Jörg Tiedemann
Modern Languages
Univerity of Helsinki
jorg.tiedemann@helsinki.fi

02-01-2017

Habilitiation Thesis Review

Candidate: Pavel Pecina
Charles University, Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

It is my pleasure to provide my assessment of Pavel Pecina's habilitation thesis as requested by the Faculty of Mathematics and Physics at Charles University in Prague. In the following report, I first summarize the thesis and its main contributions from my point of view before evaluating each part in more detail and concluding with my overall recommendation.

First of all, I would like to congratulate the candidate for the achievements presented in the thesis. I enjoyed reading his work that has been compiled into a coherent and scientifically sound collection of individual contributions in the field of statistical machine translation (SMT). The thesis addresses one of the most challenging issues in SMT and data-driven NLP in general, namely the adaptation of models to new domains and specific applications that are different from the training objectives and (most of the) data instances used for parameter estimation. Machine learning techniques are very successful when applied within the same domain assuming that the i.i.d. conditions hold, i.e. that random variables are independent and identically distributed. However, this is often strongly violated, which makes models a bad fit in test scenarios that describe new conditions. In SMT, domain adaptation is especially crucial because of the lack of sufficient in-domain training data in most real-world cases. Pavel Pecina describes his contributions in the field, which have been published in a variety of scientific papers that build the basis of the thesis. He includes eight publications jointly written with colleagues from various projects and provides a self-contained summary of the work connecting the individual pieces taken from previous research. In my assessment, I will focus on that summary as the papers are already properly reviewed before publication in well-established channels such as proceedings of main international conferences in computational linguistics, scientific journals and article compilations in published lecture notes.

The contributions roughly fall into two categories: (1) Domain adaptation by collecting small in-domain data sets and using them for training and tuning and (2) adaptation of machine translation to the task of cross-lingual information retrieval (CLIR) in

the medical domain. The framework is phrase-based SMT and the experiments rely on Moses, a commonly applied toolkit that implements standard models including training, development and testing procedures. The work presented focuses on the optimization of MT engines using well-established procedures rather than the development of novel algorithms or models. Three papers discuss the efficient use of Web-crawled data sets that augment general-purpose models for specific applications. One paper discusses the impact of in-domain tuning and analyses the effect on model parameters. Four papers investigate the development of MT in the medical domain for purposes of cross-lingually enhanced information retrieval. The work is embedded in major international projects and, therefore, concentrates efforts on practical solutions and real-world scenarios.

The thesis is structured in four parts: an introduction including a historical overview of MT research and development, a background chapter containing introductions to SMT and model adaptation techniques, a result chapter that summarizes the main contributions of the thesis and, finally, some brief conclusions and final remarks.

The first part focuses on historical developments and places the presented research with the chosen framework into the perspective of that development. It is not entirely clear why this part is necessary in connection with the presented work on adaptation but it forms a gentle introduction to the field especially for people outside of MT research. Slightly more relevant is the introduction of MT applications especially in relation to the adaptation for specific tasks discussed later. However, I would have expected a more explicit motivation of the problem of domain adaptation and a thorough (formal) definition of the task, which is, unfortunately missing in this part. Also, important concepts and terminology such as “domain” (as opposed to genre or register) could have been defined more explicitly. This is one of the shortcomings in the thesis I would like to mention together with the missing formulation of explicit challenges and research questions also from a more theoretical point of view.

The second part summarizes foundations of machine translation and previous work in domain adaptation for MT. The classical division of MT into rule-based, example-based, statistical and hybrid models is presented and upcoming neural MT is mentioned. Personally, I do not like the traditional classification but a discussion of the established terminology is beyond the scope of the thesis (and this assessment). To add some minor notes, I would like to point out that it may be wise to make clear what is meant by “knowledge” extracted from parallel texts and why that knowledge does not refer to rules and dictionaries in the case of SMT and EBMT. I would also claim that EBMT does not extract models from the data but rather uses the data as a model but that is yet another minor comment. Other fields like syntax-based SMT, hybrid models, system combinations, neural MT and the incorporation of linguistic information are mentioned very briefly and I wonder how valuable those parts are for the work presented in the thesis.

The following part introduces SMT, motivating the general framework and after that presenting some details of standard models. Similar to the previous section, I wonder about the relevance of presenting details like IBM models for the purpose of the thesis. Instead, important components of phrase-based SMT (the model used in all presented experiments) such as lexicalized reordering, lexical weighting, phrase penalties and word penalties are not properly motivated and explained even though

they become important, for example, in the discussions on parameter tuning. The introduction of decoding is also quite brief and I wonder if more details could have been added to better support discussions on adaptation and practical applications. In general I miss a clear connection of this part to the adaptation work presented later on. Some explicit links could have been added to stress the relevance of the presented background.

The following sections are obviously important for the thesis introducing related work in domain adaptation and MT in specific applications with the focus on the medical domain and cross-lingual information retrieval. They provide a comprehensive overview over main techniques that have been proposed in the literature. The presentation emphasizes in-domain data acquisition and its use in model adaptation, which is quite natural in a data-driven framework. As part of this, web crawling techniques are introduced. Surprisingly, domain-specific tuning is not explicitly mentioned even though tuning is extensively discussed in the contributions of the thesis. A section on cross-lingual IR concludes the background chapter. A summary that connects the background with the subsequent foreground is, unfortunately, missing.

Chapter three describes the contributions of the research and provides an overview of the results. The chapter is organized as a summary of each paper attached to the thesis in separate thematic sections. Each publication is a result of collaborative efforts but Pavel Pecina clearly states that he is the main contributor of them all. In some cases, however, I would have liked more specific information about the division of labor especially in connection with the development of tools and resources.

The first three sections present the use of web-crawled in-domain data in domain adaptation. The approach is investigated and analyzed extensively and the results are impressive and convincingly supported. Very interesting is the significant improvement that can be achieved with in-domain tuning (which is slightly contradictory to the background description in which it is claimed that domain adaptation is mainly about increasing lexical coverage). The most important contribution of this part is the detailed analyses of model parameters and how they are influenced by adaptation techniques. The experiments also provide valuable guidance for the development of real-world applications, which is very much appreciated.

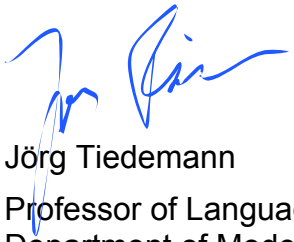
The following three sections focus on the medical domain and cross-lingual IR as the main application. The thesis clearly motivates the challenges of translating search queries (instead of grammatical sentences) and back-translating answer snippets, which are dense and more compact than common texts. The experiments are thoroughly set up and the results show the impact of adaptation on system performance. Most of the work is embedded in international evaluation campaigns and, therefore, apply well-developed procedures and are properly evaluated. The platform that has been implemented performs well and reflects professional research and development. The final approach includes a re-ranking approach that achieves a performance that is close or even extends the use of the reference translation in cross-lingual IR, which is quite impressive.

The final chapter concludes the summary with some additional remarks and further links to publications and related contributions by the author of the thesis. The latter demonstrate the strong activity of the candidate in the research community.

To sum up, the thesis presents extensive work in the field of statistical machine translation and contributes valuable ideas for adapting practical systems to real-world applications. The work has been carried out in a professional way and the collaborations show the ability of the candidate to work in teams and to lead research projects in international environments. Shortcomings in theoretical considerations are compensated by detailed analyses of the empirical results and the valuable resources and tools provided. The thesis certainly fulfills the requirement of a habilitation in computational linguistics and demonstrates the professional skills of the candidate.

With this, I have thus no reservations to recommend the thesis to be accepted by the university and to appoint the candidate as an associate professor at your institute.

With best regards,



Jörg Tiedemann

Professor of Language Technology
Department of Modern Languages
University of Helsinki
jorg.tiedemann@helsinki.fi