LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Prof. Dr. Hinrich Schütze
Chair of Computational Linguistics
Center f. Language & Information Processing
LMU Munich
Oettingenstr. 67
D-80538 München

Tel: +49 89 2180 9720
Tel: +49 89 2180 9721 (Frau Hobmaier)
Fax: +49 89 2180 9701
hs2014@cislmu.org

February 11, 2018

Opponent review of the habilitation thesis of RNDr. Ondřej Bojar, Ph.D.

Dear Sir or Madam,

It is my pleasure to provide this opponent review of the habilitation thesis entitled "English-to-Czech MT: Large Data and Beyond" by RNDr. Ondřej Bojar, Ph.D.

I have known Dr. Bojar for several years, but we have not collaborated and there are no joint publications.

The thesis is based on twelve publications that Dr. Bojar has authored or coauthored. One publication appears in the prestigious Oxford Handbook series. Two publications have appeared in "Prague Bulletin of Mathematical Linguistics", a highly regarded journal. One publication appears at the top conference for Natural Language Processing (ACL). This conference has strict reviewing criteria and acceptance is therefore seen as proof of quality in the NLP community, similar to the role journals play in other communities. Five publications have appeared at WMT, the most important – in my personal opinion – conference for machine translation. (It was initially a workshop, but was then reclassified as a conference.) Dr. Bojar's significant presence at WMT is evidence for the respect he enjoys in the machine translation community.

Several of the publications have high citation numbers according to Google Scholar. For example, "The joy of parallelism with CzEng 1.0" is cited 59 times and "English-to-Czech factored machine translation" is cited 47 times.

In the remainder of this review, I will focus on two areas of research covered in the thesis since they seem of particular relevance and impact to me: incorporating morphology and syntax into machine translation.

As the title of the thesis suggests, the topic of the thesis is machine translation, with a special focus on English-to-Czech machine translation. The difficulty of machine translation varies widely depending on source and target language. English-to-Czech machine translation is particularly difficult because Czech is a morphologically rich language. This increases the number of error sources in generation because it is easy to get morphological features of a rich morphology wrong. Most Czech features have no equivalent in English, so that the source language is a poor source of information for generation of Czech.

Dr. Bojar has published extensively on how to solve this problem. This research is summarized in Chapter 4. The key design choice is how to translate morphology. Dr. Bojar demonstrates that there is no hope in trying to translate complex forms directly in phrase-based machine translation because of sparseness, which results in unseen forms even for very large training sets. Instead, he proposes to also translate morphological tags and treat them as pseudowords. We can then use standard NLP technologies, including language models, to either check that the generated tags make sense on the Czech side or choosing the best translation based on both the translation model and the Czech language model. This is important and leading work in NLP on how to deal with translation into a morphologically rich language.

A second contribution described in the thesis concerns the use of "deep syntax" in machine translation, which means the use of any type of syntax beyond part-of-speech. Two counteracting forces make the effective use of syntactic analysis difficult. (i) Syntactic analysis would be of enormous help if it were 100% correct, but it has a relatively high error rate. (ii) If training corpora are large enough, then raw statistics are frequently sufficient for good translation. Thus, syntactic analysis is most helpful for sparse-data situations.

Dr. Bojar devises an ingenious way of employing syntactic analysis in the face of these difficulties. He uses the syntax-based system TectoMT for translating development and test data – this is a potentially big win in situations where the training data is not huge or where there is domain mismatch between train and test. He then learns a phrase table from the artificially created parallel data. This is a good strategy for making sure that a certain error rate can be tolerated because the output of parsing is postprocessed by a statistical "association" procedure.

The empirical results in the WMT evaluation campaigns are impressive: The Prague system was best 2013-2015 on the most widely used metric in machine translation and still better than Google Translate in 2016.

In summary, the work covered in the thesis, especially the two areas I have discussed in detail in this review, is of high-quality, has had international impact and constitutes an important contribution to natural language processing in general and machine translation in particular.

I hereby explicitly recommend that RNDr. Ondřej Bojar, Ph.D. be appointed as an associate professor.

Sincerely,

Prof. Dr. Hinrich Schütze