Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Neural Retrievers for Question Answering Systems in the Biomedical Domain

Study Project Exposé

Robert Martin

March 9, 2023

1 Introduction

Search engines are an essential tool for research in biology and medicine. Given a user query, they return a ranked list of potentially relevant documents. PubMed is a search engine and database of *"more than 35 million citations for biomedical literature"* [1]. Many of these publications can be relevant for deciding the best treatment for a patient in Evidence-Based Medicine (EBM) [Athenikos and Han (2010)]. It is left to the researcher to study, filter, and combine search results to derive answers [Tsatsaronis et al. (2015)]. Russell-Rose and Chamberlain (2017) found that medical professionals spent hours to complete search tasks on PubMed and other related search engines. In contrast to a list of retrieved candidates, question answering (QA) systems provide a direct answer, which can help biomedical researchers navigate the vast amount of literature more efficiently.

Traditional QA systems like "Watson" [Ferrucci (2012)] process queries in multiple stages of analysis and rely on structured and semi-structured data sources, such as knowledge bases, to find answers [Baradaran et al. (2022)]. Making information accessible to the machine often involves a large amount of manual work in annotating literature. Knowledge representations need to be updated and refined continuously in a dynamic field such as the biomedical one.

To overcome this issue, there has been a trend towards applying machine learning models to source information directly from natural language texts. This is achieved by pretraining deep neural networks on language modeling (LM) objectives, such as masked-language modeling (MLM) [Devlin et al. (2018)]. The goal of LM is to learn a vector embedding representation for text that encapsulates its semantics, topic or meaning. Without the need of manual annotation, LMs can be trained on vast amounts of document collections, containing books [Zhu et al. (2015)], massive web crawls [Raffel et al. (2019)] or Wikipedia articles to learn meaningful text representations.

Pretrained Language Models (PLMs) [Peters et al. (2018), Devlin et al. (2018)] are the most prominent building block to improve language understanding in biomedical QA (BQA) systems [Jin et al. (2022)], with PLMs being specifically trained on literature from the biomedical domain [Lee et al. (2019)]. PLMs can be finetuned on the objective of downstream tasks such as machine reading comprehension (MRC) to identify answers for questions in a given text context.

However, applying MRC models to identify the answer in an entire corpus, e.g. PubMed, is computationally expensive and may lead to poor results since the majority of documents are likely to be irrelevant to the given question. Consequently, the approach commonly involves the following two steps: first, a context retriever extracts a small subset of text units from the corpora which probably contain the answer. These can be sentences, passages or whole documents. Secondly, the retrieved text units and question are fed to an MRC model to identify the correct answer [Karpukhin et al. (2020)]. Chen et al.

[1] https://pubmed.ncbi.nlm.nih.gov *last access: February 14, 2023*

(2017) and Karpukhin et al. (2020) find that the quality of the retrieval stage has a significant impact on question answering performance in the second stage.

A standard approach for the retrieval step involves the use of term matching algorithms, such as TF-IDF and BM25 [Robertson and Zaragoza (2009)], which score documents based on how many words are shared with the question. Text for retrieval is represented by high-dimensional, sparse vectors. Each dimension depicts the presence of a word from the vocabulary with an applied weighting based on relevance heuristics. The vector is sparse, because most terms in the vocabulary do not exist in an individual document and the corresponding dimension values are zero. A downside of this encoding is the disregard of semantic relations and the context of words. These shortcomings become more apparent in QA. For instance, it is not possible to encode a request for quantities, such as "How many" [Luo et al. (2022)]. When ignoring context, biomedical terms encompass a high level of ambiguity. The term "promoter" refers to an entire different concept in biology then in chemistry [Spasic et al. (2005)]. Additionally, a high amount of abbreviations are used which can be expanded to many different terms depending on the underlying document.

As an alternative to term matching algorithms, PLMs produce dense contextual vector representations that are not affected by the limitations of sparse encoding, but have a much higher computational impact. PLMs have been applied for re-ranking the top candidate results of conventional retrieval systems [Lei et al. (2016)], but documents with synonyms that rank low in preliminary retrieval might not be part of the selected candidates.

Neural Retrievers (NR) derive relevant documents or passages directly in response to a query. They mitigate the computational expense of PLMs by encoding documents separately from input questions to be indexed ahead of retrieval. A popular method is to apply a dual encoder model architecture and represent query and document individually using two PLMs [Karpukhin et al. (2020)]. The encoders are finetuned together so that dense representation pairs of question and relevant document are close in a joined vector space and have a higher similarity to each other than the irrelevant ones. During interference, only questions have to be encoded by the model and relevant documents can be found by applying a fast neighbor look-up in an index of precomputed document representations.

The application of neural retrievers has shown to outperform term matching approaches like BM25 in the open domain [Karpukhin et al. (2020)]. However, the biomedical domain presents unique challenges. Annotating large biomedical corpora is expensive and current expert-annotated BQA datasets are small in size [Jin et al. (2022)], making it more difficult to train and evaluate neural retrieval methods. As a result of this, the application of Neural Retrievers in the biomedical domain is still an area of limited research. For this project we plan to adapt two neural retrieval systems for biomedical QA on PubMed articles and compare them to the well-established term matching approach BM25.

2 Statement of the problem

Given a document collection $D = \{d_1, \ldots, d_n\}$ and a question q, the goal is to train a retrieval model f_r that calculates a relevance score conditioned on each document d_i so that

$$f_r(q, d_i, \Theta) = s_i \ ,$$

where $\Theta = \{\theta_1, \ldots, \theta_m\}$ is the set of trainable model parameters. The score s_i approximates the probability $p(\text{rel}|d_i, q)$ that a document d_i is relevant to the question q. As a result set $r_q = \{\hat{d}_1, \ldots, \hat{d}_k\}$ we select the top-k documents for the question q from D with the highest relevance scores. Let d^+ be a positive (relevant) document and d^- a negative one from a gold standard. A common choice to optimize a retrieval model f_r as applied in the DPR approach is to minimize the contrastive loss defined as

$$L(f_r, q, d^+, d^-) = \log \frac{e^{f_r(q,d^+,\Theta)}}{e^{f_r(q,d^+,\Theta)} + \sum_{i=1}^{N} e^{f_r(q,d_i^-,\Theta)}}$$

[Shen et al. (2022)].

To analyze and compare results of the three retrieval approaches for this project, we test them on a dataset Q of questions annotated with relevant documents from D. Let $\text{rel}_q = \{\tilde{d}_1, \ldots, \tilde{d}_l\}$ be the set of true relevant documents for a question q and $\hat{d}_k$ the ranked top K results of the model f_r. A performance measure for f_r can be determined by first calculating the average precision at k (AP@K) for each question q_i so that

$$AP@K(q_i) = \frac{1}{r_k} \sum_{k=1}^{K} \frac{||\text{rel}_{q_i}|| \cdot \text{rel}(k)}{k} \ , \quad \text{where rel}(k) = \begin{cases} 1 & \text{if } \hat{d}_k \in \text{rel}_q \\ 0 & \text{otherwise} \end{cases}$$

and r_k is the total number of relevant documents in the top K results. Subsequently we can report the Mean Average Precision at k (MAP@K) value for the whole dataset by averaging over the AP@K values for each question.

3 State of the art

Neural retrievers (NR) have been leading in performance over traditional retrieval approaches such as TF-IDF and BM25 [Robertson and Zaragoza (2009)] on open domain QA datasets. Prominent NR approaches apply cross-attention and dual encoder model architectures. Cross-attention models have been used for document ranking [MacAvaney et al. (2019), Nogueira and Cho (2019)], but are computationally expensive and cannot be applied directly on large corpora. Dual encoder models, such as Dense Passage Retrieval (DPR) by Karpukhin et al. (2020), employ separate encoders for query and document

context, which allows to precompute encodings for the whole corpus. The final retrieval score is calculated using a simple dot product between query and context vectors.

A disadvantage of this approach is the lack of model interaction for calculating the final retrieval score. Thakur et al. (2021) observe that DPR encounters issues when the model input deviates too much from the training data. Khattab and Zaharia (2020) propose ColBERT, adding a late token-level model interaction step over query and context representations, which improves generalization performance at the cost of higher retrieval latency [Thakur et al. (2021)]. However, storing token vectors for each context representation greatly increases the space requirement for the document index. Santhanam et al. (2021) apply a residual compression mechanism to reduce the space footprint while preserving approximately the same quality as uncompressed embeddings.

Comparing the generalization ability of DPR with late-interaction models, it is easy to presume that the dot-product is a bottleneck and not powerful enough to capture semantic relevance. However, Ni et al. (2021) are able to consistently improve generalization ability of DPR by increasing the encoder model size while keeping the bottleneck embedding size fixed.

Tay et al. (2022) demonstrate a new paradigm for building retrieval models. They propose Differentiable Search Index (DSI) that learns to generate a unique identifier for each document in the corpus. Afterwards, beam search can be used on a question to generate a ranked list of potentially relevant document ids. In order to ingrain semantics into document ids, the authors apply a hierarchical clustering algorithm so that each consecutive id digit references an increasingly subdivided cluster in the embedding space. Wang et al. (2022) propose a Prefix-Aware Weight-Adaptor (PAWA) decoder to better leverage the hierarchical document id structure. The authors employ special decoder tokens that uniquely identify a cluster in the hierarchical tree structure and constrain beam search to only output tokens of the corresponding level.

Due to comparatively small gold standard dataset sizes for training, it is not trivial to bring the successes of NRs from the open domain to the biomedical one. Luo et al. (2022) apply DPR on the biomedical domain and compile insights over domain specific issues of neural retrievers. Notably, dense retrievers have a reduced emphasis on exact word matching which affects the performance more negatively on biomedical datasets such as BioASQ. Furthermore, dense approaches struggle with representing larger contexts [Yang et al. (2019)]. To mitigate this, Luo et al. (2022) represent context not by a single BERT representation for the whole document, but use the context representations for each token to calculate a maximum similarity score between query and document. They also leverage additional pretraining strategies which reflect larger context sizes and improve exact word matching performance. Additionally, the authors employ a hybrid model by accumulating DPR model and BM25 scores. The retrieval results based on this joined similarity score improve, showing that DPR and BM25 tend to different signals in the document and complement each other.

4 Methodology

Since their introduction by Karpukhin et al. (2020), dual encoder models have been the de facto standard approach for neural retrieval. While the authors have shown that their neural retrieval approach is comparable or better than BM25 in certain tasks, e.g. open domain QA, their application to the biomedical domain is still understudied. For this project we plan to adapt the Dense Passage Retrieval (DPR) approach by Karpukhin et al. (2020) for the application on PubMed articles. As an immediate step for applying DPR on biomedical corpora, we can replace the underlying BERT model with the domain specific model BioBERT [Lee et al. (2019)], which is pretrained on PubMed articles. What makes biomedical QA especially challenging for PLMs is a lack of large annotated datasets. While most gold standard datasets annotate relevant document passages for each question, DPR additionally requires negative training examples. A common strategy for a given question is to utilize positive documents for other questions in the same training batch as negatives. The main disadvantage is that random negative documents are likely to address an entirely different topic and are often not hard to distinguish from a positive one. Providing hard negatives during training significantly improves the performance of dense retrieval models [Lu et al. (2021)]. Karpukhin et al. (2020) suggest to use false positive documents retrieved by BM25 that do not contain the answer of the question as additional hard negative training examples.

Along with the dual encoder model approach, we plan to implement the DSI model by Tay et al. (2022) since it has shown promise on the open domain Natural Questions (NQ) dataset of Wikipedia articles and does not require negative training examples. The biomedical corpus PubMed is a much larger document collection and we can investigate how corpus size affects its ability to memorize documents. In this scenario, the model also has to deal with the challenge of having much less training data. The DSI model is built upon the T5 sequence to sequence model [Raffel et al. (2019)] and we will use an equivalent one specifically trained on the biomedical domain, i.g. SciFive [Phan et al. (2021)].

Finally, since term matching algorithms based on bag of words representations are still frequently used in the biomedical domain, we will compare them to our neural retriever results. For this purpose we apply the Pyserini toolkit by Lin et al. (2021) which includes an implementation of BM25.

A prominent challenge of the biomedical domain is the lack of large expert annotated datasets. For this project we plan to train and evaluate the neural retrieval methods on the BioASQ 10b dataset [Tsatsaronis et al. (2015)]. It contains approximately 4 000 questions, annotated with PubMed articles containing the answer. In contrast to the biomedical domain, the open domain Natural Questions (NQ) [Kwiatkowski et al. (2019)] dataset for Wikipedia articles includes over 300 000 question-document pairs. The aforementioned datasets show the disproportion of training data availability between the domains and we will evaluate the methods on both as a comparison.

Training language models is computational expensive and time consuming. While

experimenting with different parameters it is crucial to reduce the corpus to a more manageable size. Tay et al. (2022) split the NQ dataset into the three different size categories NQ10K, NQ100K, and NQ320K, with the numbers denoting the total document count. To reduce the corpus size, articles are sampled uniformly from the complete set and NQ320K contains all articles from the corpus. Similarly to the NQ dataset we reduce the PubMed corpus to 1 million and 100 000 documents respectively. Additionally to the reduction in computational expense while training the models, we can simulate the effects of different corpus sizes on model performance and will report results for each corpus size.

References

Athenikos, S. J. and Han, H. (2010). Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24.

Baradaran, R., Ghiasi, R., and Amirkhani, H. (2022). A survey on machine reading comprehension systems. *Natural Language Engineering*, pages 1–50.

Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading wikipedia to answer open-domain questions.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Ferrucci, D. A. (2012). Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4):1:1–1:15.

Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., Chen, M., Huang, S., Liu, X., and Yu, S. (2022). Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys*, 55(2):1–36.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering.

Khattab, O. and Zaharia, M. (2020). ColBERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Lei, T., Joshi, H., Barzilay, R., Jaakkola, T., Tymoshenko, K., Moschitti, A., and Màrquez, L. (2016). Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Lin, J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., and Nogueira, R. (2021). Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Lu, J., Abrego, G. H., Ma, J., Ni, J., and Yang, Y. (2021). Multi-stage training with improved negative contrast for neural passage retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Luo, M., Mitra, A., Gokhale, T., and Baral, C. (2022). Improving biomedical information retrieval with neural retrievers.

MacAvaney, S., Yates, A., Cohan, A., and Goharian, N. (2019). Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Ni, J., Qu, C., Lu, J., Dai, Z., Ábrego, G. H., Ma, J., Zhao, V. Y., Luan, Y., Hall, K. B., Chang, M.-W., and Yang, Y. (2021). Large dual encoders are generalizable retrievers.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with bert.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

Phan, L. N., Anibal, J. T., Tran, H., Chanana, S., Bahadroglu, E., Peltekian, A., and Altan-Bonnet, G. (2021). Scifive: a text-to-text transformer model for biomedical literature.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Russell-Rose, T. and Chamberlain, J. (2017). Expert search strategies: The information retrieval practices of healthcare information professionals. *JMIR Medical Informatics*, 5(4):e33.

Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., and Zaharia, M. (2021). Colbertv2: Effective and efficient retrieval via lightweight late interaction.

Shen, X., Vakulenko, S., del Tredici, M., Barlacchi, G., Byrne, B., and de Gispert, A. (2022). Low-resource dense retrieval for open-domain question answering: A comprehensive survey.

Spasic, I., Ananiadou, S., McNaught, J., and Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251.

Tay, Y., Tran, V. Q., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., Schuster, T., Cohen, W. W., and Metzler, D. (2022). Transformer memory as a differentiable search index.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artiéres, T., Ngomo, A.-C. N., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., and Paliouras, G. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1).

Wang, Y., Hou, Y., Wang, H., Miao, Z., Wu, S., Sun, H., Chen, Q., Xia, Y., Chi, C., Zhao, G., Liu, Z., Xie, X., Sun, H. A., Deng, W., Zhang, Q., and Yang, M. (2022). A neural corpus indexer for document retrieval.

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., and Lin, J. (2019). End-to-end open-domain question answering with bertserini.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.