University of Colorado Boulder

Department of Computer Science
College of Engineering and Applied Science
430 UCB
Boulder, Colorado 80309-0430

t 303-492-3552
f 303 492 2844
james.martin@colorado.edu

Professor Jan Trlifaj, Vice-Dean
Faculty of Mathematics and Physics
Charles University

February 7, 2017

Dear Colleagues,

It is my pleasure to provide this opponent review of **Dr. Pavel Pecina's** habilitation thesis. After reviewing the materials provided to me, I can state with confidence that Dr. Pecina's research over the last 6 years has been outstanding. I fully support the proposed action to promote him to associate professor.

By way of introduction, I am a Professor of Computer Science and a Fellow of the Institute of Cognitive Science here at the University of Colorado Boulder. I currently serve as Chair of the Department of Computer Science. My research is in the area of computational linguistics with a particular focus on computational semantics and its applications to educational technology, medical informatics, and social media. I am the co-author with Prof. Dan Jurafsky of *Speech and Language Processing* (2008), the premier reference text in our field.

Dr. Pecina's work addresses fundamental problems in multilingual processing including machine translation and cross-linguistic information retrieval, with a particular focus on adaptation of statistical machine translation systems with respect to both genre and application. These are areas of research with significant impact in terms of basic language science and practical application.

The work described in Pecina et al. (2011, 2012a, 2012b, and 2015) covers a series of experiments designed to improve the performance of statistical machine translation systems (SMT) in the context of domain specific application. In particular, the focus of these efforts is on adapting generic SMT systems, trained on large amounts of general domain data, to specific domains.

The focus of Dr. Pecina's early work (2011) demonstrated that effective and relatively painless web-crawling methods could be used to acquire monolingual and parallel training data useful for adapting generic SMT systems. The follow-up experiments described in Pecina (2012a) explore **how** such data can be used to improve the performance SMT systems. In particular, Pecina and colleagues explore the relative benefits to acquiring small amounts of parallel training data to improve the translation model of a system vs. acquiring data to improve the language model with monolingual data. Here they find that small amounts of parallel data can make a big difference, while enormous amounts of monolingual data are required to assist with domain adaptation. This is a critical and non-obvious finding that will guide future system development.

Pecina et al. (2015) provides further detailed experiments that document the degree to which, and the mechanism by which, phrase-based MT systems degrade when applied to specific domains. Unlike many such data directed efforts, Dr. Pecina provides the results of a detailed investigation into the mechanisms of this degradation by analyzing the feature weights in a trained SMT system. This analysis shows that a key strength of the phrase-based approach leads to its downfall when applied to specific domains. In particular, the preference for small numbers of longer phrases during decoding, coupled with the low translation score for these phrases leads to poor translation of domain specific text. This is true even when the SMT models have access to better translations consisting of a larger number of shorter more appropriate phrases. This is an important result since it demonstrates that models trained on general corpora, in theory, have access to better translations, even in the absence of large amounts of in-domain data. This leads to the insight that proper parameter with small amounts of data can overcome the data scarcity issue in narrow domains.

Having determined the mechanisms by which SMT systems degrade in specific domains, Pecina (2015) investigates the various ways to improve language models, translation models and how they can be optimally combined. Here they find that, even with small amounts of domain specific training data, a log-linear combination of separate translation models (general and domain specific) provides significant improvement. In addition, these improvements to the translation model can be combined with improvements to the language model in a largely independent fashion. Again, these results provide a clear roadmap to practitioners wishing to apply generic SMT approaches to specific application domains.

The next body of work described in his thesis involves his work on medical applications, specifically his work with the Khresmoi Project. The first significant contributions of this work are the resources created for it. These include medically relevant queries in Czech, French and German along with a corresponding set of documents with relevance judgments. Such collections are enormously important to advancing the field. Historically, the creation of such datasets has proven to be more important than many of the initial algorithmic techniques that were proposed for use with them.

The experiments detailed in Pecina et al. (2014) represent an impressive and systematic exploration of the role of MT in cross-lingual medical information retrieval. The evaluation of systems that involve both machine translation and information retrieval components is difficult and time consuming. The interactions between the components are complex and there is often little reason to assume that component-level evaluations and improvements will lead to overall system improvements. For these reasons it is important to be extremely careful in the evaluation of such systems. The paper presents both intrinsic and extrinsic evaluations for a series of MT adaptation techniques in the form of MT translation quality and IR retrieval effectiveness for the medical domain.

The final, and most recent, area of research presented centers around the notion of re-ranking of translation hypotheses for use in cross-linguistic retrieval. Interest in re-ranking in the context of large complex systems such as MT, speech recognition and information retrieval has increased over the last decade. This is largely because it provides an opportunity to improve the performance of large monolithic systems without significant re-engineering efforts.

The work presented In Saleh and Pecina (2016) leverages the ability of current statistical MT systems to provide a ranked list of N-best translations. The information used to generate this list lies solely on the translation side of things. Pecina and colleagues cleverly use training data to re-rank the potential queries based on their precision scores in a retrieval task, thus bringing information from the retrieval side to bear on the translation. The features used to train the re-ranking system are quite diverse, ranging from features provided by the MT system to Wikipedia and concepts from standard medical informatics terminologies. The net effect of these efforts is to significantly improve retrieval performance without having to radically alter the underlying MT or retrieval engines.

Finally, I should note that Dr. Pecina has also been active in numerous efforts to organize and run "shared tasks" for various language tasks. Such efforts have become a critical driving force for progress in speech and language processing. However, organizing such tasks is an enormous and complex undertaking for which there is often little reward or recognition. Dr. Pecina deserves applause for his numerous efforts along these lines.

In sum, Dr. Pecina's habilitation thesis, presents an impressive body of work that addresses a set of important problems in multilingual language processing. The research presented makes significant contributions to the core science of machine translation as well as to important practical applications that make use of machine translation techniques. I am confident that the work presented in this thesis warrants promotion to Associate Professor – it would certainly do so at most institutions I am familiar with.

Sincerely,

James H. Martin
Professor and Chair of Computer Science
Fellow of the Institute of Cognitive Science
University of Colorado Boulder