Ways to Improve the Quality of English-Czech Machine Translation
Řešitel: Martin Popel
Vedoucí: Ing. Zdeněk Žabokrtský, Ph.D.
ÚFAL MFF UK

Diplomová práce Martina Popela si klade za cíl odhalit nejvážnější chyby ve strojovém překladu produkovaném systémem TectoMT a pokusit se je opravit. Hned v úvodu zdůrazňuji, že tento cíl je bez výhrady dosažen, přestože finální kvalita překladu zůstává i nadále … poněkud strojová.

Péči, jaké se TectoMT dostalo díky diplomové práci M. Popela, by si zasloužil každý systém strojového překladu. Martin Popel nejen pečlivě identifikoval chyby a opravoval je, ale mimoděk také velmi detailně zdokumentoval předchozí i nový stav TectoMT. Následující seznam proto představuje spíše doporučení pro budoucí výzkum než výtky k předkládané práci.

- Autor se oprávněně vyhýbá slovu „significant" a hovoří o podstatných (substantial) zlepšeních. U udávaných hodnot BLEU a NIST by se totiž patřilo uvést (empirické) intervaly spolehlivosti. Větší část pozorovaných zlepšení by pak zejména s ohledem na malou množinu testovacích dat byla statisticky nesignifikantní.
- K autorově vlastní implementaci lematizace mám hned dvě připomínky: (1) nevšiml jsem si, jak je řešena víceznačnost vstupu, např. identifikace, zda slovo na začátku věty je vlastní jméno, a potřebuje-li tedy lemma s velkým písmenem, a (2) bylo by vhodné vlastní lematizaci vyhodnocovat i na ručně anotovaných datech, třeba velmi malého rozsahu.
- Zajímavé by bylo vědět, jakého zlepšení bylo v HMTM dosaženo díky převěšování na efektivní rodiče (kapitola 6.2.5).
- Popis HMTM je velmi pěkný a zvláště oceňuji rozbor omezení daného přístupu. Bylo by možné a užitečné trénování HMTM rozšířit o např. algoritmus expectation-maximization pro zpřesnění odhadu přechodových a emisních pravděpodobností místo stávajícího podílu frekvencí?
- Jako postesknutí spíše než věcnou připomínku zmiňuji, že řada provedených vylepšení jsou „nesystematické" korekce, jejichž nedostatkem je neúplnost pokrytí reálných případů. S tím se ale dá málo dělat, a vítám každý krok směrem k datově-orientovaným metodám, jako např. realizované HMTM.
- Do diskuse přidávám i připomínku autorům PEDT k pozn. pod čarou č. 33 na str. 47: Proč konstrukce „I can't not obey" není řešena speciální hodnotou deontmod raději než výjimečnou strukturou?

Práce je velmi přehledná a psaná dobrou angličtinou, neobsahuje prakticky žádné typografické chyby a velmi málo chyb jazykových (namátkou str. 67: „effective" místo „efficient implementation", „an usual" a „an universal" místo krátkého tvaru neurčitého členu).

Jednoznačně diplomovou práci Martina Popela **doporučuji k přijetí** a navrhuji celkovou známku **výborně**.

V Praze dne 26. 8. 2009,

Ondřej Bojar