10th January 2018

Opponent Report on Habilitation Thesis of Dr. Ondřej Bojar:
"English-to-Czech MT: Large Data and beyond"

This habilitation thesis is a compilation of 12 published papers authored or co-authored by the candidate, preceded by short overviews of (i) Machine Translation (MT), (ii) collecting data for MT, (iii) improving the grammaticality of MT output, especially with repect to morphological coherence, (iv) utilising deeper linguistic information, (v) MT evaluation, both human and automatic, and (vi) work done on shared tasks, which has driven forward the development of high-quality MT systems in the field.

The presentation and structure of the habilitation thesis is clear, and well organised. It is a comprehensive and coherent document detailing both the depth and breadth of the work undertaken by Dr. Bojar over a number of years, on a topic of huge importance today. While written primarily for the purposes of obtaining promotion to associate professor, the document also serves as an excellent summary of work on translation of morphologically rich languages which would be of value to any PhD student starting out on a similar topic; when one embarks on research in this area, one finds pretty quickly that data is not available, certainly not in the quantities needed to train today's large-scale machine learning-based models, so one has two choices: develop one's own resources – to be shared with the wider community – or choose a different topic. Once the data has been assembled (cf. Chapter 3, and Bojar et al. (2012b) in the Appendix to the thesis), engines need to be trained, evaluated, and improved, ultimately by competing on a level playing field with world-leading teams from other countries (cf. Bojar & Tamchyna (2013)). This thesis contains influential papers on all these sub-tasks.

While the papers selected by Dr. Bojar for inclusion in this habilitation thesis are all relevant, interesting papers in their own right, and extremely pertinent to the topic at hand, it is to the candidate's credit that the one, single paper for which he is best known is not included. Dr. Bojar was one of the developers of the Moses statistical MT (SMT) system (Koehn et al., 2007), which led MT R&D both in academia and in industry for ten years. It is a *hugely* influential paper, having over 4000 citations, more than most researchers have for *all* their papers combined! Without Moses, the field would not be at the advanced point where we are today; while it appears that SMT has had its day, and neural MT looks like being the new state-of-the-art, the Moses paper will continue to be cited for the forseeable future as most researchers are likely to provide a Moses SMT baseline in their experiments to demonstrate the improvement of their techniques over a strong, well-recognised engine.

Whenever a new paradigm comes along in any scientific discipline, it attracts newcomers to the field who are not well-versed in what research has been done before, nor what are the main problems that

people have been tackling over a long period. In Chapter 2, the candidate points to one of his papers included in the Appendix to the thesis (Bojar, 2015) which talks about how certain problems become even harder when one contemplates translation incolving a morphologically complex language. We see in Chapter 4 how Bojar (2007) explains how a factored approach to translation in such cases can be beneficial, while Bojar & Kos (2010) present a two-phase translation process to help handle unseen words. This is improved still further in Bojar & Tamchyna (2011b), where the SMT language model supports the translation model to improve the handling of such out-of-vocabulary items.

Of course, string-based methods can only get us so far, so in Chapter 5, Bojar et al. (2013c) presents the first in a long line of work that resulted in the Chimera tree-based system that operated at deeper, grammatical levels. Once one has built what one considers to be a 'better' system, this has to be demonstrated objectively using both automatic and (preferably) human evaluation metrics (cf. Chapter 6). Dr. Bojar has for some time now been a strong proponent of the use of shared tasks (cf. Chapter 7) to promote the field (cf. Bojar et al., 2016c); however, comparing different models of MT and their output is non-trivial, and both Bojar et al. (2010a) and Bojar et al. (2013b) demonstrate why MT evaluation is so difficult. Ultimately, of course, one needs to know what particular phenomena one's MT system can handle well, and what areas need further attention. Bojar (2011) reports on two techniques for capturing errors that MT systems make, noting, interestingly, that different system types make different types of errors.

In sum, it is abundantly clear that the results of Dr. Bojar's research are a very clear contribution to the area of MT. He has written a number of influential research papers, produced and shared data sets for the wider language community, and built the most influential MT infrastructure – Moses – that we have ever seen. All of this is captured in a very well-written, coherent and comprehensive habilitation thesis on the topic of MT involving a morphologically rich language. It is, therefore, my strong contention that this thesis should be accepted for habilitation, and as an outstanding researcher in the field of MT, Dr. Ondřej Bojar fully deserves to be promoted to Associate Professor.

Yours Faithfully,

Prof. Andy Way