The Knowledge and Language Gap in Medical Information Seeking

Luca Soldaini
Amazon Alexa AI
lssoldai@amazon.com

Abstract

Interest in medical information retrieval has risen significantly in the last few years. The Internet has become a primary source for consumers looking for health information and advice; however, their lack of expertise causes a language and knowledge gap that affects their ability to properly formulate their information needs. Health experts also struggle to efficiently search the large amount of medical literature available to them, which impacts their ability of integrating the latest research findings in clinical practice. In this dissertation, I propose several methods to overcome these challenges, thus improving search outcomes.

For queries issued by lay users, I introduce *query clarification*, a technique to identify the most appropriate expert expression that describes their information need; such expression is then used to expand the query [5]. I experiment with three existing synonym mappings, and show that the best one leads to a 7.3% improvement over non-clarified queries. When a classifier that predicts the most appropriate mapping for each query is used, an additional 5.2% improvement over non-clarified queries is achieved.

Furthermore, I introduce a set of features to capture semantic similarity between consumer queries and retrieved documents, which are then exploited by a learning to rank framework [2]. This approach yields a 26.6% improvement over the best known results on a dataset designed to evaluate medical information retrieval for lay users.

To improve literature search for medical professionals, I propose and evaluate two query reformulation techniques that expand complex medical queries with relevant latent and explicit medical concepts [1, 4]. The first is an unsupervised system that combines a statistical query expansion with a medical terms filter, while the second is a supervised neural convolutional model that predicts which terms to add to medical queries. Both approaches are competitive with the state of the art, achieving up to 8% improvement in inferred nDCG.

Finally, I conclude my dissertation by showing how the convolutional model can be adapted to reduce clinical notes that contain significant noise, such as medical abbreviations, incomplete sentences, and redundant information [3]. This approach outperforms the best query reformulation system for this task by 27% in inferred nDCG.

This dissertation was published on May 12th, 2018.

Advisor: Dr. Nazli Goharian
Committee members: Dr. Der-Chen Chang, Dr. Ophir Frieder, Dr. Elad Yom-Tov, Dr. Wenchao Zhou.
Available at: `http://ir.cs.georgetown.edu/downloads/diss-luca-soldaini.pdf`

References

[1] Soldaini, L., Cohan, A., Yates, A., Goharian, N., and Frieder, O. Retrieving Medical Literature for Clinical Decision Support. In *Advances in Information Retrieval* (Mar. 2015), Lecture Notes in Computer Science, Springer, Cham, pp. 538–549.

[2] Soldaini, L., and Goharian, N. Learning to Rank for Consumer Health Search: a Semantic Approach. In *European Conference on Information Retrieval (ECIR)* (2017), Springer.

[3] Soldaini, L., Yates, A., and Goharian, N. Denoising Clinical Notes for Medical Literature Retrieval with Convolutional Neural Model. ACM Press, pp. 2307–2310.

[4] Soldaini, L., Yates, A., and Goharian, N. Learning to reformulate long queries for clinical decision support. *Journal of the Association for Information Science and Technology 68*, 11 (Nov. 2017), 2602–2619.

[5] Soldaini, L., Yates, A., Yom-Tov, E., Frieder, O., and Goharian, N. Enhancing web search in the medical domain via query clarification. *Information Retrieval Journal 19*, 1-2 (Apr. 2016), 149–173.