DISSERTATION THESIS REVIEW

Applicant: Ing. Shadi Saleh
Thesis title: Cross-lingual Information Retrieval in the Medical Domain

Review structure

Faculty of information technology asked to reflect the following criteria in the review:
1) critical analysis of the thesis, listing its advantages and shortcomings,
2) up-to-datedness of the topic,
3) assessment of the methods used in the dissertation,
4) evaluation of the results, listing which contributions of the dissertation are novel,
5) formal quality of the thesis,
6) significance of the thesis for the future development of the research field and possible applications of the results in practice,
7) the overall evaluation of whether the dissertation shows that the author is able to perform scientific work independently.

Ad 1) *Critical analysis of the thesis, listing its advantages and shortcomings*
The thesis presents several contributions related to the topic of cross-lingual information retrieval. The thesis starts with a general and well-readable introduction to the topic, including detailed motivation for the focus on medical information retrieval. The text of the thesis contains many helpful examples. The main contributions of the thesis are presented in Chapter 4 "Test Collection" and Chapter 6 "Experiments". Evaluation of the contributions is under point 4 below. As I am not expert in the field of machine translation, I cannot fully appreciate the algorithmic advancements presented in the thesis. Overall, my impression is that the thesis mostly presents solid evaluations and benchmarks of existing approaches with incremental improvements and their combinations.

Weaker points of the thesis include limited availability of the code for replicating the experiments (or not stated) and some caveats in the logical organization of the content (section Experiments also contains the description of the author's approach, which would be better placed in a separate chapter). I am unable to judge to what extent state-of-the-art systems are included in the evaluation, but it seems to me that they are underrepresented. The statement "Finally, I want to emphasize that I use we in this work to refer to me and you (the reader), and it does not imply any collaborative work." (p. 6) is not always accurate, as the thesis does feature some collaborative work, e.g., the content of chapter 4 seems to be partly based on the work [Saleh and Pecina, 2019a], cf. e.g.,"... The extended dataset contains a total of 38, 109 document-query pairs, 14, 368 pairs of them are assessed by us." on p. 48.

Ad 2) *Is the topic of the thesis up-to-date with respect to current research.*
Regarding the state-of-the-art character of the research, cross-lingual information retrieval is a current area of research. The main contribution of the thesis is achieved using neural network language models, which is a state-of-the-art approach.

Ad 3) *Assessment of the methods used in the dissertation*
Given the fact that the thesis makes contributions in multiple fields, it uses a variety of research methods. As the main method to achieve the first contribution (new dataset), the author had to organize annotation of linguistic content. The quality of annotations is measured using Cohen's Kappa, which is a standard approach.
To prepare the dataset, the author applied several preprocessing algorithms (such as HTML-Strip and Boilerpipe). The queries used in the work were adapted from the CLEF contest, in which the

author participated several times. CLEF (Conference and Labs of the Evaluation Forum) is an internationally recognized event mostly focused on information retrieval evaluation.
For the experiments, the author used both statistical machine translation and neural machine translation algorithms. The experiments utilized existing systems such as UDPipe, eman and Marian library, some of which are developed at the author's institution. The use of existing frameworks improves transparency and repeatability of the experiments, as well as facilitates reuse. Some of the results are also evaluated for statistical significance (Wilcoxon signed-rank test).

Ad 4) *Evaluation of the results, listing which contributions of the dissertation are novel*
The thesis makes several incremental contributions, I've found the following ones the most notable:
- Survey and application of useful methods for text cleaning (section 4.1).
- A new dataset included in the Lindat/CLARIN repository.
- Replication of work of Dušek et al. [2014].
- Development of an SMT system that produces lemmatized sentences.

Possibly the most interesting result for me, as a researcher with a very limited background in machine translation, is the conclusion of the main comparative study. This showed that none of the Document Translation approaches was able to outperform the Query Translation approach. As the author notes, such result was unexpected and contradicting earlier studies done in the 1990s. Overall, I find the multifaceted contributions appropriate for a PhD thesis. Contribution to the field can be evaluated by publication record, which includes eleven papers out of which in ten, the applicant is the first author. While most are workshop contributions, several papers are in proceedings of conferences published in esteemed conference series by leading publishers (ACL, Springer).

Ad 5) *Formal quality of the thesis.*
The thesis is well-written, the language is nearly flawless to the extent to which I, as a non-native speaker, can judge. I could find only a few typos, e.g., (e.g., BELU instead of BLEU on p. 76, "tuning it did not" on p. 78). The organization of the text is also logical, although sometimes, as I noted earlier, not systematic. For example, query expansion is reviewed on p. 28 and within a different section on p. 31. Part of the evaluation is present already in Chapter 5 (sec. 5.6.1), instead of the designated Chapter 6 ("Experiments").
Formulas and equations are appropriately explained with some exceptions (e.g. in Equation 3.4). Overall, the thesis satisfies this criterion.

Ad 6) *Significance of the thesis for the future development of the research field and possible applications of the results in practice.*
Accessibility of the created dataset in the Lindat/CLARIN repository allows easy reuse by other researchers. I would expect that the significance of the thesis would be even higher if the code used for the experiments is released. Since I am not expert on machine translation, I do not feel fully qualified to further judge the thesis on this criterion.

Ad 7) *The overall evaluation of the dissertation*
The author of the dissertation proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (4) of the Law Nr. 111/1998 (The Higher Education Act) I do recommend the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

Question to be answered during the defense:
In your work, you build upon query expansion using Wikipedia (via fulltext Wikipedia search). What is your opinion on future work focusing on structured data sources for query expansion, such as DBpedia, Wikidata, Babelnet or ConceptNet? In particular, the last resource can be directly used to retrieve synonyms for medical terminology.

August 4, 2020

Doc. Ing. Tomáš Kliegr, Ph.D.
Faculty of informatics and statistics
University of Economics, Prague