Cross-Lingual Information Retrieval in the Medical Domain

Shadi Saleh

In recent years, there has been an exponential growth of the digital content available on the Internet, which has correlated with the increasing number of non-English Internet users due to the spread of the Internet across the globe. This raises the importance of unlocking resources for those who want to look up information not limited to the languages they understand. For example, those who want to use the Internet to find medical content related to their health conditions (self-diagnosis) but they do not have access to resources in their language. Cross-Lingual Information Retrieval (CLIR) breaks the language barriers by allowing search for documents written in a language different from the query language.

This thesis tackles the task of CLIR in the medical domain and investigates the two main approaches: query translation (QT) where queries are machine translated to the language of documents and document translation (DT) where documents are translated to the language of queries. We proceed with our research by employing Statistical Machine Translation (SMT) systems that are tuned for the QT approach and the DT approach in the medical domain for seven European languages (Czech, German, French, Spanish, Hungarian, Polish and Swedish) and empirically show that DT does not outperform QT (the contrary to what had been assumed since the late 1990's).

We develop a machine-learning-based system to rerank the translation hypotheses provided by an SMT system towards better CLIR performance. The system is first designed for Czech, French and German CLIR systems and then is adapted to Spanish, Hungarian, Swedish and Polish. Our findings suggest that the best translation produced by SMT is not necessarily the best translation to construct a query in CLIR. Our reranker system produces translations that are optimized towards CLIR, and significantly outperforms the baseline QT system without reranking. To remedy the vagueness of translated queries in CLIR, we present a novel approach that reformulates base queries by adding useful terms to them. The terms are scored for usefulness using a linear regression model. Our approach improves both the performance of CLIR systems in all languages and of the monolingual IR (English reference queries).

To compare the performance of SMT versus NMT (predicting a translation using deep neural networks) in the context of CLIR, we train a task-oriented NMT model to translate medical queries. The presented NMT-based QT model significantly outperforms the SMT-based QT one in all languages.

During the progress of our research, we developed an extended dataset for CLIR in the medical domain, which is based on existing datasets from the IR tasks of the CLEF eHealth Labs Series 2013–2015, and we make the dataset publicly available via the Lindat/CLARIN repository.