Opponent's Review of Diploma Thesis

Title of the thesis:

Query expansion for medical information retrieval

Author:

Feraena Bibyna

---

1. Task specification

The submitted thesis focuses on automatic information retrieval in full-text medical databases. One of the existing issues is the terminological gap between the language of the stored medical documents (which are commonly written by medical professionals) and the language of common user's queries (which are typically formulated by non-professionals, who widely use layman vocabulary). In the submitted thesis, the author investigates the effect of query expansion. She uses a domain-specific knowledge resource, namely the *Unified Medical Language System* (UMLS), a repository of biomedical vocabularies containing medical terms and a system of semantic relations among the terms.

The subject of the thesis is implementation and experimental evaluation of query expansion in the medical domain. The main goal has been defined as *"evaluating the use of semantic relations between terms for query expansion"* and examining its effect on the precision of medical information retrieval.

2. Thesis structure

The first part of the thesis (Introduction and Chapter 1) gives a brief introduction to the field and describes especially the UMLS, existing retrieval models, known approaches to query reformulation/expansion, and usual evaluation metrics.

Then, in Chapters 2 and 3, the student focuses on the description of the used data set, and on the implementation of the examined retrieval models. The one million document collection used consists of web pages that cover a broad variety of medical health topic. It was provided by the organizer of CLEF eHealth 2015 Task 2. For both training and test purposes the author utilizes test queries distributed by CLEF eHealth organizer for Task 3a of 2014 and Task 2 of 2015. As the two set of test queries differ a lot, she combines them for training and then evaluates separately.

For the retrieval implementation the author makes use of the freely available open source information retrieval platform *Terrier* (Terrabyte Retriever), which provides both indexing and retrieval modules, and a sufficiently flexible query language as well. In Chapter 3 she describes in details both the implemented retrieval models, and several different kinds of query expansions, as well as several other techniques to improve retrieval performance, namely blind relevance feedback, field weighting, and linear interpolation. As to exploiting the UMLS, she utilizes two of its resources: the Metathesaurus and the Semantic Network.

Main author's contributions are described in experimental Chapters 4 (experiments on traning set) and 5 (experiments on two test sets). Here she thoroughly describes the experiments and the results, and discusses the differences in the different models' outputs. Finally, in Conclusion the author summarizes what has been done and gives some ideas about future work.

3. Formal aspects and evaluation

The submitted thesis fulfills all general formal requirements. It is written in good English and all text is well comprehensible. Minor grammatical mistakes are not too frequent. The list of bibliography consists of 31 items. The lists of Tables, Figures, and Abbreviations are added at the end.

In the thesis, the author convincingly demonstrated good knowledge of the field and practical ability to perform a large number of non-trivial experiments. The resulting data are presented in numerous tables with rich commentaries. The student is also a co-author of a workshop contribution to an international competition on the given topic, which is to be presented soon at *CLEF 2015 Labs and Workshops*.

During the defence I would like to hear more about parameter learning. The data used was strictly divided into training and test parts, which were used in a standard way. However, in the thesis there is no discussion about possible overfitting the training queries. The author simply chooses the best models according to the training results and applies them to the test portion. Are the parameters learned robust enough?

In my view, another weaker part of the presented work is the Conclusion, which is quite long, rather complicated and a bit unclear. As a result of the whole work, I would like to see simplier, better structured and more clear conclusions or recommendations or maybe directions for future research.

4. Conclusion

To conclude my review, I appreciate that the student has successfully contributed to the information retrieval research area by performing a systematic and quite exhaustive series of experiments with query expansion in the medical domain. In fact she did a lot of work, and the work has been both done and presented thoroughly.

I definitely recommend the submitted thesis for the defense and I suggest accepting this work as a diploma thesis.

Prague, August 31, 2015

RNDr. Martin Holub, Ph.D.

Institute of Formal and Applied Linguistics
Charles University in Prague