Posudek *oponenta* diplomové práce

Jméno a příjmení autora posudku: Mgr. Michal Novák

Jméno a příjmení autora práce: Duc Tam Hoang

Název práce: Pivoting Machine Translation for Vietnamese

Vlastní text (sem prosím napište text posudku, délka textu posudku není omezena):

Thesis description

The goal of the presented thesis is to create a statistical Machine Translation (MT) system between Czech and Vietnamese in both translation directions. A particular focus is devoted to so-called pivoting methods, in which the translation is performed indirectly over a third, pivot, language.

The thesis is structured as follows: An introduction in Chapter 1 presents a motivation for conducting machine translation between Czech and Vietnamese and the goals of the thesis. Chapter 2 gives a brief introduction to SMT and pivoting methods and shows basic features and distinctions of the languages involved. In Chapter 3, the author presents the parallel and monolingual corpora required for SMT methods. These also include collections of noisy nature, such as film subtitles and conference talks. The methods used to normalize and filter such data are described in Chapter 4. The pivoting methods, elaborated in Chapter 5, are evaluated and analyzed in experiments in Chapter 6. Conclusion and Bibliography wrap up the thesis. The thesis consists of 68 pages including Bibliography and List of Figures and Tables. No CD is attached.

Comments

The thesis is written in very good English , using a reasonably wide vocabulary, such as in the sentence: "... the English-Vietnamese corpus size slightly dwindles … and then it plunges" (p. 48).
On the other hand, I also spotted a non-marginal amount of typos (e.g., frequent missing space after words "Vietnamese" and "Czech"; p. 17), grammar errors and slips of the pen, e.g., "Based on the *natural* of the raw data" instead of "nature" (p. 18), "former … *later*" instead of "latter" (p. 52). One of the misprints is quite funny – supposedly, MT "has been one of the earliest and most active areas in NLP research since 1700s."
However, the problem with this kind of mistakes becomes more serious as soon as they concern language names in translation pairs, names of phrase tables and systems. For instance, it took me a few minutes to decipher the paragraph explaining Table 6.11 on p.52-53. I was not sure whether the names of phrase tables ("PT041", "PT044") or the language pair ("English-Vietnamese") was incorrect. The latter was true; "English-Czech" should have appeared there. In addition, it seems to me not a clever idea to denote systems and phrase tables by numbers. Such an approach is prone to errors for the author, and it makes the text more difficult to read since a(the) reader is forced to constantly look at the legend (on p. 41, 49).
Presenting related works on pivoting techniques in Section 2.2.2 (p. 10), the methods of phrase translation and phrase triangulation are described chaotically, so one gets the impression that the two methods are in fact identical. This is rectified later in Chapter 5.
Regarding tables and figures, some numbers that I expected to be the same do not agree, e.g., the size of unreviewed unnormalized English-Vietnamese corpus in Table 6.5 (p. 44) agrees with none of the numbers in Table 3.3 (p. 20), neither in the number of sentences nor tokens. What is more, in Table 6.4 (p. 43), presenting the English-Vietnamese corpora again, the column denoted

as “Line count” does not agree with any of the figures in the former two tables.
Furthermore, what does the red dashed line denoted as “baseline” mean in Figures 4.1 (p. 24) and 4.2 (p. 25)? I could not find it explained. In Figures 6.2 and 6.3 (p. 47) only the scale for BLEU scores is present for the corpus size it is missing.
Most of these errors could have been avoided if the thesis was proofread.

As for the content and structure of the thesis, it leaves me with very good impression. The text is well-structured, the author's own work is clearly separated from the rest of the thesis. The thesis proposes and evaluates direct baseline systems in each translation direction between Czech, English and Vietnamese, 10 different pivoting methods and their variants, each one in both translation directions of Czech-Vietnamese, an experiment on phrase table interpolation and several experiments testing the effect of corpus normalization and filtering. Most experiments are accompanied with analyses of the results that shed more light on possible reasons of their performance. The impressive number of possibilities explored and experiments performed shows that the author has deeply immersed himself into the topic.
A great deal of manual work has been carried out by the author. Especially, I find the translation of 3,000 sentences from WMT'13 test set into Vietnamese admirable and beneficial for potential future research on Vietnamese MT. Do you plan to release the corpus? Furthermore, for the experiments on supervised filtering two sets of 1,000 English-Vietnamese and Czech-Vietnamese sentence pairs must have been manually labeled with the indication whether the sentences in a pair are mutual translations or not. I expect this portion of data had been annotated by the author, which was not sufficiently emphasized in the thesis, though. Or am I wrong?
On the other hand, all the MT performance scores are reported using the automatic BLEU measure only, although according to the thesis' specification, manual evaluation should have been conducted as well. I recommend comparing manually at least some systems on a small-sized dataset and presenting it during your defense to confirm/refute the findings based on BLUE scores and to fulfill what has been promised.
When describing the phrase table triangulation method using the co-occurrence count of phrase pairs, the author mentioned several times that different sizes of the source and target corpora negatively affect probabilities calculated from the merged phrase table. Do you think that normalizing the counts by the sizes of corpora would solve this issue and how similar it would be to the other approach of obtaining these probabilities you present – the probabilistic one?

Conclusion

Despite the formal shortcomings of the thesis, its content shows that the author is capable of and, moreover, keen on conducting scientific research on his own. Except for the manual evaluation, which I recommend the author to supply for the defense, the work also fulfills the specifications. All in all, the work satisfies the requirements set on a master thesis. Thus, I recommend that this thesis is accepted for the defense.

Doporučení k obhajobě:

Z výše uvedených důvodů práci *doporučuji* k obhajobě.

Vynikající práce vhodná pro soutěž studentských prací	ANO ☒

Seznam soutěží studentských prací, viz http://www.mff.cuni.cz/studium/bcmgr/prace/

Pokud jste výše zaškrtli ANO, zdůvodněte prosím svůj návrh, případně uveďte konkrétní soutěž, pro kterou je práce vhodná (rámeček lze nechat prázdný, pokud za dostatečné zdůvodnění považujete text posudku):

Navzdory početné vietnamské komunitě v České Republice je autor dle mého vědomí první, který se zabýval strojovým překladem mezi češtinou a vietnamštinou. Autor prozkoumal a v práci popsal mnoho možných variant a až na několik formálních nedostatků je zpracována velmi kvalitně.

V Praze dne: 3. září 2015

Podpis:**

** nehodící se škrtněte (vymažte)*

*** do SISu vkládejte formulář nepodepsaný (ve formátu PDF), podpis je potřeba doplnit až na vytištěný posudek.*