Supervisor's review of doctoral thesis

Charles University, Faculty of Mathematics and Physics
Study programme: Computer Science
Study branch: Mathematical Linguistics

Student: **Shadi Saleh**
Title: **Cross-lingual Information Retrieval in the Medical Domain**

Supervisor: **doc. RNDr. Pavel Pecina, Ph.D.**
Department: Institute of Formal and Applied Linguistics MFF UK

The doctoral thesis by Shadi Saleh belongs to the area of Cross-lingual Information Retrieval (CLIR), a sub-field of Information Retrieval which attempts to overcome the language barrier when users formulate search queries in a language different from the language of documents the information is searched in. The presented work focuses on a specific domain – non-expert search in medical documents (users are typically patients looking for information on some disease, symptom, or treatment in on-line documents) which becomes more and more popular among internet users.

The thesis thoroughly investigates two main approaches to this task: query translation (where queries are machine-translated to the language of documents) and document translation (where documents are translated to the language of queries). Shadi's work builds on the department's history of research projects on machine translation in the medical domain (Khresmoi, KConnect) but goes far beyond them and brings several important contributions. First, Shadi proposed two methods improving query translation: reranking of query translations and term expansion of query translations, both significantly outperforming strong baselines set by the two above-mentioned research projects. The second major contribution is a large-scale analysis and comparison of the two main approaches to CLIR (query translation, document translation) which resulted in a very interesting and unexpected findings. The third major contribution is the data set (test collection) that was created during the work and significantly extends an existing resource from the CLEF eHealth Lab series. The complete relevance assessment of all experiments presented in the thesis makes the evaluation very reliable. The data set has been published and became a valuable resource for the research community.

The thesis is well structured. It is split into 6 numbered chapters plus a (unnumbered) conclusion and several other appendices including a list of author's publications and a rich bibliography. Chapter 1 is an introduction to the thesis presenting its goals and the main contribution of the work. Chapter 2 is a brief introduction to the field of Information Retrieval describing the main retrieval models and principles of evaluation. Chapter 3 presents a research context of the thesis in a form of an overview of related work. Chapter 4 describes the test collection exploited in the experiments including details on newly added query translations and relevance assessments. Chapter 5 deals with the technologies of machine translation employed in the work. Chapter 7 presents the main experiments and findings and is followed by a conclusion, list of references and other appendices. The text is readable, written in English with infrequent errors in grammar and style. The experiments are well described with a good level of details.

Shadi's bibliography is quite rich. It contains 11 papers, all related to the work presented in his doctoral thesis. Nine of them are indexed by Scopus and some of them were presented at highly-ranked conferences (2xECIR, 1xACL).

Although Shadi’s thesis belongs to the area of Information Retrieval, most of the work that was done rather belongs to the area of Machine Translation. Shadi’s first experiments were based on the traditional Phrase-Based Statistical Machine Translation models but during the course of his work, the field observed a significant paradigm shift and now the neural models are considered the state of the art. Shadi adopted this change in his later experiments and used this new, very different technology and conducted experiments comparing two main approaches to CLIR, using two MT paradigms, something that was probably never done on such a scale. I recommend Shadi Saleh’s thesis to be defended.

doc. RNDr. Pavel Pecina, Ph.D. 20.9. 2020, Prague