

Opponent Report on Habilitation Thesis of Pavel Pecina
"Adaptation of Machine Translation to Specific Domains and Applications"

This habilitation thesis is a compilation of published papers, preceded by a comprehensive introduction and extensive summary of the research performed by the applicant, showing the consistency of the research work pursued by Dr. Pavel Pecina.

The presentation and structure of the thesis is clear and introduces the topic in a didactic and smooth way. The topic, domain and application adaptation, is currently a very active research area. It is also of a high relevance because of the explosion of applications based on Natural Language Processing (NLP), and Machine Translation in particular, that is currently taking place. European multilingualism is considered to be a barrier to the full deployment of applications based on Language Technologies while Machine Translation is looked at as the solution. Dr. Pecina research line is exactly going in the direction of proposing how to make these technologies operatives in real world applications. His clear statements in form of practical conclusions and development guidelines demonstrate his interest for contributing to actual solutions.

Moreover, in his work adaptation is combined with the issue of lack of resources, in particular parallel corpora, which is one of the most serious problems faced by Statistical Machine Translation applications and, in general, by methods based on machine learning techniques. The work shows different domain adaptation options depending on the availability of resources, what shows a global understanding of the problems that application development faces. Dr. Pecina global understanding of the area is also demonstrated with his contributions to MT evaluation, and to the design of the MT evaluation extrinsically in other applications, such as Crosslingual Information Retrieval.

The results of his research are a clear contribution to the area. In particular, I appreciate that the research done provides with actual empirical data about different practical aspects that are normally stated but not demonstrated in the literature:

- The accurate data provided about the loss of quality suffered by general Machine Translation systems when used in particular domains is of great interest. In terms of BLEU, the most used evaluation measure, on average it is of 54%.
- About 500 sentence pairs of the addressed domain can be used as development data for adapting a general MT system to a particular domain. If no parallel data, in-domain monolingual data can also be used to enrich the language model. If no in-domain data is available at all, parameter-tuning strategy can also be applied.

- Convincing data about the useless of cleaning crawled parallel texts are provided.

The results reported in the different published papers also include conclusions of negative experiments and guidelines for performing the operations, what makes his research replicable and demonstrates applicant's interest in supporting other researchers and developers.

In summary, the thesis convincingly shows that Dr. Pecina has an outstanding understanding of the area and applications of Machine Translation and Natural Language Processing. I recommend that the thesis is accepted for habilitation and that the applicant is appointed as an associate professor.



Núria Bel
Associate Professor
Department of Translation and Language Sciences
Institute for Applied Linguistics
Universitat Pompeu Fabra
Roc Boronat, 138
E-08018 - Barcelona