Contextual embedding and model weighting by fusing domain knowledge on Biomedical Question Answering

Yuxuan Lu  
Beijing University of Technology  
Chaoyang Qu, Beijing Shi, China  
luyuxuanleo@gmail.com

Jingya Yan  
Beijing University of Technology  
Chaoyang Qu, Beijing Shi, China  
yanjy1998@163.com

Zhixuan Qi  
Beijing University of Technology  
Chaoyang Qu, Beijing Shi, China  
zhixuanqi@outlook.com

Zhongzheng Ge  
Beijing University of Technology  
Chaoyang Qu, Beijing Shi, China  
bubble327670@gmail.com

Yongping Du  
Beijing University of Technology  
Chaoyang Qu, Beijing Shi, China  
ypdue@bjut.edu.cn

ABSTRACT

Biomedical Question Answering aims to obtain an answer to the given question from the biomedical domain. Due to its high requirement of biomedical domain knowledge, it is difficult for the model to learn domain knowledge from limited training data. We propose a contextual embedding method that combines open-domain QA model AoA Reader and BioBERT model pre-trained on biomedical domain data. We adopt unsupervised pre-training on large biomedical corpus and supervised fine-tuning on biomedical question answering dataset. Additionally, we adopt an MLP-based model weighting layer to automatically exploit the advantages of two models to provide the correct answer. The public dataset biomRc constructed from PubMed corpus is used to evaluate our method. Experimental results show that our model outperforms state-of-the-art system by a large margin.

CCS CONCEPTS

• Applied computing → Bioinformatics; • Computing methodologies → Natural language processing.

KEYWORDS

biomedical question answering, contextual embedding, model weighting, domain knowledge.

1 INTRODUCTION

Question answering is a classic task in Natural Language Processing, requiring a model to understand natural languages. Cloze-style question answering problem has been a popular task because it is relatively easier to build cloze-style datasets. The cloze style question aims to select the best candidate answer regarding the specified context and fill in the blank of the question. Multiple cloze-style datasets have been published, such as CNN/Daily Mail [5], Children’s Book Test [6], etc. Datasets focused on biomedical domain have been published, such as BioASQ [12], BioREAd [9], BioMRC [10] and CORD-19 [11] Models based on neural networks are then proposed, such as AS READER [7], CAS READER [3], AoA Reader [2] and BERT [4]. These models have achieved good performance on several datasets. However, they do not perform well when facing domain-oriented problems. The main reason is that domain-oriented questions require more background knowledge to give an answer, and a large dataset is needed to allow the models to learn the required domain knowledge.

We make improvements to the existing model, AoA Reader, and validate our results on the BioMRC dataset [10], the public biomedical dataset constructed from corpus from PubMed. We put forward the Contextual Word Embedding method and the MLP-based model weighting strategy for the biomedical question answering task. By combining the open-domain QA model and domain-oriented contextual word embedding, the proposed method outperforms state-of-the-art system on biomedical domain question answering significantly, setting up a new state-of-the-art system.

The main contributions of this paper are listed as follows:

• Combining BioBERT and AoA Reader, which can take full advantage of contextual word embedding model pre-trained on large domain corpus and mining semantic and contextual information to choose the best answer. In particular, multiple aggregation methods are adopted and evaluated.

• An MLP-based model weighting strategy is proposed, which can automatically learn the preferences and biases of different models and exploit the advantages of both models to provide the correct answer.

• Our method is evaluated on the BioMRC dataset, and the results show that it outperforms state-of-the-art system significantly. Our code is available at https://github.com/leoleosd/MLP-based-weighting.

2 METHOD

We propose a pre-training strategy based on the scientific pre-training model (SciBERT) and open-domain QA model (AoA Reader) to obtain the final answer to the question. In particular, different embedding and weighting strategies are used in the training process. Fig. 1 shows the full structure of our model.
2.1 Formal Task Description

This model is aiming at tasks that comprise cloze-style questions. This task can be formalized as a triplet \((C, Q, \mathcal{A})\) that is inclusive of the given context \(C = \{w_1, w_2, \ldots, w_n\}\) made of words \(w_i\), a query \(Q = \{q_1, q_2, \ldots, [MASK], \ldots, q_m\}\) where the special token [MASK] marks the position where the answers are supposed to be placed, and answer candidates \(\mathcal{A} = \{a_1, a_2, \ldots, a_n\}\). A function \(F\) is expected to be learned by the model to predict the answer \(\mathcal{A}\) of question \(Q\) based on its comprehension of the proffered context \(C\):

\[
\forall a \in \mathcal{A}, P(a|C, Q) = \begin{cases} 
1 & a \text{ is the correct answer} \\
0 & a \text{ is not the correct answer} 
\end{cases} 
\]

(1)

\[
F(C, Q, \mathcal{A}) = \max_{a \in \mathcal{A}} P(a|C, Q) 
\]

(2)

2.2 Training of SciBERT

We use the SciBERT [1], which has been pre-trained on the Semantic Scholar corpus. The unsupervised pre-training process using the large-scale corpus allows the model to obtain the semantic information of biomedical texts. Further, in order to let the model adapt to the cloze-style question answering task, BioMRC dataset is used to fine-tune the model. We adopt the answer extraction strategy Pappas et al. used in their BioMRC dataset [10]. For each context-question pair, we first divide the context into sentences. Each sentence is concatenated to the question by [SEP] token, and they are fed to SciBERT respectively. In this way, we obtain the top-level embedding of the candidate entities and the placeholder in the question. The embeddings of each entity in the sentence are connected to the placeholder’s embedding and are sent to a multi-layer perceptron to obtain the score for the particular entity. If an entity appears multiple times in the paragraph, we choose the maximum value of its score.

2.3 Training of AoA Reader

In order to make AoA Reader [2] achieve better performance on domain-oriented tasks, we adopt different contextualized word embedding and attention aggregation strategies.

2.3.1 Contextualized word embedding. In order to make AoA Reader [2] aware of biomedical domain knowledges and terms, we adopt BioBERT [8] to generate contextualized word embedding. Domain knowledge and the meaning of biomedical terms can be learned during the pre-training process. For tokenization, WordPiece [13] through which new words can be represented by known tokens is used. Further, it can solve the Out-Of-Vocabulary issue and allow the model to better understand domain terms made by word-formation methods.

The context and the question is then fed into BioBERT to obtain \(E(C, Q)\), which is the contextual embedding of both the context \(C\) and the query \(Q\).

To obtain the embeddings of context \(C\) and the query \(Q\), we’ve applied a masking operation on \(E(C, Q)\) for segmentation. This would conceal the representation of the other segment by zero vectors, leaving the desired half acquired:

\[
E(C)_i = \begin{cases} 
E(C, Q)_i & i \text{ is a context token} \\
0 & i \text{ is a question token} 
\end{cases} 
\]

(3)

\[
E(Q)_i = \begin{cases} 
E(C, Q)_i & i \text{ is a question token} \\
0 & i \text{ is a context token} 
\end{cases} 
\]

(4)

We adopt bi-directional RNN to further obtain the contextual representations \(h_{context}\) of the context \(C\) and \(h_{question}\) of the question \(Q\).

2.3.2 Pair-wise Matching Score. Pairwise matching matrix which indicates the relevance between a token in the context and question is calculated by calculating their dot product:

\[
M(i, j) = h_{context}(i)^T \cdot h_{question}(j) \quad i \in C, \ j \in Q. 
\]

(5)

2.3.3 Attentions over Attention Mechanism. To get the attended context-level attention, the Attentions over Attention Mechanism [2] is used. We first calculate the context-level attention regarding each token in the question:

\[
\alpha(t) = \text{softmax}(M(1, t), \ldots, M(|C|, t)) 
\]

\[
\alpha = [\alpha(1), \alpha(2), \ldots, \alpha(|Q|)] . 
\]
Then the question-level attention $\beta$ is calculated by an average on row-wise softmax:

$$
\beta(t) = \text{softmax}(M(t, 1), M(t, 2), \ldots, M(t, |Q|))
$$

(8)

$$
\beta = \frac{1}{n} \sum_{t=1}^{n} \beta(t).
$$

(9)

And finally we adopt the attention-over-attention mechanism, by merging these two attentions to get the "attended context-level attention":

$$
s = \alpha^T \cdot \beta
$$

(10)

where $s$ denotes the importance of each token in the context.

2.3.4 Answer Predictions. The AoA Reader model is used to predict the answer of the question. It uses sum attention mechanism proposed by Kadlec et al. [7] to get the confidence score of each candidate entity. However, in our model which uses WordPiece to obtain contextualized word embeddings, an entity may be either segmented into multiple tokens or composed of multiple words, and each token of the entity may occur multiple times in the context. So the confidence score of each candidate answer $a$ is calculated by aggregating all the occurrences of all its tokens in the context:

$$
P(a|C, Q) = \frac{F_1}{|T(a)|} • \frac{F_2}{|I(t, C)|}
$$

where $T(a)$ is the result of segmenting the candidate answer $a$ using WordPiece; $F_1$ and $F_2$ are aggregating functions, which can be either maximum or sum, and $I(t, C)$ indicates the position that the token $t$ appears in the context $C$.

2.4 Model Weighting strategy

After completing the training of AoA Reader and SciBERT, a model weighting strategy is used to obtain the final answer by combining the advantages of both models.

The weighting process is performed by calculating a weighted average of the answer’s confidence score and the similarity of the answer derived from two models. Further, considering that different models perform differently against data with different features, we use a simple MLP with one hidden layer to allow the model weighting layer to automatically learn this difference and to take advantage of both models.

$$
\text{score} = \text{MLP}([\text{score}_a, \text{score}_b])
$$

(12)

Where $\text{score}_a, \text{score}_b \in \mathbb{R}^{|A|}$ is the confidence score of each model. In this way, the weighting layer would be able to learn the predication and biases of each candidate model and achieve better performance.

3 EXPERIMENT

3.1 Dataset and Experiment Settings

Considering both our computing resource and the stability of the result, we conduct the experiments on the BiomRC Lite dataset [10] to verify the effect of our method. We train our model on the BiomRC Lite dataset and evaluate it both on the BiomRC Lite and Tiny dataset. We use Setting A for BiomRC, in which all pseudo-identifier like @entity1 have a global scope.

3.2 Results

3.2.1 Performance of the Contextualized Word Embedding Strategy. Contextualized word embedding strategy based on BioBERT is used to obtain the final prediction answer. The selection of aggregating functions is crucial to the model performance. Therefore, multiple combinations of different aggregating functions are evaluated, and the results are shown in tab. 1.

It can be seen that choosing sum as both aggregation functions obtains better performance, and our model outperforms the state-of-the-art model significantly, which is about 6.77% absolute improvements on the BiomRC Lite test sets.

Our model also shows an improvement on the BiomRC Tiny dataset, though the dataset contains only 30 samples, and this result may be unstable. Our performance on the larger BiomRC Lite test set still exceeds the average human expert performance on the BiomRC Tiny test set.

3.2.2 Performance of the Weighting Model. The MLP-based weighting model not only gives the correct answer derived from two models, i.e., the percentage of the union of the questions answered correctly by the two models in the total number of questions.

The results when excluding data that both models failed to answer are shown in tab. 2. As expected, both of the two models correctly answer some questions that the other model failed to answer. The proposed MLP-based weighting model not only gives the correct answer to the question that at least one model answers correctly, but also a small number of questions that both models fail to answer.

In general, our MLP-based weighting model improves the performance by 1.26% significantly compared to the original single model. These results proved that the proposed method can automatically learn the biases and preferences and exploit the strengths of both models to achieve better performance.

4 CONCLUSIONS

We propose a contextual embedding and model weighting method, which can combine model pre-trained on a large corpus and open-domain QA model to mine semantic and contextual information in biomedical question answering. Especially, we adopt an MLP-based model weighting strategy which can automatically learn and

1https://github.com/leoleoasd/MLP-based-weighting
Table 1: The result of different aggregation functions, compared to the state-of-the-art model and human experts

| Method                  | Occurrence Aggregation | Token Aggregation | Train Time¹ | BIOMRC LITE | BIOMRC TINY² |
|-------------------------|------------------------|-------------------|-------------|-------------|--------------|
| AS-READER               | -                      | -                 | 16.56hr     | 62.29       | 62.38        |
| AoA-READER              | -                      | -                 | 60.90hr     | 70.00       | 69.87        |
| SCIBERT-MAX-READER      | -                      | -                 | 83.22hr     | 80.06       | 79.97        |
| HUMAN EXPERTS           | -                      | -                 | -           | -           | -            |
| AoA-READER with BioBERT Embedding (ours) | max | max | 1.50hr | 78.54 | 78.11 |
| SCIBERT-MAX-READER      | max | sum | 0.88hr | 83.40 | 83.36 |
| Human-based Weighting Model (ours) | sum | max | 3.60hr | 80.98 | 81.20 |
|                         | sum | sum | 1.70hr | 87.22 | 86.74 |

1: We conduct some code optimizations on the AoA-Reader model, so the training time of our implementation can not be compared to their original implementation.
2: The test set of BIOMRC Tiny dataset only contains 30 samples, and so the results on it may be unstable. On the other hand, the demonstrated accuracy of human experts comes from averaging the results of multiple experts, so it is a bit more stable than other results.

Table 2: The results of our MLP-based Weighting Model, excluding data that both models failed to answer

| Method                  | BIOMRC LITE | BIOMRC TINY¹ |
|-------------------------|-------------|--------------|
| AoA-READER with BioBERT Embedding (ours) | 93.71 | 93.21 |
| SCIBERT-MAX-READER      | 86.67       | 86.94        |
| MLP-based Weighting Model (ours) | 95.36 | 95.38 |

1: The test set of BIOMRC Tiny dataset only contains 30 samples, and so the results on it may be unstable.

utilize the preferences and biases of two models to combine their advantages. The results show that our method outperforms state-of-the-art system and has higher accuracy than experts. In future work, how to use the semantic similarity between entity tokens and context tokens in getting final predictions should be studied, i.e., a context token should contribute to the score of an entity if its semantic information is similar to that of entity token.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[2] Yiming Cui, Zhipei Chen, Shi Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-Attention Neural Networks for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vancouver, Canada, 593–602. https://doi.org/10.18653/v1/P17-1055

[3] Yiming Cui, Ting Liu, Zhipei Chen, Shijin Wang, and Guoping Hu. 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. The COLING 2016 Organizing Committee, Osaka, Japan, 1777–1786.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Ray, Mustafa Suleyman, et al. 2015. Teaching Machines to Read and Comprehend. In Advances in Neural Information Processing Systems, Vol. 28. Curran Associates, Inc.

[6] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations. In 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, Yoshua Bengio and Yann LeCun (Eds.).

[7] Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text Understanding with the Attention Sum Reader Network. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Berlin, Germany, 908–918. https://doi.org/10.18653/v1/P16-1086

[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, et al. 2019. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. Bioinformatics (Sept. 2019), btz682. https://doi.org/10.1093/bioinformatics/btz682

[9] Dimitris Pappas, Ion Androuloupoulos, and Haris Papageorgiou. 2018. BioRead: A New Dataset for Biomedical Reading Comprehension. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan.

[10] Dimitris Pappas, Petros Stavropoulos, Ion Androuloupoulos, and Ryan McDonald. 2020. BioMRC: A Dataset for Biomedical Machine Reading Comprehension. In Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. Association for Computational Linguistics, Online, 140–149. https://doi.org/10.18653/v1/2020.biomed-1.15

[11] Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, et al. 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. arXiv:2004.11339 [cs] (April 2020) arXiv:2004.11339 [cs].

[12] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, et al. 2015. An Overview of the BIOASQ Dataset for COVID-19. arXiv:1609.08144 (2016). arXiv:1609.08144 [cs].

[13] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, et al. 2016. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. CoRR abs/1609.08144 (2016).