Aleš Tamchyna
Lexical and Morphological Choices in Machine Translation

reviewed by Alexandr Rosen

The thesis describes a way of advancing phrase-based machine translation (PBMT) by adding a discriminative model, which uses context information from both the source and the target text. The information is expressed as linguistically motivated features, helping to resolve specific weaknesses of standard PBMT. In addition to introductory parts (on PBMT, machine learning background and discriminative translation models) and the main story about the integration and evaluation of the proposed model, the thesis describes and compares a number of variant solutions and experiments related to the topic. The proposed system is compared to the baseline PBMT system and to *Chimera*, until recently the best-performing English-to-Czech system.

Each part of the thesis demonstrates a remarkable degree of expertise and meticulous attention to detail. The extent of research and experimental work behind the thesis must be enormous. An MT practitioner will appreciate the exploration of all side tracks as alternatives to the adopted solutions or as arguments for the specific setup. All the experiments are described and evaluated in a clear, concise and fair way.

As far as I can tell, the clarity of the text and the author's command of English are superb. Typos and other mishaps are quite rare (see the few cases I spotted below). There is just one editorial issue I could bring up: although most abbreviations are explained on their first use I would have liked to see a glossary.

In the following I first focus on the impact of the work for the field and then proceed with other, more specific comments.

Phrase-based machine translation (PBMT) vs. deep learning

After comparing the BLEU figures and text samples for systems competing in the last WMT16 contest, including the system proposed in the thesis, the mention of deep learning/neural network approach in the introduction and the related systems part strikes a chord. The BLEU score of the winner (uedin-nmt) in the English-to-Czech direction was 26.3, compared to 21.65 of the proposed system. In this connection, it is hard to resist a reflection on the developments of the MT field, starting from the early optimism in the 1950s, followed by the post-ALPAC anticlimax, the rise of linguistics-based systems until the success of machine learning methods. Are there any parallels between the current paradigm change on the one hand and the advance of machine learning at the expense of rule-based systems in 1990s on the other? Would any achievements (findings, solutions) within the PBMT paradigm still be useful within a system based on neural networks?

p. 1: "[…] lexical choice and target-side morphology represent the basic challenge of MT, and we therefore believe that many of the findings in this work are more general and can be relevant even for the newly emerging approaches to MT"

- *Would you have any more specific suggestions? This looks like a crucial point: how can this work bring benefits in a different framework?*

p. 78: Perhaps a case in point: Google Translate, now based on neural networks also for Czech, translated (May 25 2017) *the most intensive mining took place there from 1953 to 1962* as *od roku 1953 do roku 1962 zde probíhala nejintenzivnější těžba* while the proposed system offered *nejvíce intenzivní těžba probíhala od roku 1953 do roku 1962*.

Other comments and questions

p. 1: “This work focusses on two problems in MT: lexical choice and target-side morphology. The first problem is the correct transfer of meaning from the source language to the target: [...] when translating a foreign word, the system should disambiguate its sense in the source language and choose a *lexeme* in the target language which best corresponds to its meaning.”

- *This looks like an assumption of 1:1 correspondence between the source and target words/lexemes, which is obviously not realistic in general.*

p. 1: “The second problem is the choice of the correct *surface form* of each lexeme. This task is mainly relevant for target languages with rich morphology where multiple surface forms can correspond to a single lexeme.”

- *Is this about analytical morphology? If so, there are morphology-poor languages (English or even French) where verbal morphology is more analytical than Czech.*

p. 21, par. 2: “For example, adjectives in the inflectional paradigm ‘jarní’ do not inflect at all: the same word form is used for all (3–4) genders, 7 cases and two numbers (singular and plural).”

- *This is only true about the feminine gender singular. Actually, there are 7 different forms in the paradigm, including one for the dual number in the instrumental case.*

p. 30, par. 3: “However, the reduction of errors when we add TectoMT is interesting.”

- *Can this be due to the principled capturing of longer-distance phenomena in TectoMT?*

p. 32–34: Translation examples

- *A nice, interesting part. Maybe syntactic structure could help in the tricky cases?*

Typos etc.

p. 5, par. 1: We focus **on** PBMT…

p. 17, par. 4: Factored MT was introduced ~~in~~ as …

p. 21, par. –2: **The** training data…

p. 24, par. –2: TectoMT is therefore ~~is~~

p. 31, the heading Data Sparsity in LMs – shouldn’t be abbreviated (the previous heading spells out Translation Model)

p. 36, par. –3: a sentence ~~in~~**is** written in German

p. 56, par. 4: Johnson et al.~~.~~

p. 60, par. –1: what ~~is~~ a suitable generalization for this task **is**

p. 79, Figure 7.2: *destruction_of equipment* probably –> *destruction of_equipment*

Conclusion

The dissertation of Aleš Tamchyna is a respectable work with a significant creative contribution. There is no doubt about its scientific level. The linguistic and formal level of work is excellent. The thesis exceeds the requirements for a dissertation and the author has clearly demonstrated his abilities for independent scientific work.

26 May 2017

Alexandr Rosen