Evaluating a lightweight neural reranking model for biomedical question answering

Tiago Almeida
tiagomeloalmeida@ua.pt

IEETA
Universidade de Aveiro
Aveiro, Portugal

Sérgio Matos
aleixomatos@ua.pt

DETI/IEETA
Universidade de Aveiro
Aveiro, Portugal

Abstract

Automatic searching mechanisms are essential to human progress by simplifying access to relevant information in increasingly large libraries.

In this paper, we present a lightweight searching system, that combines traditional techniques with neural networks yielding a model with only 620 trainable parameters that can be applied to any searching problem for which there is training data available.

We evaluated our system in two challenges, both on the biomedical domain, namely BioASQ 8b and TREC-Covid. In the first one, we achieved top and close to top scores in all the batches, while on TREC-Covid our best result was a second place in the third round.

1 Introduction

In today's science, we witness an unprecedented amount of new information being generated every year. So, the ability to automatically search this unstructured information, like documents, articles, or web pages, becomes a cornerstone of scientific development and progress.

As an example, in the biomedical area, scientists need to routinely search a constantly increasing amount of information, usually in the form of scientific articles, to conduct their day-to-day tasks, which becomes an extremely time-consuming effort. To give a better context, during the current pandemic situation more than 200 thousand articles exclusively related to the study of the coronavirus were published, at a rhythm of approximately one thousand new articles per day[1]. In a more global view, the most used database PubMed/MEDLINE has 30 million indexed articles and is growing at a rate of one and a half million new articles per year.

This searching challenge is addressed by the Information Retrieval (IR) field that studies and creates automatic systems capable of retrieving the most relevant piece of information (usually documents) from a set of unstructured information (set of documents also designated corpus) given a query that encodes the information need. Nowadays, the IR field is considered to be divided into traditional IR and neural IR. The former uses handcrafted rules and equations to directly compute the query-document importance, the BM25 [8] ranking equation being the most popular example. On the other hand, neural IR explores the increasing success of neural networks to approximate a (sub)optimal ranking function by exploring labeled examples. In the literature, the most successful neural architectures for this type of search are Interaction Based, which create a joint representation of the question and the documents by considering multiple matching signals. DRMM [5] and DeepRank [7] are examples of such neural architecture.

This paper presents our lightweight neural interaction-based system, with only 620 trainable parameters, to tackle the previously enunciated searching problem. This system follows a two-step approach where we combined the BM25, a traditional approach, with our lightweight neural model.

We evaluate our system on the 8th BioASQ challenge and on the TREC-Covid challenge, where for the BioASQ we achieved top and close to the top scores for all the batches, while in the TREC-Covid we achieved a second place on the third round as the best result. We are also participated in the TREC-Deep Learning and TREC-Precision Medicine challenges using this same system. However, at the time of writing the results are not available.

[1]These values are inferred from the metadata from the CORD-19 Corpus `https://www.semanticscholar.org/cord19`

2 System Description

As previously mentioned, our system follows a two-step retrieval strategy, more precisely, we adopt the BM25, as the first step, to act as a filter in order to reduce the enormous search space and select only the *top-N* most relevant documents that are further ranked by the neural model, as described in Figure 1.

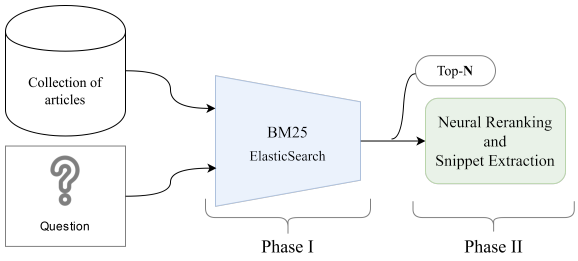

Figure 1: Overview of the system flow.

Our lightweight neural ranking model is a direct evolution of the following previous work [1, 4]. This new version was designed to weight the importance of the document sentences concerning the query by taking into consideration the context where the exact match occurs, *e.g.*, this model produces a more refined judgment of the previously exact match signal considered in the BM25.

Additionally, supported by the inner structure of the designed neural model we devised a zero-shot learning algorithm for sentence extraction, which gave us the ability to extract the sentences that are supposedly more relevant to a given question. This was engineered by looking at the activation value that the model gave to each sentence and retrieving those with higher values. So we assumed that the sentences that most contribute to the final document score must be also the most relevant sentences present in that document for that given question. For a more complete view of the neural model and the zero-shot snippet extraction, we redirect the reader to the following paper [2].

3 Evaluation

In this section, we will individually describe each competition and task following by the respective results.

3.1 BioASQ

The BioASQ challenge [9] is an annual competition on document classification, retrieval, and question-answering, currently in the eighth edition. We submitted our system to be evaluated on the document and snippet retrieval task, part of BioASQ task 8b phase A. For the document retrieval task the objective was to retrieve the most relevant articles from last year's PubMed/MEDLINE annual database. The snippet task is similar but the unit of information becomes the sentences from the PubMed/MEDLINE articles.

The organizers published, in intervals of two weeks, a total of 500 biomedical questions split into five batches. For each batch we submit our system results that were evaluated in terms of Recall, F1, MAP, and GMAP.

With respect to the system, the BM25 filter was fine-tuned with the 2700 biomedical questions provided by the organizers as the training data. The neural model was trained on the same data using a pairwise cross-entropy loss with cyclic learning rates. We also used the GenSim imple-

mentation of the word2vec [6] algorithm to train the word embeddings directly on the PubMed/MEDLINE articles.

Regarding the competition, this edition received on average a total of 25 submissions from 9 teams for the five batches.

Table 1: Summary of the results obtained in the BioASQ.

System	Document Retrieval			Snippet Retrieval		
	Rank	MAP@10	GMAP@10	Rank	MAP@10	F1@10
Batch 1						
Ours	1	33.98	**1.20**	5	29.53	15.00
Top Competitor	3	33.59	0.88	1	85.75	17.52
Batch 2						
Ours	3	31.68	**2.23**	4	27.67	14.13
Top Competitor	1	33.04	1.85	1	68.21	17.73
Batch 3						
Ours	4	43.69	**2.04**	5	41.37	16.61
Top Competitor	1	45.10	1.87	1	100.39	21.40
Batch 4						
Ours	4	40.24	1.31	7	36.59	17.23
Top Competitor	1	41.63	2.04	1	102.44	21.51
Batch 5						
Ours	1	48.42	**3.49**	5	43.79	19.60
Top Competitor	2	48.25	2.54	1	112.67	24.91

In terms of results, we achieved highly competitive scores on the document retrieval task, as shown in Table 1, being first on the first and fifth batches. On the other hand, although our results in the snippet retrieval task were comparatively lower, we consider the results to be encouraging, especially in terms of F1 score, given that these were obtained without supervision.

3.2 TREC-Covid

TREC-Covid was an initiative to rapidly promote the development of an automatic system capable of searching the fast growing literature about the novel coronavirus to aid scientists in their researches. This challenge was organized, in a matter of weeks, by the Allen Institute for Artificial Intelligence (AI2), the National Institute of Standards and Technology (NIST), Oregon Health Science University (OHSU), and others.

The challenge follows a TREC style format and relies on the CORD-19 dataset[2] as the collection of scientific articles about the novel coronavirus. The objective was to retrieve the most relevant articles from this collection for each topic given by the organizers. In TREC challenges, the topic represents the information need that in this case can be used as the query to search the collection.

The competition had a total of five rounds, each with an increasing number of topics: the first round had a total of 30 topics, with increments of 5 topics for the following rounds. The system results were evaluated in a residual manner, except for the first round, since the remaining rounds share topics that had already been evaluated. The metrics adopted were P@5, NDCG@10, Brepf, and MAP. The organizers also allowed the use of the evaluation feedback of previous rounds as training data to tune/train the submitted systems. An important note is that for the first round no training data was available since no previous evaluation had been performed.

Concerning the system, we utilized the BioASQ system on the first round, which means this was a transfer learning approach, exploiting the proximity of the domains. The only change was the embeddings that were trained on the PubMed/MEDLINE articles and the CORD-19 dataset. For the remaining rounds, we kept a similar approach but we also fine-tuned (trained) the model with the feedback data from previous rounds. For a more complete description of the strategy followed, we redirect the reader to the following work [3].

Regarding the competition, this challenge received a lot of interest from the community, resulting in one of the highest TREC participation rates ever. For example, in the first round, a total of 56 teams submitted results for a total of 143 runs.

Table 2: Summary of the two best results achieved on TREC-Covid.

System	Round 1			Round 3		
	Rank	P@5	NCDG@10	Rank	P@5	NCDG@10
Ours	9	63.33	52.98	2	**86.50**	77.15
Top Competitor	1	78.00	60.80	1	86.00	77.40

[2] https://www.semanticscholar.org/cord19

In terms of results, we achieved positive and encouraging results, being the best one on the third round as shown in Table 2. Furthermore, we show that our assumption to perform transfer learning with the BioASQ data empirically works, by beating traditional IR techniques and more recent transform-based techniques like BERT and T5. Regarding the remaining rounds, we scored approximately in the middle of the table. This lower results could be partially explained by mistakes later identified in these submissions.

4 Conclusion

In this paper, we show a two-stage retrieval system that was evaluated on two biomedical challenges, namely BioASQ 8b and TREC-Covid. Regarding the BioASQ 8b, we demonstrate the effectiveness of our system on the document task, by achieving top scores, and show a promising zero-shot learning setup for the snippet retrieval task. With respect to the TREC-Covid challenge, we demonstrate a successful transfer learning technique of our BioASQ system to this new task by leveraging the proximity of domains between the tasks.

Acknowledgements

This work was supported by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020, and by the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking undergrant agreement No 806968.

References

[1] Tiago Almeida and Sérgio Matos. Neural-based snippet extraction for biomedical question answering. *Proceedings of the 25th Portuguese Conference on Pattern Recognition*, 2019. URL http://recpad2019.dcc.fc.up.pt/wp-content/uploads/2019/05/ProceedingsRECPAD.pdf#page=79.

[2] Tiago Almeida and Sérgio Matos. BIT.UA at BioASQ 8: Lightweight neural document ranking with zero-shot snippet retrieval. *BioASQ 8 workshop, CLEF 2020*, 2020.

[3] Tiago Almeida and Sérgio Matos. Frugal neural reranking: evaluation on the covid-19 literature. 2020. URL https://openreview.net/pdf?id=TtcUlbEHkum.

[4] Tiago Almeida and Sérgio Matos. Calling attention to passages for biomedical question answering. In *Advances in Information Retrieval*, pages 69–77, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45442-5. doi: 10.1007/978-3-030-45442-5_9.

[5] Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. Deep Relevance Ranking Using Enhanced Document-Query Interactions. sep 2018. URL http://arxiv.org/abs/1809.01682.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[7] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. Deeprank. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Nov 2017. doi: 10.1145/3132847.3132914.

[8] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019.

[9] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weißenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and Georgios Paliouras. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 04 2015. doi: 10.1186/s12859-015-0564-6.