Developing Unsupervised Knowledge-Enhanced Models to Reduce the Semantic Gap in Information Retrieval

Stefano Marchesin
University of Padua, Padua, Italy
stefano.marchesin@unipd.it

Abstract

In this thesis we tackle the semantic gap, a long-standing problem in Information Retrieval (IR). The semantic gap can be described as the mismatch between users' queries and the way retrieval models answer to such queries. Two main lines of work have emerged over the years to bridge the semantic gap: (i) the use of external knowledge resources to enhance the bag-of-words representations used by lexical models, and (ii) the use of semantic models to perform matching between the latent representations of queries and documents. To deal with this issue, we first perform an in-depth evaluation of lexical and semantic models through different analyses [Marchesin et al., 2019]. The objective of this evaluation is to understand what features lexical and semantic models share, if their signals are complementary, and how they can be combined to effectively address the semantic gap. In particular, the evaluation focuses on (semantic) neural models and their critical aspects. Each analysis brings a different perspective in the understanding of semantic models and their relation with lexical models. The outcomes of this evaluation highlight the differences between lexical and semantic signals, and the need to combine them at the early stages of the IR pipeline to effectively address the semantic gap.

Then, we build on the insights of this evaluation to develop lexical and semantic models addressing the semantic gap. Specifically, we develop unsupervised models that integrate knowledge from external resources, and we evaluate them for the medical domain – a domain with a high social value, where the semantic gap is prominent, and the large presence of authoritative knowledge resources allows us to explore effective ways to address it. For lexical models, we investigate how – and to what extent – concepts and relations stored within knowledge resources can be integrated in query representations to improve the effectiveness of lexical models. Thus, we propose and evaluate several knowledge-based query expansion and reduction techniques [Agosti et al., 2018, 2019; Di Nunzio et al., 2019]. These query reformulations are used to increase the probability of retrieving relevant documents by adding to or removing from the original query highly specific terms. The experimental analyses on different test collections for Precision Medicine – a particular use case of Clinical Decision Support (CDS) – show the effectiveness of the proposed query reformulations. In particular, a specific subset of query reformulations allow lexical models to achieve top performing results in all the considered collections.

Regarding semantic models, we first analyze the limitations of the knowledge-enhanced neural models presented in the literature. Then, to overcome these limitations, we propose SAFIR [Agosti et al., 2020], an unsupervised knowledge-enhanced neural framework for IR. SAFIR integrates external knowledge in the learning process of neural IR models and it does not require labeled data for training. Thus, the representations learned within this framework are optimized for IR and encode linguistic features that are relevant to address the semantic gap. The evaluation on different test collections for CDS demonstrate the effectiveness of SAFIR when used to perform retrieval over the entire document collection or to retrieve documents for Pseudo Relevance Feedback (PRF) methods – that is, when it is used at the early stages of the IR pipeline. In particular, the quantitative and qualitative analyses highlight the ability of SAFIR to retrieve relevant documents affected by the semantic gap, as well as the effectiveness of combining lexical and semantic models at the early stages of the IR pipeline – where the complementary signals they provide can be used to obtain better answers to semantically hard queries.

Awarded by: University of Padua, Padua, Italy **on** 24 March 2021.
Supervised by: Maristella Agosti.
Available at: http://paduaresearch.cab.unipd.it/13174/.

Selected Publications

M. Agosti, G. M. Di Nunzio, and S. Marchesin. The University of Padua IMS Research Group at TREC 2018 Precision Medicine Track. In *Proc. of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018*, volume 500-331 of *NIST Special Publication*, pages 1–10. NIST, 2018.

M. Agosti, G. M. Di Nunzio, and S. Marchesin. An Analysis of Query Reformulation Techniques for Precision Medicine. In *Proc. of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 973–976. ACM, 2019.

M. Agosti, S. Marchesin, and G. Silvello. Learning Unsupervised Knowledge-Enhanced Representations to Reduce the Semantic Gap in Information Retrieval. *ACM Trans. Inf. Syst.*, 38(4): 1–48, September 2020.

G. M. Di Nunzio, S. Marchesin, and M. Agosti. Exploring how to Combine Query Reformulations for Precision Medicine. In *Proc. of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019*, volume 1250 of *NIST Special Publication*, pages 1 – 14. NIST, 2019.

S. Marchesin, A. Purpura, and G. Silvello. Focal Elements of Neural Information Retrieval Models. An Outlook through a Reproducibility Study. *Inf. Process. Manag.*, page 102109, 2019.