

Neural-based snippet extraction for biomedical question answering

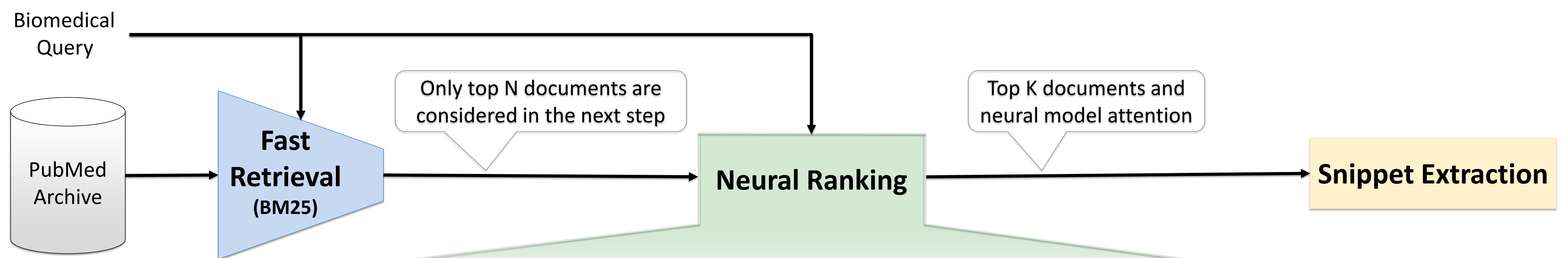
Tiago Almeida and Sérgio Matos

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, \vec{u}_i and \vec{v}_j respectively.

$$S_{ij} = \frac{\vec{u}_i^T \cdot \vec{v}_j}{\|\vec{u}_i\| \times \|\vec{v}_j\|}$$

Measurement Network

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

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

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 $D(u_i)$.

$$s_{p_j} = w^T \cdot \tanh \left(W \cdot \vec{h}_{p_j} \right)$$
$$a_{p_j} = \frac{e^{s_{p_j}}}{\sum_{p_k \in D(u_i)} e^{s_{p_k}}}$$
$$\vec{c}_{u_i} = \sum_{p_j \in D(u_i)} \left(a_{p_j} \times \vec{h}_{p_j} \right)$$

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 \vec{c}_{u_i} is weighted by the relative importance of the respective query term, a_{u_i} , calculated from its embedding representation.

$$s_{u_i} = \vec{w} \cdot \vec{x}_{u_i}$$
$$a_{u_i} = \frac{e^{s_{u_i}}}{\sum_{u_k \in q} e^{s_{u_k}}}$$
$$\vec{c} = \sum_{u_i \in q} \left(a_{u_i} \times \vec{c}_{u_i} \right)$$

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 best¹, including a top result on Batch 1. Figure 1 displays a prototype web application that exposes this system.

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.

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

Biomedical Search

Which enzyme is inhibited by imetelstat? Other Example

Attention levels of the tokenized query: enzyme inhibited **metelstat**

Document Score: 5.607 PMID: 25627551

The telomerase inhibitor imetelstat alone, and in combination with trastuzumab, decreases the cancer stem cell population and self-renewal of HER2+ breast cancer cells.

the telomerase inhibitor imetelstat alone and in combination with trastuzumab decreases the cancer stem cell population and self renewal of her2 breast cancer cells cancer stem cells cscs are thought to be responsible for tumor progression metastasis and recurrence her2 overexpression is associated with increased cscs which may explain the aggressive phenotype and increased likelihood of recurrence for her2 breast cancers telomerase is reactivated in tumor cells including cscs but has limited activity in normal tissues providing potential for telomerase inhibition in anti cancer therapy the purpose of this study was to investigate the effects of a telomerase antagonistic oligonucleotide imetelstat gm163l on csc and non csc populations of her2 breast cancer cell lines the effects of imetelstat on csc populations of her2 breast cancer cells were measured by aldh activity and cd44 24 expression by flow cytometry as well as mammosphere assays for functionality combination studies in vitro and in vivo were utilized to test for synergism between imetelstat and trastuzumab imetelstat inhibited telomerase activity in both subpopulations moreover imetelstat alone and in combination with trastuzumab reduced the csc fraction and inhibited csc functional ability as shown by decreased

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

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).