# Calling Attention to Passages for Biomedical Question Answering

Tiago Almeida and Sérgio Matos(✉)

DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
{tiagomeloalmeida,aleixomatos}@ua.pt

**Abstract.** Question answering can be described as retrieving relevant information for questions expressed in natural language, possibly also generating a natural language answer. This paper presents a pipeline for document and passage retrieval for biomedical question answering built around a new variant of the DeepRank network model in which the recursive layer is replaced by a self-attention layer combined with a weighting mechanism. This adaptation halves the total number of parameters and makes the network more suited for identifying the relevant passages in each document. The overall retrieval system was evaluated on the BioASQ tasks 6 and 7, achieving similar retrieval performance when compared to more complex network architectures.

**Keywords:** Biomedical question answering · Neural networks · Attention mechanism · Snippet extraction

## 1 Introduction

Question Answering (QA) is a subfield of Information Retrieval (IR) that specializes in producing or retrieving a single answer for a natural language question. QA has received growing interest since users often look for a precise answer to a question instead of having to inspect full documents [4]. Similarly, biomedical question answering has also gained importance given the amount of information scattered over large specialized repositories such as MEDLINE. Research on biomedical QA has been pushed forward by community efforts such as the BioASQ challenge [13], originating a range of different approaches and systems.

Recent studies on the application of deep learning methods to IR have shown very good results. These neural models are commonly subdivided into two categories based on their architecture. **Representation-based** models, such as the Deep Structured Semantic Model (DSSM) [5] or the Convolutional Latent Semantic Model (CLSM) [12], learn semantic representations of texts and score each query-document pair based on the similarity of their representations. On the other hand, models such as the Deep Relevance Matching Model (DRMM) [3] or DeepRank [10] follow a **interaction-based** approach, in which matching signals between query and document are captured and used by the neural network to produces a ranking score.


J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 69–77, 2020.
https://doi.org/10.1007/978-3-030-45442-5_9

The impact of neural IR approaches is also noticeable in biomedical question answering, as shown by the results on the most recent BioASQ challenges [9]. The top performing team in the document and snippet retrieval sub-tasks in 2017 [1], for example, used a variation of the DRMM [8] to rank the documents recovered by the traditional BM25 [11]. For the 2018 task, the same team extended their system with the inclusion of models based on BERT [2] and with joint training for document and snippet retrieval.

The main contribution of this work is a new variant of the DeepRank neural network architecture in which the recursive layer originally included in the final aggregation step is replaced by a self-attention layer followed by a weighting mechanism similar to the term gating layer of the DRMM. This adaptation not only halves the total number of network parameters, therefore speeding up training, but it is also more suited for identifying the relevant snippets in each document. The proposed model was evaluated on the BioASQ dataset, as part of a document and passage (snippet) retrieval pipeline for biomedical question answering, achieving similar retrieval performance when compared to more complex network architectures. The full network configuration is publicly available at https://github.com/bioinformatics-ua/BioASQ, together with code for replicating the results presented in this paper.

## 2 System Description

This section presents the overall retrieval pipeline and describes the neural network architecture proposed in this work for the document ranking step.

The retrieval system follows the pipeline presented in Fig. 1, encompassing three major modules, **Fast Retrieval**, **Neural Ranking** and **Snippet extraction**. The fast retrieval step is focused on minimizing the number of documents passed on to the computationally more demanding neural ranking module, while maintaining the highest possible recall. As in previous studies [1,7], we adopted Elasticsearch (ES) with the BM25 ranking function as the retrieval mechanism.

The documents returned by the first module are ranked by the neural network which also directly provides to the following module the information for extracting relevant snippets. These modules are detailed in Sects. 2.1 and 2.2.

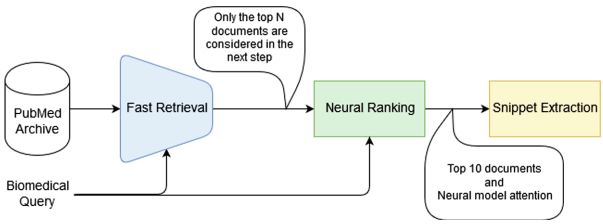


**Fig. 1.** Overview of the main modules of the proposed system. The number N of documents returned by the first module is considered an hyper-parameter.

### 2.1 Neural Ranking Model

The network follows a similar architecture to the original version of DeepRank [10], as illustrated in Fig. 2. Particularly, we build upon the best reported configuration, which uses a CNN in the measurement network and the reciprocal function as the position indicator. The inputs to the network are the **query**, a **set of document passages** aggregated by each query term, and the **absolute position** of each passage. For the remaining explanation, let us first define a query as a sequence of terms $q = \{u_0, u_1, ..., u_Q\}$, where $u_i$ is the $i$-th term of the query; a set of document passages aggregated by each query term as $D(u_i) = \{p_0, p_1, ..., p_P\}$, where $p_j$ corresponds to the $j$-th passage with respect to the query term $u_i$; and a document passage as $p = \{v_0, v_1, ..., v_S\}$, where $v_k$ is the $k$-th term of the passage. We chose to aggregate the passages by their respective query term at the input level, since it simplifies the neural network flow and implementation.

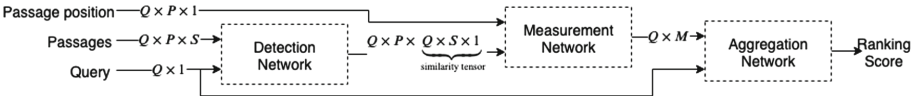


**Fig. 2.** High-level structure and data flow of the proposed version of DeepRank.

The **detection network** receives as input the **query** and the **set of document passages** and creates a similarity tensor (interaction matrix) $S \in [-1, 1]^{Q \times S}$ for each passage, where each entry $S_{ij}$ corresponds to the cosine similarity between the embeddings of the $i$-th query term and $j$-th passage term, $S_{ij} = \frac{\vec{u_i}^T \cdot \vec{v_j}}{\|\vec{u_i}\| \times \|\vec{v_j}\|}$.

The **measurement network** step is the same used in the original DeepRank model. It takes as inputs the previously computed tensors $S$ and the **absolute position** of each passage and applies a 2D convolution followed by a global max polling operation, to capture the local relevance present in each tensor $S$, as defined in Eq. 1:

$$\begin{aligned} h^m_{i,j} &= \sum_{s=0}^{x-1} \sum_{t=0}^{y-1} w^m_{s,t} \times S_{i+s,j+t} + b^m \,, \\ h^m &= \max_{i,j}(h^m_{i,j}), \;\; m = 1, ..., M \,. \end{aligned} \tag{1}$$

At this point, the set of document passages for each query term is represented by their respective vectors $\vec{h}$, i.e, $D(u_i) = \{\vec{h_{p_0}}, \vec{h_{p_1}}, ..., \vec{h_{p_P}}\}$, where $\underset{M \times 1}{\vec{h}}$ encodes the local relevance captured by the $M$ convolution kernels of size $x \times y$, plus an additional feature corresponding to the position of the passage.[1]

[1] For simplicity, we consider that the dimension $M$ already accounts for the concatenated feature, i.e, $\underset{M \times 1}{\vec{h}} \leftarrow \underset{(M+1) \times 1}{\vec{h}}$.

The next step uses a self-attention layer [6] to obtain an aggregation $\underset{M\times 1}{\vec{c_{u_i}}}$ over the passages $h_{p_j}$ for each query term $u_i$, as defined in Eq. 2. The weights $a_{p_j}$, which are computed by a feed forward network and converted to a probabilistic distribution using the softmax operation, represent the importance of each passage vector from the set $D(u_i)$. The addition of this self-attention layer, instead of the recurrent layer present in the original architecture, allows using the attention weights, that are directly correlated with the local relevance of each passage, to identify important passages within documents. Moreover, this layer has around $A \times M$ parameters, compared to up to three times more in the GRU layer (approximately $3 \times A \times (A + M)$), which in practice means reducing the overall number of network parameters to half.

$$\begin{aligned} s_{p_j} &= \underset{1\times A}{w^T} \cdot \tanh\left(\underset{A\times M}{W} \cdot \underset{M\times 1}{\vec{h}_{p_j}}\right), \\ a_{p_j} &= \frac{e^{s_{p_j}}}{\sum_{p_k \in D(u_i)} e^{s_{p_k}}}, \\ \underset{M\times 1}{\vec{c_{u_i}}} &= \sum_{p_j \in D(u_i)} \left(\underset{1\times 1}{a_{p_j}} \times \underset{M\times 1}{\vec{h}_{p_j}}\right). \end{aligned} \tag{2}$$

Finally, the **aggregation network** combines the vectors $\underset{M\times 1}{\vec{c_{u_i}}}$ according to weights that reflect the importance of each individual query term $u_i$. We chose to employ a similar weighting mechanism to the term gating layer in DRMM [3], which uses the query term embedding to compute its importance, as defined in Eq. 3. This option replaces the use of a trainable parameter for each vocabulary term, as in the original work, which is less suited for modelling a rich vocabulary as in the case of biomedical documents.

The final aggregated vector $\vec{c}$ is then fed to a dense layer for computing the final ranking score.

$$\begin{aligned} s_{u_i} &= \underset{1\times E}{\vec{w}} \cdot \underset{E\times 1}{\vec{x_{u_i}}}, \\ a_{u_i} &= \frac{e^{s_{u_i}}}{\sum_{u_k \in q} e^{s_{u_k}}}, \\ \underset{M\times 1}{\vec{c}} &= \sum_{u_i \in q} \left(\underset{1\times 1}{a_{u_i}} \times \underset{M\times 1}{\vec{c_{u_i}}}\right). \end{aligned} \tag{3}$$

**Optimization.** We used the pairwise *hinge loss* as the objective function to be minimized by the *AdaDelta* optimizer. In this perspective, the training data is viewed as a set of triples, $(q, d^+, d^-)$, composed of a query $q$, a positive document $d^+$ and a negative document $d^-$. Additionally, inspired by [14] and as successfully demonstrated by [16], we adopted a similar negative sampling strategy, where a negative document can be drawn from the following sets:

– **Partially irrelevant set**: Irrelevant documents that share some matching signals with the query. More precisely, this corresponds to documents

retrieved by the fast retrieval module but which do not appear in the training data as positive examples;
- **Completely irrelevant set**: Documents not in the positive training instances and not sharing any matching signal with the query.

### 2.2 Passage Extraction Details

Passage extraction is accomplished by looking at the attention weights of the neural ranking model. As described, the proposed neural ranking model includes two attention mechanisms. The first one computes a local passage attention with respect to each query term, $a_{p_i}$. The second is used to compute the importance of each query term, $a_{u_k}$. Therefore, a global attention weight for each passage can be obtained from the product of these two terms, $a_{g_{(k,i)}} = a_{u_k} \times a_{p_i}$, as shown in Eq. 4:

$$\begin{aligned} \underset{M\times 1}{\vec{c}} &= \sum_{u_k \in q} \left( \underset{1\times 1}{a_{u_k}} \times \sum_{p_i \in D(u_k)} \left( \underset{1\times 1}{a_{p_i}} \times \underset{M\times 1}{\vec{h}_{p_i}} \right) \right) \\ &= \sum_{u_k \in q} \left( \sum_{p_i \in D(u_k)} \left( \underbrace{\underset{1\times 1}{a_{u_k}} \times \underset{1\times 1}{a_{p_i}}}_{global\ attention} \times \underset{M\times 1}{\vec{h}_{p_i}} \right) \right). \end{aligned} \tag{4}$$

## 3 Results and Discussion

This section presents the system evaluation results. We used the training data from the BioASQ 6b and 7b phase A challenges [13], containing 2251 and 2747 biomedical questions with the corresponding relevant documents, taken from the MEDLINE repository. The objective for a system is to retrieve the ten most relevant documents for each query, with the performance evaluated in terms of **Map@10** on five test sets containing 100 queries each.

### 3.1 Experiments

At first, a study was conducted to investigate the performance of the proposed neural ranking model. After that, the full system was compared against the results of systems submitted to the BioASQ 6 and 7 editions for the document retrieval task. Finally, we investigate if the attention given to each passage is indeed relevant.

In the results, we compare two variants of DeepRank: **BioDeepRank** refers to the model with the modified aggregation network and weighting mechanism, and using word embeddings for the biomedical domain [15]; **Attn-BioDeepRank** refers to the final model that additionally replaces the recurrent layer by a self-attention layer.[2]

[2] Configuration details of both variants, including all the hyperparameters used, are available in the code repository.

**Neural Ranking Models.** We compared both neural ranking versions against BM25 in terms of MAP@10 and Recall@10, on a 5-fold cross validation over the BioASQ training data. Table 1 summarizes the results.

Both models successfully improved the BM25 ranking order, achieving an increase of around 0.14 in MAP and 0.31 in recall. Results of Attn-BioDeepRank, although lower, suggest that this version is at least nearly as effective at ranking the documents as the model that uses the recursive layer.

**Table 1.** Evaluation of the retrieval models on 5-fold cross validation on the BioASQ 7b dataset. Results are presented as the average ± standard deviation over the 5 validation folds.

| | BioASQ 7b | |
|---|---|---|
| | MAP | RECALL |
| BM25 | 0.153 ± 0.006 | 0.329 ± 0.013 |
| BioDeepRank | 0.298 ± 0.008 | 0.643 ± 0.035 |
| Attn-BioDeepRank | 0.289 ± 0.009 | 0.639 ± 0.038 |

**Biomedical Document Retrieval.** We report results on the BioASQ 6b and BioASQ 7b document ranking tasks (Table 2). Regarding BioASQ 6b, it should be noted that the retrieved documents were evaluated against the final gold-standard of the task, revised after reevaluating the documents submitted by the participating systems. Since we expect that some of the retrieved documents would have been revised as true positives, the results presented can be considered a lower bound of the system's performance. For BioASQ 7b, the results shown are against the gold-standard before the reevaluation, since the final annotations were not available at the time of writing. In this dataset both systems achieved performance nearer to the best result, including a top result on Batch 1.

**Table 2.** Evaluation of the retrieval models on BioASQ 6b and 7b test sets

| **6B** Systems | Batch 1 | | Batch 2 | | Batch 3 | | Batch 4 | | Batch 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAP | RANK | MAP | RANK | MAP | RANK | MAP | RANK | MAP | RANK |
| Best result | 0.2327 | – | 0.2512 | – | 0.2622 | – | 0.1843 | – | 0.1464 | – |
| BioDeepRank | 0.2051 | (5/15) | 0.2065 | (11/22) | 0.1857 | (19/24) | 0.1554 | (11/21) | 0.1116 | (17/23) |
| Attn-DeepRank | 0.1944 | (5/15) | 0.2080 | (10/22) | 0.2071 | (15/24) | 0.1556 | (11/21) | 0.1210 | (12/23) |
| **7B** Systems | Batch 1 | | Batch 2 | | Batch 3 | | Batch 4 | | Batch 5 | |
| | MAP | RANK | MAP | RANK | MAP | RANK | MAP | RANK | MAP | RANK |
| Best result | 0.0809 | – | 0.0849 | – | 0.1199 | – | 0.1034 | – | 0.0425 | – |
| BioDeepRank | 0.0874 | (1/12) | 0.0760 | (7/23) | 0.1006 | (6/21) | 0.0922 | (5/17) | 0.0344 | (9/18) |
| Attn-BioDeepRank | 0.0865 | (1/12) | 0.0764 | (7/23) | 0.0995 | (6/21) | 0.0882 | (6/17) | 0.0373 | (3/18) |

**Passage Evaluation.** Finally, we analysed whether the information used by the model for ranking the documents, as given by the attention weights, corresponded to relevant passages in the gold-standard. For this, we calculated the precision of the passages, considering overlap with the gold-standard, and evaluated how it related to the confidence assigned by the model. Interestingly, although the model is not trained with this information, the attention weights seem to focus on these relevant passages, as indicated by the results in Fig. 3.

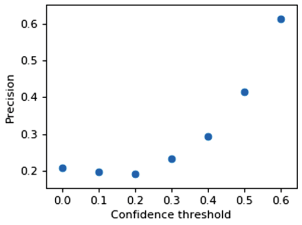


**Fig. 3.** Quality of retrieved passages as a function of the confidence attributed by the model.

## 4 Conclusion

This paper describes a new neural ranking model based on the DeepRank architecture. Evaluated on a biomedical question answering task, the proposed model achieved similar performance to a range of others strong systems.

We intend to further explore the proposed approach by considering semantic matching signals in the fast retrieval module, and by introducing joint learning for document and passage retrieval.

The network implementation and code for reproducing these results are available at https://github.com/bioinformatics-ua/BioASQ.

**Acknowledgments.** This work was partially supported by the European Regional Development Fund (ERDF) through the COMPETE 2020 operational programme, and by National Funds through FCT – Foundation for Science and Technology, projects PTDC/EEI-ESS/6815/2014 and UID/CEC/00127/2019.

## References

1. Brokos, G.I., Liosis, P., McDonald, R., Pappas, D., Androutsopoulos, I.: AUEB at BioASQ 6: Document and Snippet Retrieval, September 2018. http://arxiv.org/abs/1809.06366
2. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805

3. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM 2016, pp. 55–64. ACM Press, New York(2016). https://doi.org/10.1145/2983323.2983769, http://dl.acm.org/citation.cfm?doid=2983323.2983769
4. Hirschman, L., Gaizauskas, R.: Natural language question answering: the view from here. Nat. Lang. Eng. 7(04), 275–300 (2001). https://doi.org/10.1017/S1351324901002807, http://www.journals.cambridge.org/abstract_S1351324901002807
5. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management - CIKM 2013, pp. 2333–2338. ACM Press, New York (2013). https://doi.org/10.1145/2505515.2505665, http://dl.acm.org/citation.cfm?doid=2505515.2505665
6. Lin, Z., Feng, M., dos Santos, C.N., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. CoRR abs/1703.03130 (2017). http://arxiv.org/abs/1703.03130
7. Mateus, A., González, F., Montes, M.: Mindlab neural network approach at bioasq 6b, November 2018. 10.18653/v1/W18-5305
8. McDonald, R., Brokos, G.I., Androutsopoulos, I.: Deep Relevance Ranking Using Enhanced Document-Query Interactions, September 2018. http://arxiv.org/abs/1809.01682
9. Nentidis, A., Krithara, A., Bougiatiotis, K., Paliouras, G., Kakadiaris, I.: Results of the sixth edition of the BioASQ challenge. In: Proceedings of the 6th BioASQ Workshop A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering, pp. 1–10. Association for Computational Linguistics, Brussels, November 2018. https://doi.org/10.18653/v1/W18-5301, https://www.aclweb.org/anthology/W18-5301
10. Pang, L., Lan, Y., Guo, J., Xu, J., Xu, J., Cheng, X.: DeepRank. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM 2017, pp. 257–266. ACM Press, New York (2017). https://doi.org/10.1145/3132847.3132914, http://dl.acm.org/citation.cfm?doid=3132847.3132914
11. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389, April 2009. https://doi.org/10.1561/1500000019, http://dx.doi.org/10.1561/1500000019
12. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: A latent semantic model with convolutional-pooling structure for information retrieval. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM 2014, pp. 101–110. ACM Press, New York (2014). https://doi.org/10.1145/2661829.2661935, http://dl.acm.org/citation.cfm?doid=2661829.2661935
13. Tsatsaronis, G., et al.: An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinform. **16**, 138 (2015). https://doi.org/10.1186/s12859-015-0564-6
14. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. CoRR abs/1404.4661 (2014). http://arxiv.org/abs/1404.4661

15. Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z.: BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci. Data **6**(1), 52 (2019). https://doi.org/10.1038/s41597-019-0055-0
16. Zhu, M., Ahuja, A., Wei, W., Reddy, C.K.: A hierarchical attention retrieval model for healthcare question answering. In: The World Wide Web Conference, pp. 2472–2482. WWW 2019. ACM, New York (2019). https://doi.org/10.1145/3308558.3313699, http://doi.acm.org/10.1145/3308558.3313699