Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

DFKI GmbH, Campus D3 2, Stuhlsatzenhausweg 3, 66123 Saarbrücken

Prof. RNDr. Jan Kratochvil CSc.
Dean
Faculty of Mathematics and Physics
Charles University
Praha 2
Czech Republic

Prof. Dr. Josef van Genabith
Multilinguality and
LanguageTechnology

DFKI GmbH
Campus D3 2
Stuhlsatzenhausweg 3
66123 Saarbrücken

Telefon: +49 (0)681 85775-5282
Telefax: +49 (0)681 85775-4700
E-Mail: mlt-sek@dfki.de
Internet: www.dfki.de

22/11/2019

Review PhD Thesis

Ing. Tom Kocmi

Exploring Benefits of Transfer Learning in Neural Machine Translation

Neural machine translation has tremendously improved the quality of automatic translation, so much so that for some language pairs and domains human parity has been claimed. At the same time, to achieve good results, NMT needs substantial amounts of training data. The training data required are human translations from which the machine is able to learn how to translate. Typically, top performing NMT models require millions of previously translated sentences to train. Such data is not available for the vast majority of the world's 7000 languages. In fact, lack of sufficient amounts of training data is the most challenging obstacle preventing the vast majority of human languages to participate in what modern language technologies are able to offer. Research on methods that help modern language technologies to achieve improved performance on low resource languages (LRLs) is therefore very important. An additional very welcome side aspect some of the techniques developed to address low resource scenarios may also be increased energy efficiency: modern deep learning technologies are often not just "data-hungry" but require substantial amounts of electricity during training. "Greener" versions of machine learning are an important research objective.

Tom Kocmi's thesis is situated in the context descried above: the objective of the thesis is to explore transfer-learning based approaches to support NMT for LRLs. Transfer learning here involves a parent model trained on a high-resource language (HRL) pair and adapting this model in such a way so as to support NMT for a LRL pair. A number of scenarios are explored in the

thesis, including sharing or no-sharing of a language between the parent NMT XX->YY and child NMT ZZ->VV, cold-start and warm-start models, where cold-start models train the parent model (and then adapt) without any recourse to information from the child model, while warm start models already use information relevant to the child model in training the parent model (e.g. using a shared vocabulary between language in the parent and child model), linguistic closeness and distance of the languages involved in the parent and child models etc. The final part of the thesis provides a detailed analysis into some aspects of the transfer learning based approaches investigated in the thesis, including negative transfer, impact of the position of the languages (here position means Source->Target) in parent and child models, the impact of data size vs. linguistic relatedness and the question whether the positive impact of the transfer based approaches explored in the thesis can be ascribed to linguistic features (from the parent model) that are transferred to the child model or whether the impact observed in the experiments is rather due to improved initialisation (compared to random initialisation).

The work reported in the thesis includes many interesting, novel and effective ideas. Many of them are in fact very simple (this is not intended to take away from the research presented in the thesis – on the contrary this is a rather "refreshing" aspect of the thesis). Even just a few years ago, many seasoned researchers in MT would not have considered these ideas as holding much prospect: training a (parent) system on one language pair and then refining it to another (which may or may not contain a language shared with the original parent system, using the fact that for languages with the same script, byte-pair encodings can in principle capture any word in the languages involved – in the worst case decomposing it into a sequence of characters), (very simple) transfer of the parent vocabulary keeping shared (sub-)words and replacing parent sub-words not in child by the "next" child (sub-)word not already used in the replacement and thereby associating it with an embedding learned for something potentially completely unrelated in another language. The approaches mentioned above are mainly the cold-start approaches and show surprisingly good performance in the experiments presented in the thesis. The warm-start approaches are slightly more "sophisticated", in that they involve shared byte-pair (or wordpieces) vocabularies between parent and child systems and for many of the language pairs investigated improve results over cold-start approaches.

The research presented in the thesis is in a positive and refreshing sense "irreverent", creative, innovative and a contribution to knowledge. It is well presented and coupled with a thorough investigation into some of the aspects involved in the chosen transfer learning settings in the Analysis section of the thesis. There is no doubt that the work presented in the thesis constitutes a PhD. This assessment is further backed up by an impressive list of international publications coming from the work reported in the thesis.
I would have liked to see a bit more discussion on the extent to which a parent model should be trained prior to transfer: should it be fully trained or not, for it to be more malleable and therefore better suitable for transfer? Part of the

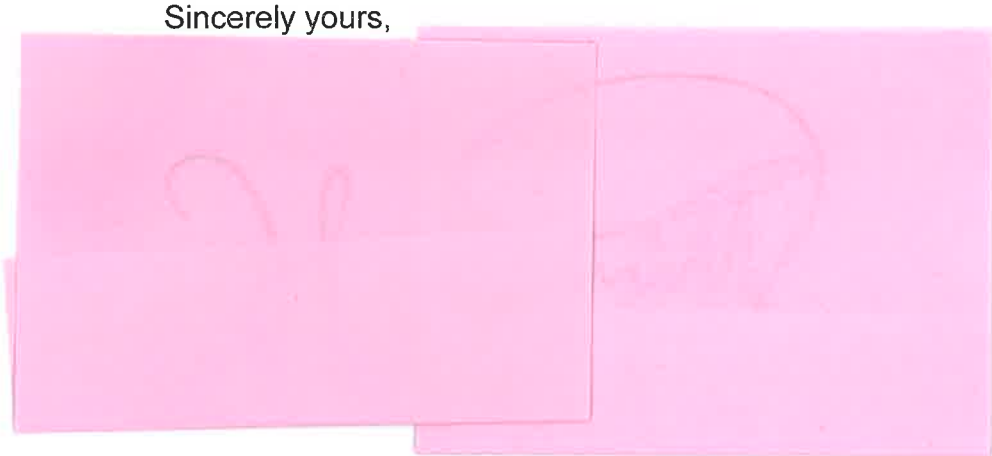
research reported in the thesis focused on another scenario, namely the use of an externally provided parent model (that has to be trained only once – at least in the cold-start scenarios). It would also have been great to see more direct comparison with the "competition" in this space, including (Zoph et al., 2016, Nguyen and Chiang, 2017, Lakew et al., 2018, Neubig and Hu 2018, Kim et al., 2019).

Having said this, again, there is no doubt that the work presented in the thesis constitutes a PhD. I very much enjoyed reading the thesis and wish the candidate every success in the upcoming defence.

References:

- Surafel M Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. Transfer learning in multilingual neural machine translation with dynamic vocabulary. IWSLT 2018. Bruges, Belgium.
- Yunsu Kim, Yingbo Gao, and Hermann Ney. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. ACL 2019. Florence, Italy.
- Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. EMNLP 2018, pages 875–880, Brussels, Belgium.
- Toan Q. Nguyen and David Chiang. Transfer learning across low-resource, related languages for neural machine translation. IJCNLP 2017.
- Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. EMNLP 2016, Austin, Texas.

Sincerely yours,



Prof. Dr. Josef van Genabith
Scientific Director
Multilingual Technologies
DFKI GmbH