Dual Model Weighting Strategy and Data Augmentation in Biomedical Question Answering

Yongping Du*, Jingya Yan*, Yiliang Zhao*, Yuxuan Lu* and Xingnan Jin*
*Faculty of Information Technology
Beijing University of Technology, Beijing, China
Email: ypdu@bjut.edu.cn;yanjy1998@163.com;ylzhao7@yeah.net;luyuxuanleo@gmail.com;jinxingnan@outlook.com

***Abstract*—Biomedical Question Answering aims to extract an answer to the given question from a biomedical context. Due to the strong professionalism of specific domain, it's more difficult to build large-scale datasets for specific domain question answering. Existing methods are limited by the lack of training data, and the performance is not as good as in open-domain settings. We propose a model weighting strategy for the final answer prediction in biomedical domain, which combines the advantage of two models, open-domain model QANet and BioBERT pre-trained in biomedical domain data. Especially, we adopt effective data augmentation strategies to improve the model performance, including round-trip translation and summarization. The public biomedical dataset collected from PubMed provided by BioASQ is used to evaluate our approach. The results show that the model performance has been improved significantly on BioASQ 6B, 7B and 8B datasets compared to the single model.**

***Index Terms*—biomedical question answering, data augmentation, summarization, model weighting**

I. Introduction

As a classic task in Natural Language Processing (NLP), question answering has achieved more attention. It is naturally used to test the machine's ability to understand natural language. Presently, many question answering datasets have been published, such as CNN/Daily Mail [1], SQuAD [2] and TriviaQA [3]. Driven by these high-quality large-scale datasets, models based on deep neural network have been proposed, such as BiDAF [4], QANet [5] and AoA [6].

Existing question answering models have achieved excellent performance and surpassed human performance in open-domain such as SQuAD challenge. However, the existed models do not perform well and still have limitations in specific domains. The main reason is that building a large-scale dataset in specific domain is difficult as it needs enough professional knowledge and domain experts are required to label the data, so the limited available training data restricts the performance of the model.

We conduct experiments on BioASQ [7], the public biomedical dataset collected from PubMed. In this paper, we put forward the weighting strategies based on two models, QANet which performs well in open-domain question answering and BioBERT [8] which has been pre-trained in large-scale biomedical texts. We aim to give full play to the advantages of the two models, and improve the generalization ability of the model and the performance of biomedical domain question answering. The answers given by two models, BioBERT and QANet, will be considered comprehensively for the final answer prediction. This method brings significant improvements compared with the original two models. Especially, several data augmentation strategies are adopted to support model for better performance.

The main contributions of this paper are as follows:

- A dual model weighting strategy is proposed, which can fully play the respective advantages of two models, and improve both the model's generalization ability and the understanding of biomedical context.
- The effective data augmentation strategies are designed to make full use of the limited training data, including round-trip translation and summarization.
- Our approach is evaluated on BioASQ dataset, and the results show that the model weighting strategy with data augmentation outperforms the single model's performance significantly in biomedical domain.

II. Method

We put forward the weighting strategy based on biomedical pre-training model BioBERT and open domain QA model QANet to get the final answer of the question. Especially, different embeddings and data augmentation strategies are used during the training process. Fig 1 shows the full structure of our method.

A. Training of BioBERT

We use the BioBERT [9] model for training, which has been pre-trained on PubMed corpus on BERT model. The PubMed corpus contains millions of biomedical literatures from MEDLINE, life science journals, and online books. Pre-training on PubMed corpus aims at obtaining the semantic information of biomedical text. The weights pre-trained on SQuAD v1.1 dataset on top of BioBERT v1.1 are chosen to let the model better apply to the task of question answering. Based on these weights, the training samples of factoid type question in BioASQ training dataset are used for fine-tuning. Here, batch size is set to 5.

B. Training of QANet

In order to make QANet achieve better performance after training, and further perform well in model weighting with BioBERT, we adopt different word embeddings and data augmentation strategies during the training on the large-scale open-domain dataset SQuAD and biomedical dataset BioASQ.

978-1-6654-0126-5/21/$31.00 ©2021 IEEE

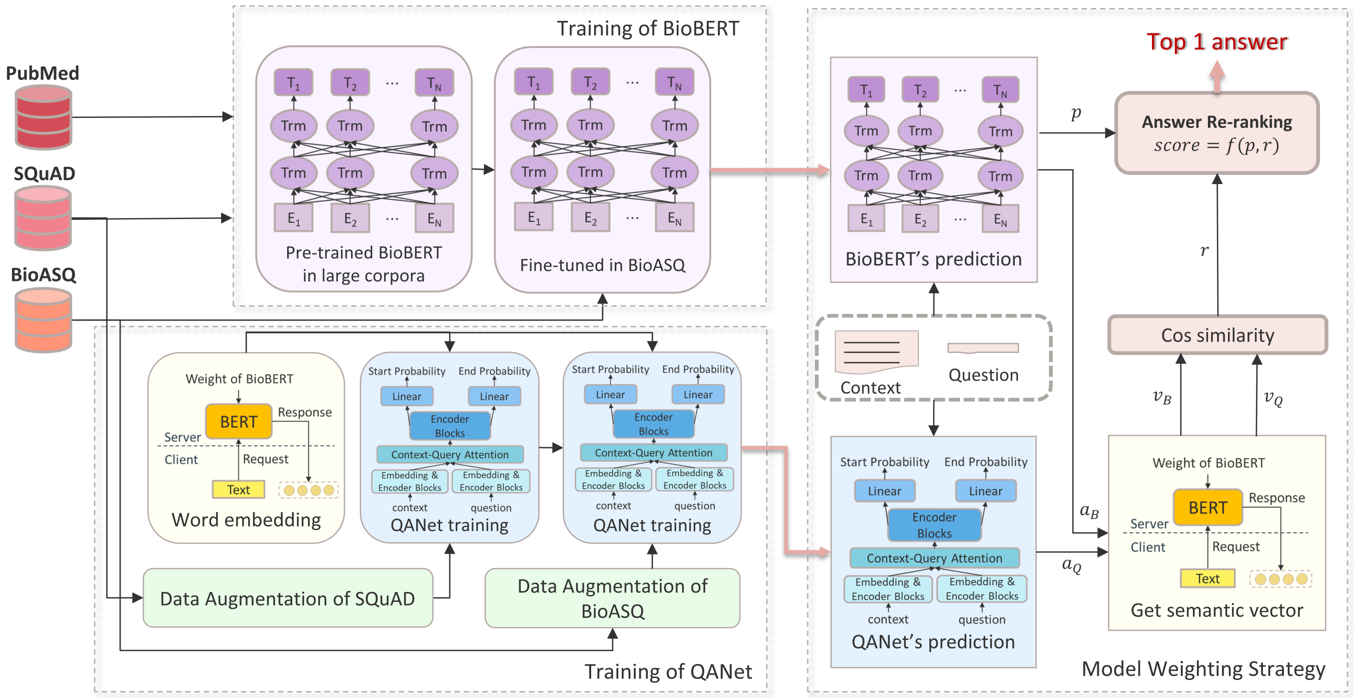


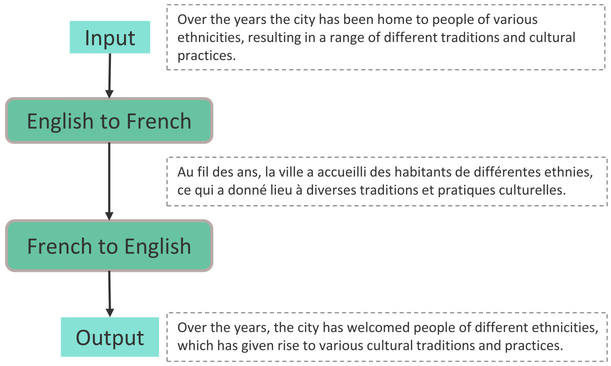
Fig. 1. Model structure based on Dual Model Weighting Strategy and Data Augmentation

1) Word embedding: QANet model uses GloVe [10] as word embedding, where most biomedical vocabularies do not appear. At the same time, some words used in open domain may have different meaning in biomedical field. To deal with this problem, the tool ”bert-as-service” [11] is used to generate new 768-dimensional word embedding (denoted as BioBERT 768 embedding) for each word in GloVe and the new words appeared in BioASQ dataset. The weights used by ”bert-as-service” are the same as the pre-training weights used by BioBERT. GloVe and BioBERT 768 embedding will be used as the word-level embedding in our following experiments.

2) Data Augmentation:

a) Round-trip Translation Method: Round-trip translation method is used to enrich SQuAD dataset [12], choosing French as the bridge language to create more training samples. After the original text in English is translated into French, it is translated into English in turn as the context of the new sample, which is illustrated in Fig 2. To find the answer to the original question in new context, character-level 2-gram method is used to get the start/end position of the new answer by comparing the first two characters of the start/end word of the original answer span with the first two characters of each word in new context.

b) Summarization: QANet model cannot deal with long context, which will be discarded when the context exceeds the default length limit. However, there are many long contexts in BioASQ dataset, and nearly 40% of the training data will be discarded due to the length limitation by QANet training. Therefore, inspired by the task of summarization, slide window method and TextRank method [13] are used to shorten the context, so as to make full use of the training data.



Over the years the city has been home to people of various ethnicities, resulting in a range of different traditions and cultural practices.

Au fil des ans, la ville a accueilli des habitants de différentes ethnies, ce qui a donné lieu à diverses traditions et pratiques culturelles.

Over the years, the city has welcomed people of different ethnicities, which has given rise to various cultural traditions and practices.

Fig. 2. Round-trip Translation Method with French as the Bridge Language

C. Model Weighting Strategy

After completing the training of QANet and BioBERT, dual model weighting strategy is adopted to predict the final answer to the question. BioBERT performs well in biomedical QA, but as the pre-training corpora are mainly unlabeled data, it's still hard for BioBERT to answer all of the questions correctly. To deal with this problem, an open-domain model QANet which has rich labeled data for training is needed to correct the predicted results of BioBERT.

BioBERT gives the top k candidate answers, and each answer carries the prediction probability p, denoted as $answerset_{BioBERT} = \{(a_{B_1}, p_1), (a_{B_2}, p_2), \cdots (a_{B_k}, p_k)\}, p_1 > p_2 > \cdots > p_k$.

We set k to 20 in our experiments. On the other hand, QANet give the unique prediction answer a_Q.

BioBERT's "bert-as-service" is used to get the semantic vector of each candidate answer, and the cosine similarity r_i between answer a_Q predicted by QANet and each candidate answer a_{B_i} predicted by BioBERT is calculated.

We use p and r to calculate new score for every candidate answer predicted by BioBERT. They will be used to re-rank k answers and the top one is regarded as the final answer of our model.

$$answer = a_{B_i}, i = argmax\left(score_i\right) \tag{1}$$

It is noticed that the higher p value indicates that BioBERT has stronger certainty of the current answer as the final answer, and it is more likely to be correct. While the lower p value shows that BioBERT is not sure about it and needs to refer to the answer given by QANet to adjust. The value of cosine similarity r is concentrated in the range of [0.7,1]. On the other hand, the value of p represents the prediction probability of BioBERT, and it has a larger variation range. In order to make the new score reflects the importance of p and r at the same time, we consider the combination of the product item and linear addition, the new score formula of the *i-th* candidate answer predicted by BioBERT is as follows:

$$\begin{gathered} score_i = \left(1-\alpha\right) p_i^x r_i^y + \alpha r_i^n \\ \left(x, y, n > 0, 0 < \alpha < 1, 1 <= i <= k\right) \end{gathered} \tag{2}$$

Here, α, x, y, n are the parameters which need to be determined later. The item $p_i^x r_i^y$ considers the influence of p value and also gives the adjustment by the value r appropriately. The item αr_i^n is used to increase the influence of r and further fine-tune the result. The value of α is used to adjust and balance the proportion of the two items.

III. Experiment

We conduct the experiments on BioASQ dataset to verify the effect of model weighting strategy and data augmentation.

TABLE I
THE STATISTICS OF SQUAD AND BIOASQ DATASETS

Dataset	SQuAD	BioASQ 6B	BioASQ 7B	BioASQ 8B
Context	18896	901	1084	1295
Question	87599	901	1084	1295
Source	Wikipedia	PubMed	PubMed	PubMed
Answer Type	Text span	Text span	Text span	Text span

A. Datasets

The statistics of the datasets used in our experiment are shown in Table 1.

SQuAD (Stanford Question Answering Dataset) is a large-scale span-based machine reading comprehension dataset, containing 107,785 question-answer pairs on 536 articles (contexts). After data augmentation, it contains over 190,000 question-answer pairs.

BioASQ is a challenge providing training data for biomedical semantic indexing and question answering task. The training sets of BioASQ 2018 (6B), 2019 (7B) and 2020 (8B) are utilized in this paper. The contexts in each sample are extracted from PubMed corpus, and the answers to the questions are denoted by biomedical experts. In our experiments, factoid questions are used for fine-tuning BioBERT, and both factoid and list questions are used for QANet's training. Our method is tested on factoid questions. Since it is found that around 70% of the factoid questions have at least one extractable answer, we view this problem as a span extraction task.

B. Experiment settings

During the process of fine-tuning BioBERT, the batch size is set to 5 and other parameters remain unchanged.

During the training of QANet, the model is firstly trained on SQuAD training set with the learning rate of 0.001, and the number of max training step is 60000. The current model will be saved every 1000 steps. The best model will be chosen to train on BioASQ dataset with the learning rate of 0.001 continually, the number of max training step is set to 4000, and the current model will be saved every 1000 steps also. Finally, the best QANet model will be chosen by the following steps.

The model is evaluated on the 5 batches of each year's BioASQ challenge dataset using the same evaluation metrics as SQuAD: Exact Match (denoted by EM below) and F1. These metrics regard each token as an evaluation unit, the number of tokens in the predicted answer appearing in the correct answer span is measured. EM value indicates the proportion of the words in the predicted answer appearing in the ground truth answer. Recall is calculated as the proportion of the words in the ground truth answer appearing in the predicted answer. The F1 value is calculated by the above

TABLE II
EM AND F1 SCORE ON BIOASQ 6B, 7B AND 8B AFTER USING MODEL WEIGHTING STRATEGY WITH DIFFERENT PARAMETERS. QANET USES BIOBERT 768 EMBEDDING, AND DIFFERENT DATA AUGMENTATION METHODS ARE USED IN THE EXPERIMENTS.

Dataset	Parameters				Results	
	x	y	n	α	EM	F1
BioASQ 6B	2	1	2	0.05	57.756	**74.164**
	2	1	2	0.1	57.756	73.995
	2	1	2	0.15	57.756	73.260
	3	2	2	0.05	**58.556**	74.062
BioASQ 7B	2	1	2	0.05	49.523	**65.760**
	2	1	2	0.1	**50.213**	64.819
	2	1	2	0.15	49.261	63.894
	3	2	3	0.1	49.261	62.884
BioASQ 8B	2	1	2	0.1	43.976	60.878
	2	1	2	0.15	43.235	60.011
	3	2	3	0.15	46.881	**62.058**
	3	2	3	0.2	**46.967**	61.473

TABLE III

EM and F1 score of BioBERT, QANet and model weighting strategy on BioASQ 6B, 7B and 8B. QANet uses BioBERT 768 embedding, and the three experiments use different data augmentation methods and model weighting parameters. The model weighting parameters are chosen from Table II.

Model	**BioASQ 6B**		**BioASQ 7B**		**BioASQ 8B**	
	EM	F1	EM	F1	EM	F1
BioBERT	53.777	70.928	**49.523**	65.247	41.071	58.956
QANet	37.889	55.035	35.571	52.057	39.917	53.940
Dual Model Weighting (Ours)	**57.756**	**74.164**	**49.523**	**65.760**	**46.881**	**62.058**

EM and recall. The experimental results below will show the average results on the 5 batches.

C. Results

The model weighting strategy is used to achieve the final prediction answer based on BioBERT and QANet. The selection of parameters has a great influence on the performance. Therefore, experiments on different parameter settings are carried out on BioASQ 6B, 7B and 8B datasets, and the results are shown in Table II.

According to the above results, the parameters with the best performance are diverse on different datasets. This is due to the use of different embedding and data augmentation methods on these three datasets to achieve the best results.

After selecting the best parameters for each experiment, the comparison results are shown in Table III. It can be seen that EM and F1 score of our dual model weighting strategy are generally improved than the original single model. Especially, EM and F1 score are significantly improved by over 4% and 3% over BioBERT model on BioASQ 6B and 8B. These results indicate that the model weighting strategy can really integrate the advantages of the two models, even if the performance of QANet is not as good as BioBERT, the prediction of QANet can still inspire BioBERT to choose a better answer.

IV. Conclusions

We propose a dual model weighting strategy, which takes full advantage of two models for answer prediction in Biomedical Question Answering. They are open-domain model QANet and BioBERT model pre-trained in biomedical domain data. Especially, we adopt different data augmentation strategies to improve the model performance, including round-trip translation and summarization. Experimental results show that our method achieves the best performance compared to the single model on BioASQ 6B, 7B, and 8B datasets. In future work, how to extract the key context with high quality using the summarization method to inspire the training of the model is a problem that needs to be further studied. In addition, the model weighting parameters could be obtained by training for each group of experiments. The parameters could be adjusted continuously according to the prediction result, so as to find a set of parameters with the best performance using model weighting strategy.

Acknowledgment

This work is supported by Beijing Natural Science Foundation under grant NO.4212013 and the National Key R&D Program of China under grant NO. 2019YFC1906002.

References

[1] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, “Teaching machines to read and comprehend,” *Advances in neural information processing systems*, vol. 28, pp. 1693–1701, 2015.

[2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ questions for machine comprehension of text,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, 2016, pp. 2383–2392.

[3] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, “TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 2017, pp. 1601–1611.

[4] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “Bidirectional attention flow for machine comprehension,” *arXiv preprint arXiv:1611.01603*, 2016.

[5] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, “QANet: Combining local convolution with global self-attention for reading comprehension,” in *International Conference on Learning Representations*, 2018.

[6] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, “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)*, Vancouver, Canada, 2017, pp. 593–602.

[7] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos *et al.*, “An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition,” *BMC bioinformatics*, vol. 16, no. 1, pp. 1–28, 2015.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[9] W. Yoon, J. Lee, D. Kim, M. Jeong, and J. Kang, “Pre-trained language model for biomedical question answering,” in *19th Joint European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD 2019*, 2020, pp. 727–740.

[10] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

[11] H. Xiao, “bert-as-service,” https://github.com/hanxiao/bert-as-service, 2018.

[12] Y. Du, W. Guo, and Y. Zhao, “Hierarchical question-aware context learning with augmented data for biomedical question answering,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, pp. 370–375.

[13] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, “Variations of the similarity function of TextRank for automated summarization,” *CoRR*, vol. abs/1602.03606, 2016.