TU WIEN Informatics

Report on the Doctoral Thesis "Cross-Lingual Information Retrieval in the Medical Domain" by Shadi Saleh

Allan Hanbury

This thesis covers the extremely important topic of searching for online medical information that is written in a language that is not necessarily that of the searcher – Cross-Lingual Information Retrieval (CLIR). A challenge with lower-resourced languages is usually a lack of extensive medical information written in that language on the web. While the necessary information is more likely to be available in a more highly resourced language, such as English, it is often challenging to formulate a search query in English if English is not a commonly used language of the user. The main topic of the thesis is the automated translation of queries about medical information into other languages in order to find documents these languages. Here the thesis places an emphasis on how well machine translation methods work for information retrieval (the aim is to find the translation approach that leads to the best metric for the search results) rather than how good the translation itself is (in terms of a machine translation metric). A key insight of the thesis is that improving the quality of machine translation results (based on machine translation metrics) does not necessarily lead to improved CLIR with this machine-translated result (based on information retrieval metrics).

The proposed approach also has potential for example in pharmacovigilance (to discover previously undocumented adverse reactions to medication) and in systematic reviews in medicine (to gather evidence on the efficacy of an intervention by reviewing multiple relevant randomized controlled trials). For both of these applications, finding documents written in languages other than English is potentially important – the approach proposed in this thesis allows this to be done based on a single query in English.

The thesis makes the following key contributions to the scientific field (reduced to the key contributions from the longer list in Chapter 1 of the thesis):

1. Creation of a more comprehensive test collection for CLIR in consumer health search by enhancing the CLEF eHealth 2015 test collection by queries in additional languages as well as additional relevance assessments. This collection is available for download and has been described in a 2019 publication. Releasing useful test collections is generally associated with a large impact on the research community due to the effort and expense associated with the creation of test collections (Chapter 4).
2. A minor contribution to the Statistical Machine Translation pipeline in which the Machine Translation system is trained to produce lemmatized words. This should be a better approach for the translation of morphologically poor languages, such as English, to morphologically rich languages, such as Czech, German and Swedish (Chapter 5).
3. A new approach is proposed to learn to rerank translation hypotheses from an MT system with the target of improving the search results in the CLIR system. Extensive

experiments show the improvement obtained through the use of this approach (Chapter 6).
4. Evidence is provided that in the medical domain, query translation approaches outperform document translation approaches, contradicting a long accepted result in the field of CLIR (Chapter 6).
5. A query expansion approach is proposed that includes alternative translations from machine translation, English Wikipedia abstracts, and the titles of PubMed articles. It is experimentally shown that this query expansion approaches improves search results in CLIR, as well as in monolingual IR (Chapter 6).

The thesis is divided into the following 7 chapters:

- Chapter 1 presents a very brief justification of the importance of the thesis, followed by a list of the four thesis goals, and of 9 contributions. It is followed by a description of the structure of the thesis. It would have been useful to go more into detail on the background of the thesis in this chapter, in order to create a full picture of the work for the reader, before diving into the background theory in Chapter 2 and related work in Chapter 3.
- Chapter 2 presents the theoretical background for information retrieval (IR), beginning with a useful overview of the terminology used. There is then a brief overview of consumer health search, the part of information retrieval on which the thesis concentrates. Finally, an overview of retrieval models and evaluation metrics is given – these are key parts of the standard IR theory necessary for following the work in the thesis.
- Chapter 3 presents related work of importance to the thesis. The overview of cross-lingual Information Retrieval (CLIR) is very comprehensive, covering the two main approaches to CLIR: document translation and query translation. Query translation is particularly well covered, being a key topic in the thesis. Sections are included on dictionary-based query translation, corpus-based query translation, query translation using a machine translation system, and neural approaches. This is followed by succinct overviews of Query Expansion in IR and in CLIR. The chapter is concluded by a very detailed overview of test collections and evaluation campaigns for CLIR, including short overviews of CLIR in the main evaluation campaigns: TREC, NTCIR, FIRE, and CLEF. Finally, the evaluation tracks producing the main data used in the thesis are described: CLEF eHealth 2013–2018. The chapter conclusion section gives a very brief summary of the chapter.
- Chapter 4 describes the process of creating the extended consumer health CLIR test collection (Contribution 1 above), including combining the queries from multiple CLEF eHealth test collections, creating manual translations of queries into further languages (8 languages in total), calculating machine translations of the queries (including releasing lists of the 1000 top-ranked translation hypotheses), and collecting further relevance assessments. Some information is missing, such as the number of assessors that participated and their qualifications. In addition, the fact that the assessment was run three times is mentioned in the chapter conclusion, but not

discussed further. Overall, this chapter describes a useful contribution to the field in the form of a publicly available test collection.

- Chapter 5 deals with machine translation (MT) for CLIR, and presents a mixture of a description of the state-of-the-art, new developments, and experiment results. The chapter begins with a description of techniques for Statistical Machine Translation (SMT). This is followed by a detailed description of the dataset that was created to be used for training machine translation systems in the thesis, including data selection, and data preprocessing. The text then reverts to method background, with a relatively detailed introduction to Neural Machine Translation (NMT). This is followed by a presentation of how to evaluate MT, in particular the BLEU and PER scores. It is not clear why a description of MERT, a parameter tuning algorithm for MT algorithms, is described in the section on MT evaluation. The training of both SMT and NMT systems is described next. For SMT training, previous work is replicated, but a minor contribution in the form of developing an SMT system that produces lemmatized sentences is made, leading to the training of three SMT systems (two variations of the lemmatization system). For NMT, a standard implementation is used for the training. Finally, the MT evaluation results are presented, for both document translation and query translation. The values for BLEU and PER are presented in multiple tables, but unfortunately there is no discussion of the results – this would have been particularly interesting as it was pointed out that the MT systems were optimised on CLIR results (Precision@10), so a discussion on the effect of this on BLEU and PER is missed. Overall, there is extensive information in this chapter, but unfortunately it suffers from a rather poor structure.
- Chapter 6 presents a mixture of proposed approaches and experiment results. It first presents the experiment baseline in the form of monolingual (English) retrieval results (P@10, MAP, BPREF) on the test collection. Then it presents extensive material on query translation. The baseline results are obtained by using query translation and taking the top-ranked translation hypothesis as the query. A new approach is then proposed to learn to rerank query translation hypotheses obtained from the MT system with the target being to obtain the best search results. The proposed reranking is shown to produce significantly better search results for CLIR. The use of NMT and public MT systems (Google, Bing) is investigated, although they are only compared to the baseline system and not the proposed approach. Following this, experiments on document translation for CLIR are presented. These experiments provide evidence that in the medical domain, query translation outperforms document translation, contradicting a long accepted result in the CLIR field. Finally, the proposed query expansion approach is described and the experimental results are presented. While this chapter presents a large amount of material, the material could be better structured.
- The final chapter (unnumbered) provides a well-written summary of the main results of the thesis, followed by a short list of potential future work. The thesis ends with a list of publications of Shadi Saleh, lists of figures and tables, and a bibliography of 183 references.

Shadi Saleh lists 11 publications, of which he is first author of 10 of these. Five of the publications are in conferences with peer review (SIGIR, ACL, CLEF), one is in a SIGIR Workshop with peer-review, and five are in the CLEF Working Notes, describing participation in the CLEF eHealth evaluation track. This shows very active publication of the results of thesis.

The work presented in the thesis demonstrates the author's ability for creative scientific work.

Univ.-Prof. Dr. Allan Hanbury

Baden, 19.08.2020