Large language models for information retrieval in scientific literature – LLM4SciLit

Oleg Bakhteev, Luis Salamanca, Fernando Pérez-Cruz, Rachel Ward

April 2023

1 Project description

An important challenge when conducting research lies in the search of relevant literature and finding facts of interest in it. With the development of the Internet and search engines, the literature search has become much easier: instead of searching for books of interest in libraries, materials nowadays can be found using specialized search sites and websites of scientific publishing houses. However, even with the aid of search engines, the researcher is seldom presented with a comprehensive and high-quality search output: an extensive collection of sources must be carefully studied, to determine their relevance to the research goals and to authenticate their reliability. The problem becomes more complicated when the researcher needs to form an opinion about some fact based on a meta-analysis of articles or discover some fact that is rarely found in the literature.

At the same time, the rapid development of language model technologies [1] provides us with optimism to streamline the laborious process of literature examination for researchers. Studies [2] have evinced that language models possess the ability to quite effectively answer questions from the literature. Furthermore, there have been promising findings illustrating the proficiency of language models in answering questions sourced from scientific papers [3, 5].

Nonetheless, not all challenges have been overcome with the aid of language models. One such obstacle is obtaining results in the few-shot learning mode, wherein the language model is trained to respond to questions with a limited number of examples. This problem remains unsolved, even though language models have demonstrated good performance as few-shot learners. From an applicational standpoint, few-shot learning is essential for question-answering, fact extraction, and information retrieval in the scientific domain, given that the number of examples that researchers can potentially annotate for model training is generally restricted. Another problem to be faced is scaling the model for deployment on large collections of documents. The problem of fast and effective vector search in large document collections is still open, particularly when operating under limited computational resources. Building and supporting a vector index for sentences is a resource-intensive task, and using the approximate nearest neighbor search method leads not only to lower resource requirements but often to a significant reduction in search quality [4].

In this project, we plan to develop a framework for searching relevant literature to support research on a specific domain field. For this particular project, and as use-case, we focus on the field of biomedicine, but the proposed framework must be also extensible to any other scientific domain. We plan to use large language models to obtain relevant scientific papers, and search for answers to questions in these leveraging the few-shot learning capabilities of large language models. A portion of the PubMed corpus, containing over 8 million scientific articles, will

be employed for the project. We have access to a pre-annotated sets of papers (a predefined set of questions with the answers extracted from the papers) that act as queries for training language models in the few-shot learning mode. These queries will be employed to train a model for effectively ranking documents and finding correct answers for pre-defined questions in the documents. We will start the exploration by using simpler language models, upscaling them in intermediate stages, to analyze then the dependence of the retrieval quality to the complexity of the language model. When vectorizing text, we will also explore many aspects and details of search: how to produce the most adequate segmentation of text into parts for efficient search, how to take into account the structure of the document, as well as metadata about the document, such as the domain of the document or the citation of this document. Besides, we expect the student to perform a clean and modular implementation of the pipeline, as we aim at later embed it into a more usable and general methodology.

The present project will allow the student to gain a greater familiarity with the use of different NLP methods, especially methods of few-shot learning and the ability to work with large language models, as well as methods of vector search in large-scale document collections. In the framework of the present interdisciplinary collaboration, we expect to unveil many interesting insights from the data that will be published in conferences and/or journals in the domain field. Beyond, the code implemented by the student is intended to be used as back-end for a full-fledged methodology for scientific papers' exploration, where the he/she will be also co-author.

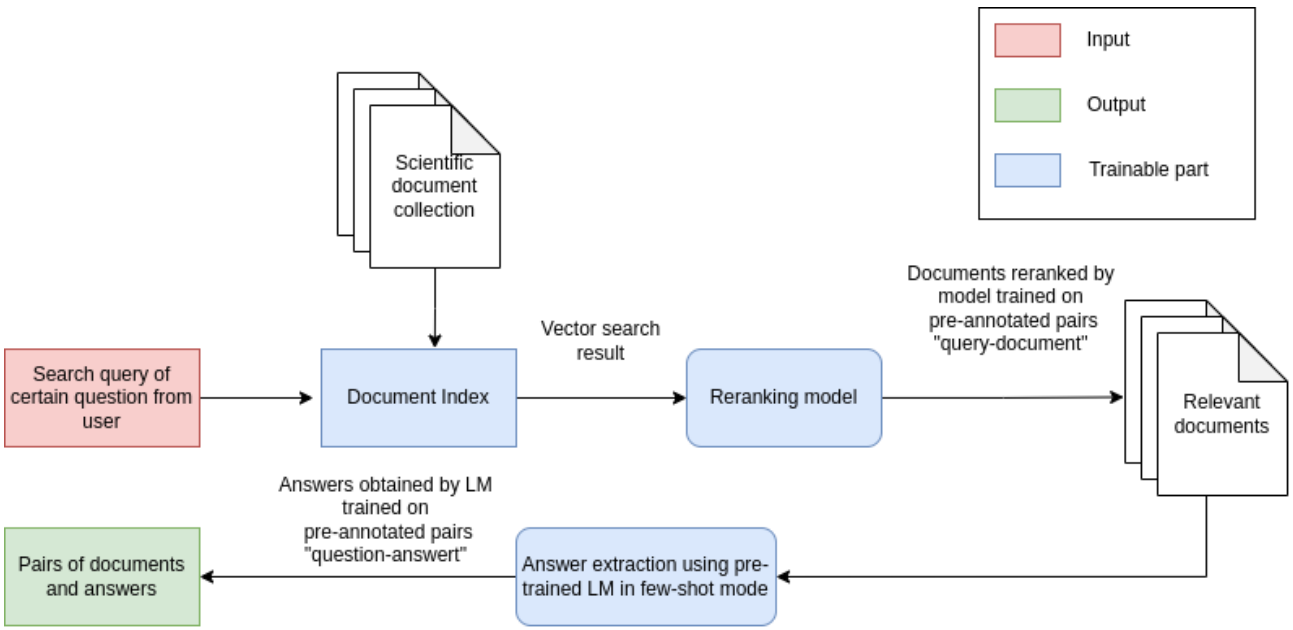

Figure 1: Scheme of the proposed framework: we use language models in two regimes, for searching and re-ranking indexed documents and for question answering. In both ways, we use annotated datasets to train our models.

2 Additional information

- **Difficulty of the project:** From moderate to very challenging.
- **What will you learn?** natural language processing, language models, vector search.

- **Requirements:** Good python level, some knowledge of either Tensorflow or PyTorch, experience with git, machine learning fundamentals, creative thinking, **interest in working with large scientific document collections and scientometrics**.
- **Supervisors:** Oleg Bakhteev (oleg.bakhteev@epfl.ch),
  Luis Salamanca (luis.salamanca@sdsc.ethz.ch),
  Fernando Pérez-Cruz (fernando.perezcruz@sdsc.ethz.ch),
  Rachel Ward (rachel.ward@hest.ethz.ch).

External references

- **PubMed corpus**. https://pubmed.ncbi.nlm.nih.gov/.
- **Paper QA project.** Similar to ours and can be used as a starting point for the project development. https://github.com/whitead/paper-qa.

Bibliography

References

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[2] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: `https://openreview.net/forum?id=H1eA7AEtvS`.

[3] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[4] Harsha Vardhan Simhadri, George Williams, Martin Aumüller, Matthijs Douze, Artem Babenko, Dmitry Baranchuk, Qi Chen, Lucas Hosseini, Ravishankar Krishnaswamny, Gopal Srinivasa, et al. Results of the neurips'21 challenge on billion-scale approximate nearest neighbor search. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 177–189. PMLR, 2022.

[5] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. Pre-trained language model for biomedical question answering. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 727–740. Springer, 2020.