Neural-based snippet extraction for biomedical question answering

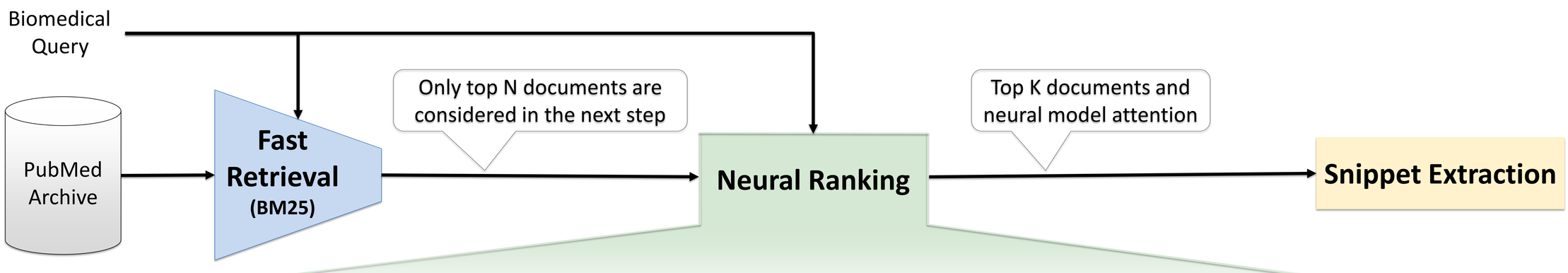
Tiago Almeida and Sérgio Matos

University of Aveiro, DETI/IEETA, 3810-193 Aveiro, Portugal

Introduction

- Literature growth poses challenges to biomedical researchers, who need to routinely examine a wide amount of scientific documents.
- Current search engines such as PUBMED do not support natural language queries, and display results as lists of documents that users must inspect to find the desired information.
- This work presents an end-to-end retrieval system, applied to the biomedical domain, that combines the efficiency of a traditional model with the efficacy of a neural ranking model in order to retrieve relevant documents with their relevant snippets highlighted.

Proposed System



Detection Network

Given a query and a document, this network finds relevant snippets that match each query terms and creates a matrix *S* for each match (query-snippet), where each entry corresponds to the cosine similarity between the embeddings of a i^{th} query token and a j^{th} snippet token, $\overrightarrow{u_i}$ and $\overrightarrow{v_j}$ respectively.

$$S_{ij} = \frac{\overrightarrow{u_i}^T \cdot \overrightarrow{v_j}}{\|\overrightarrow{u_i}\| \times \|\overrightarrow{v_j}\|}$$

Measurement Network

This network uses a 2D convolution followed by a global max polling operation, in order to capture the local relevance present in each matrix S.

$$h_{i,j}^m = \sum_{l=0}^{x-1}\sum_{n=0}^{y-1} w_{s,t}^m \times S_{i+l,j+n} + b^m$$

$$h^m = \max_{i,j}(h_{i,j}^m),\ m = 1,\ldots,M$$

Vector $\vec{h}$ captures local relevance using K convolution kernels with size *x-by-y*.

A self-attention layer is used to aggregate query-snippet pairs for each query term u_i, described by the set $D(u_i)$.

$$s_{p_j} = \underset{1\times A}{w^T} \cdot tanh\left(\underset{A\times M}{W} \cdot \underset{M\times 1}{\vec{h}_{p_j}}\right)$$

$$a_{p_j} = \frac{e^{s_{p_j}}}{\sum_{p_k \in D(u_i)} e^{s_{p_k}}}$$

$$\underset{M\times 1}{\vec{c}_{u_i}} = \sum_{p_j \in D(u_i)} \left(\underset{1\times 1}{a_{p_j}} \times \underset{M\times 1}{\vec{h}_{p_j}}\right)$$

Here, a_{p_j} are normalized attention weights for the j^{th} snippet token with respect to the query term u_i.

Aggregation Network

Each vector $\overrightarrow{c_{u_i}}$ is weighted by the relative importance of the respective query term, a_{u_i}, calculated from its embedding representation.

$$s_{u_i} = \underset{1\times E}{\vec{w}} \cdot \underset{E\times 1}{\vec{x}_{u_i}}$$

$$a_{u_i} = \frac{e^{s_{u_i}}}{\sum_{u_k \in q} e^{s_{u_k}}}$$

$$\underset{M\times 1}{\vec{c}} = \sum_{u_i \in q} \left(\underset{1\times 1}{a_{u_i}} \times \underset{M\times 1}{\vec{c}_{u_i}}\right)$$

The resulting vector is fed to a dense layer to compute the final ranking score.

Snippet Extraction

This is accomplished by looking at both attention weights of the neural ranking model (query and query-snippet). The global attention weight for each snippet can be derived from the product of these two terms, $a_{g_{(i,j)}} = a_{u_i} \times a_{p_j}$.

Experiments and Results

The BioASQ 7 dataset [2] (18M articles and 2747 questions) was used to train and evaluate the system. Compared against the original BM25 ranking order, the proposed system achieved an improvement of 0,14 in MAP and 0,31 in recall. The system was also evaluated with the available BioASQ 7 test sets, achieving performance levels comparable to the best [1], including a top result on Batch 1. Figure 1 displays a prototype web application that exposes this system.

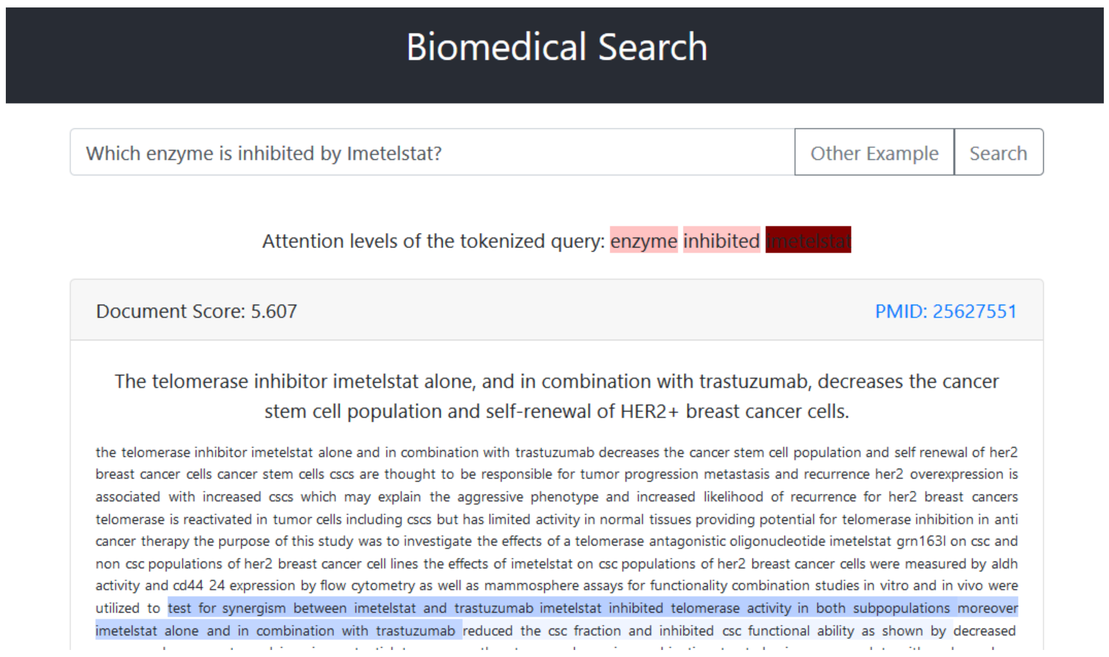


Figure 1: Biomedical literature retrieval tool, with snippet highlighting.

Conclusion

- The overall end-to-end retrieval system shows promising results when applied to the biomedical domain.
- The neural ranking model allows exploring the idea of snippet extraction with respect to the final document score.

References

[1] Lian Pang, et al.: Deeprank. In *Proceedings of the 2017 ACM on conference on Information and Knowledge Management – CIKM' 17*.

[2] Tsatsaronis, G., et al.: An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 16, 138 (04 2015).

1 Complete results can be found here: https://tinyurl.com/y2kuu26b

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