

Petra Barančíková:

Targeted Paraphrasing of Czech Sentences for Machine Translation Evaluation

Oponentský posudek

Hlavním cílem předložené práce je vytvořit softwarovou implementaci automatické metriky pro měření kvality strojového překladu, která by umožnila přesnější odhad kvality překladu na základě schopnosti nalézt parafráze k referenčnímu překladu.

Celkové hodnocení:

- Práce je dobře motivovaná, jasně strukturovaná, experimenty jsou promyšlené a prezentace jejich výsledků je přehledná.
- Práce je psána srozumitelně a dobrou angličtinou, nicméně v textu zůstalo přehlédnuto asi deset zjevných jazykových chyb (např. shoda podmětu s přívlastkem), případná další publikace textu by si tedy vyžadovala jazykovou korekturu.

Komentáře a otázky k obhajobě:

- Spatřuji jistou překvapivost použité argumentace v tom, že za hlavní problém lidského hodnocení kvality překladu je v úvodní kapitole prohlášena nízká anotátorská shoda (tedy nízká kvalita “měření” kvality překladu člověkem), přičemž v navazujících sekcích je korelace s lidským hodnocením naopak jediným kritériem pro srovnávání kvality automatických metrik a je měřena s přesností na tři platné číslice. Bylo by (alespoň teoreticky) možné doložit vyšší kvalitu nově navržených metrik ještě nějak jinak?
- Východisko předložené práce spočívá v tom, že klasické automatické metriky nedovedou rozeznat situaci, kdy automatický překlad není řetězcově identický s referenčním překladem, ale je jeho obsahově totožnou nebo blízkou parafrází. Nebylo by ale možné řešit tento problém jednodušším způsobem tak, že by se k vyhodnocení použil řádově větší vzorek paralelních dat než je dnes obvyklé? CzEng je asi o pět desítkových řádů větší než evaluační sady používané ve WMT. Šum zanesený do měření tím, že daný překladový systém použil v překladu dané konkrétní věty jinou alternativu než automatický překladač, by se při použití velkých testovacích dat mohl víceméně vyhladit.
- Očekával bych, že množství možných parafrází roste zhruba exponenciálně s délkou věty. Vedlo to při experimentech založených na parafrázování k problémům?
- Na konci sekce 5.1.2 jsou uvedeny hodnoty skóre ve frázové tabulce, v poznámce pod čarou 6 je řečeno, že tato skóre jsou považována za “log-probabilities”, jak ale může mít logaritmus pravděpodobnosti hodnotu e ?
- Prováděla autorka frekvenční analýzu typů parafrází, se kterými klasické překladové metriky “zápasí” nejčastěji? Bez empirického podkladu bych například očekával, že odlišná přichylnost češtiny a angličtiny k trpnému rodu sloves povede k tomu, že půjde o jednu z nejčastějších situací, kdy by možnost parafráze s opačným slovesným videm přišla vhod. Může některý z postupů popsaných v článku zachytit parafrázi způsobenou změnou slovesného rodu?

V Praze 18. května 2015

doc. Ing. Zdeněk Žabokrtský, Ph.D.