Habilitation Committee
Institute of Formal and Applied Languages
Charles University
Prague, Czech Republic

17 September 2018

Subject: Review of the habilitation thesis *English-t-Czech MT: Large Data and Beyond* by Ondrej Bojar, Charles University, Czech Republic.

The dissertation, consisting of 198 pages including 12 publications selected by the author, presents research activities conducted by the candidate over a significant period of time, spanning from 2007 to 2016. The core subject of the thesis is machine translation (MT) into morphologically rich languages, in general, and Czech in particular. This has been and still is a relevant research topic in the field of MT, both from the theoretical and practical point of views, as witnessed by the abundant literature available on this subject.

In this review, I will go through the single chapters of the dissertation, trying to point out the main points and commenting about strong and weak aspects I found. I will conclude my review with some general opinions about the thesis and the scientific contributions of the candidate in general.

The manuscript starts with an introduction (Chapter 1) summarizing each of the following chapters. Hence, Chapter 2 briefly introduces the statistical (feature-based) machine translation approach considered in this thesis and outlines three issues related to this method: feature granularity, search complexity and quality evaluation. For each problem, directions explored by the author that are described in the thesis are sketched out and links to the corresponding chapters are provided.

Chapter 3 describes efforts made to collect and annotate English-Czech parallel data for the purpose of training and evaluating MT systems. The author focuses in particular on the progression made for the last three (out of six) published versions of the corpus, which started with 900 thousand sentence pairs and finally reached 62.5 million sentence pairs. Six papers, respectively describing each version of the corpus and spanning from year 2006 to year 2016, document this activity. My impression is that the impact of these corpora on the research community has been significant: it increased worldwide the interest on English-Czech MT, as witnessed by the participants in the WMT evaluation campaigns that have been working on this translation direction, and enabled researchers exploring other NLP tasks for Czech, e.g. co-reference resolution, or developing new language resources, e.g. dictionaries.

Chapter 4 starts by explaining, through an example, the data-sparseness problem caused by the rich morphology of Czech. In particular, only a limited number of cases of each word can be observed in a corpus. This clearly impacts on the available *translation options* that a phrase based statistical MT model can exploit when translating from English to Czech. As the correct translation of a word also depends on the context of use, data sparseness becomes even worse if we look at word-form occurrences within a specific text window. The latter issue was attacked by the candidate by investigating factored models (4.2), an extension of phrase-based SMT that tries to leverage less sparse lexical information, like lemmas and morphological tags of words. This approach has shown however limited benefit, as it only alleviates context-dependent data sparseness but not the problem of generating word forms unseen in the training data. In the following section (4.3), two-step translation is pursued, which decomposes the translation process into two consecutive phases: (i) reordering and lexical choice and (ii) morphological choices. In order to align knowledge leveraged by the language model and the translation model, back-translated monolingual target data are added to the parallel training data. This addition produced further performance improvements; in fact, language model data shows to reduce the percentage of word forms not observed in the original parallel data. Finally, discriminative models were investigated within a PhD thesis, in order to directly generated factored output representations that can be deterministically converted into word forms by a morphological analyser. All the discussed enhancements seem to help only under little-to-medium training data conditions, but stop to be beneficial as training data becomes sufficiently large. Although the partially disappointing results, I find this strand of research well conducted and useful. In terms of publications, I also count one publication at a top conference (EACL 2010).

Chapter 5 digs the data-sparseness problem further by investigating the integration of deep syntactic knowledge into phrase-based statistical MT (SMT), topic that was already explored by the candidate in his PhD dissertation. After reviewing the issues and failures of their initial tree-to-tree (T2T) approach, in Section 5.1, the author introduces the Chimera system, which combines tree-to-tree MT with phrase-bases SMT in a very simple but effective way: T2T MTs of the dev and test data are used to create additional phrase-tables which are jointly used by the SMT system to tune the feature weights and to translated the test set. In this way, the higher precision of the T2T can be combined with the higher recall of the SMT systems. Remarkably, this approach has been the state-of-the-art of English-Czech MT until the advent of neural MT. Quality comparisons of the approach with neural MT let emerge limitations of automatic metrics, such as BLEU and TER, which in general tend to unduly reward SMT over neural MT. Commenting on this chapter, I agree that this strand of research has been quite successful in terms of results and impact. I like the fact that a rather simple idea turned out to be such rewarding in terms of performance. Given these results, what I would also have liked to see are some publications in highly ranked journals or conferences, such as ACL, EACL, EMNLP or COLING.

Chapter 6 investigates automatic metrics to evaluate the quality of MT. The author claims that metrics should meet different requirements depending on their use: day-to-day assessments, MT system training, and system selection. Focusing on system evaluation, the author explains two main issues of MT evaluation: (1) the large number of very different correct translations that can be generated for a given sentence, and (2) the large number of very similar but incorrect translation that can be generated from a correct translation. An automatic metric can be evaluated by looking at how well it correlates with human quality judgements. Experimentally, the author found that correlation of BLEU show trade-offs along two dimensions: the number of used test sentences and the number of used reference translation. A more refined analysis was performed by manually tagging MT errors and looking what type of errors BLEU is sensitive to. The main finding is that BLEU seems to be less sensitive to missing content words. Another results shows that low BLEU scores (<20) give lower correlations with human judgements. The author improved this situation by extending the calculation of n-gram matches to coarser representations of both the MT output and the references. Finally, the author discusses efforts in evaluating the adequacy of translations by means of the HMEANT metric, which builds on manual annotations that permit to match the semantic structure of a sentence.
Personally, I rate these contributions well and particularly appreciate the collaborations established with external researchers, one of which resulted in a publication at EMNLP 2016. In addition, among the numerous published papers, I also point out one publication at ACL 2010.

Chapter 7 presents contributions to the evaluation campaigns organised by the WMT ACL conference. Contributions for best practice in evaluation were the identification of hidden biases in the procedure, e.g. excluding ties in scores of manual ranking of systems, as they tend to unduly favour mainstream systems over more "original" systems; improving the reliability of human scores, e.g. moving from ranking to direct assessment. Moreover, contributions were also in the co-organization of shared tasks: news translation tasks involving Czech, metrics tasks, tuning tasks, neural MT training tasks. As the author specifies, all the activities described in this chapter should be considered as service to the scientific community more than proper research activities. I shall however remark that these contributions had a significant impact on MT in general, including industrial impact, as witnessed by the high number of participants in the evaluation campaigns in more than a decade, and the high number of citations that the evaluation report papers have collected over the years.

Chapter 8 finally summarizes the main contributions of the candidate.

In general, I consider the scientific contributions outlined by the thesis consistent and impactful. After reading the thesis, I prominently see successful efforts conducted over a decade to (1) promote in the research community machine translation between English and Czech, by developing language resources, experimental frameworks to run experiments and develop machine translation systems, and (2) develop state of the art technology for English-Czech machine translation. I also positively evaluate the presented scientific publications, a precise assessment of which would however require objective criteria, which were not given to me. To my view and knowledge, the candidate, thanks to his (great) service activities within the WMT conference has definitely gained an international reputation, has co-authored several well-cited papers at international conferences, also in collaboration with international researchers, has given presentations at recognized international conferences, has significantly contributed to the development of language resources used worldwide, to the implementation of state of the art machine translation systems for English-Czech, and last but not least has participated to several successful research projects funded by the European Union. In conclusion, on the basis of the content of the thesis and my personal knowledge of the candidate, I would favourably consider his application for promotion.

Best regards,

Marcello Federico

Head of research unit (on temporary leave)
Fondazione Bruno Kessler
Via Sommarive 18
I-38123 Povo (Trento) - Italy