Improving Biomedical Question Answering by Data Augmentation and Model Weighting

Yongping Du,Jingya Yan,Yuxuan Lu,Yiliang Zhao, and Xingnan Jin

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, especially degrading when facing to the adversarial sample. We try to resolve the above issues. Firstly, effective data augmentation strategies are adopted to improve the model training, including slide window, summarization and round-trip translation. Secondly, 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. Finally, we give adversarial training to reinforce the robustness of the model. The public biomedical dataset collected from PubMed provided by BioASQ challenge is used to evaluate our approach. The results show that the model performance has been improved significantly compared to the single model and other models participated in BioASQ challenge. It can learn richer semantic expression from data augmentation and adversarial samples, which is beneficial to solve more complex question answering problems in biomedical domain.

Index Terms—biomedical question answering, data augmentation, deep learning, model weighting

1 Introduction

Question Answering aims to make the machine has the ability to understand natural language. For each given question, it is needed to extract or generate the answer from the related context. The research on this task can help us extract useful information from a massive number of literatures efficiently. Question answering task can be categorized into four types: cloze test, span extraction, multiple choice and free answering. Cloze test task requires the model to fill in the gap according to the context. Span extraction task needs to find the answer span from the context. Multiple choice task selects the correct choice in the candidate list and free answering task further generates the answer that doesn't appear directly in the original context. In recent years, many question answering datasets have been published, such as CNN/Daily Mail dataset [1], SQuAD dataset [2] and so on. Driven by these high-quality large-scale datasets, models based on deep neural network have been proposed, such as BiDAF [3], QANet [4] and AoA [5].

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 [6], 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 [7] which has been pre-trained in large-scale biomedical texts. We aim to give full play to the advantages of the two models, and to 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. We aim to further improve the performance of BioBERT by fusing the QANet model which performs well in open domain, so that the model can fully mine the biomedical context information and further improve the generalization ability. This method brings significant improvements compared with the original two models. Especially, several data augmentation strategies are adopted to support the model for better performance.

The main contributions of this paper are as follows:

- The proposed model weighting strategy can fully play the respective advantages of two models, including open-domain model and pre-trained model in biomedical domain data. It improves both the model's generalization ability and the understanding of biomedical context.
- The effective data augmentation strategies are designed to further improve the performance on biomedical domain question answering, aiming to

- *Yongping Du is with the Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China. E-mail:ypdu@bjut.edu.cn.*
- *Jingya Yan is with the Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China. E-mail:yanjy1998@163.com.*
- *Yuxuan Lu is with the Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China. E-mail:luyuxuanleo@gmail.com.*
- *Yiliang Zhao is with the Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China. E-mail:ylzhao7@yeah.net.*
- *Xingnan Jin is with the Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China. E-mail:jinxingnan@outlook.com.*

Corresponding author: Jingya Yan.

make full use of the limited domain training data. The performance is improved significantly on the specific domain task.
- Adversarial training is utilized to further improve the robustness of the model. The experimental result shows that adversarial training has the potential ability to inspire the model to answer complex questions.

2 Related Work

Question Answering has achieved rapid development recently, which benefits from the publication of several large-scale and high-quality datasets. The expansion of dataset supports the training of large-scale neural network models. Seo et al. propose BiDAF [3], which represents context at different levels and uses bidirectional attention mechanism to obtain a query-aware context representation. Yu et al. propose QANet [4], whose encoder consists exclusively of convolution which models local interactions, and self-attention which models global interactions. Devlin et al. propose a new pre-training model BERT [8], which makes machine reading comprehension model surpass human performance for the first time on the SQuAD 1.1 dataset [2]. Seonwoo et al. [9] regard context prediction as an auxiliary task in multi-task learning manner, and propose block attention method for context prediction, aiming to let model predict the correct answer from the context relevant to the question.

In recent years, researchers have focused on the automatic construction of new pre-training corpora, the efficiency of the model and the processing of long contexts. Deriu et al. [10] propose a new method to construct QA dataset from structured data, which introduces middle representation operation tree, and a new corpus OTTA is created. Cao et al. [11] improve the speed of model processing by decomposing pre-trained transformers. Liu et al. propose RikiNet [12], which contains a dynamic paragraph dual-attention reader and a multi-level cascaded answer predictor and reads Wikipedia pages for natural question answering. In addition, new tasks of question answering are appearing, such as multi-hop question answering based on knowledge graph. Saxena et al. propose EmbedKGQA [13], which is particularly effective in performing multi-hop knowledge graph question answering over sparse knowledge graphs. For conversation QA, Kundu et al. [14] introduce a new follow-up question identification task, and propose a three-way attentive pooling network to capture pair-wise interactions between the associated passage, the conversation history, and a candidate follow-up question.

The existing neural network models often need large-scale training data to achieve better performance. But for the specific domain, the lack of training data makes it difficult to play the advantages of the above models, and the performance of the model declines significantly compared with the open-domain QA task. In order to resolve the issue, Lee et al. propose BioBERT [7], which is pre-trained on large-scale biomedical corpora to learn the characteristics of biomedical texts, and achieves better performance in several biomedical NLP tasks. Chen et al. [15] introduce knowledge abstraction matching method for medical question answering which contains frequent segment N-gram mining, medical knowledge abstraction, medical segment matching and answer re-retrieval. Kommaraju et al. [16] explore the suitability of unsupervised representation learning methods on biomedical text, and introduce a new pre-training task from unlabeled data designed to reason about biomedical entities in the context. Jeong et al. [17] focus on transferring the knowledge of natural language inference to biomedical QA, presenting a sequential transfer learning method based on BioBERT.

Data augmentation technology is widely used in natural language processing, so that the model can make full use of limited data resources and mine semantic information deeply. Yang et al. [18] use cross-attention supervised data augmentation to improve their retrieval model for question answering and the performance has been significantly improved. Xu et al. [19] propose a novel data augmentation approach that combines a self-trained neural retrieval model with a few-shot learned natural language understanding model for natural language generation, and this method outperforms the state-of-the-art methods. These methods inspire the model to mine richer semantic information with the help of a new model, so as to achieve the effect of data augmentation. Xu et al. [20] utilize a series of data-augmentation approaches on information-seeking dialogue systems to enable the model obtaining better representations, and achieve excellent performance. The multi-task learning strategy is adopted to make more datasets participate in the training process for data augmentation. These methods adopt the data augmentation technology on NLP tasks by utilizing the enhanced data or optimizing the model to obtain richer information. This paper adopts the data augmentation strategy on biomedical question answering task, making the limited data resource expanded and the model has a better understanding of question answering task and domain knowledge. Lichtarge et al. [21] use round-trip translations strategy for data generation on grammatical error correction task and achieve the best performance on CoNLL-2014 evaluation. We adopt this method in our work to enrich the limited training data and further improve the performance on biomedical question answering.

Adversarial attack is introduced to reveal the statistical bias in machine reading comprehension models. Kaushik et al. [22] conduct a large-scale controlled study focused on question answering, and find that models trained on adversarial data usually perform better on other adversarial datasets but worse on a diverse collection of out-of-domain evaluation datasets. To deal with this problem, adversarial training is proposed to improve the robustness of the model. Lin et al. [23] demonstrate a simple yet effective method to attack MRC models and reveal the statistical biases in these models, and find that when interfered by the irrelevant options, the performance of MRC models is degraded significantly. In our work, we adopt adversarial training strategy to investigate the robustness of our method and further explore the potential ability of the model to give answer correctly.

3 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,

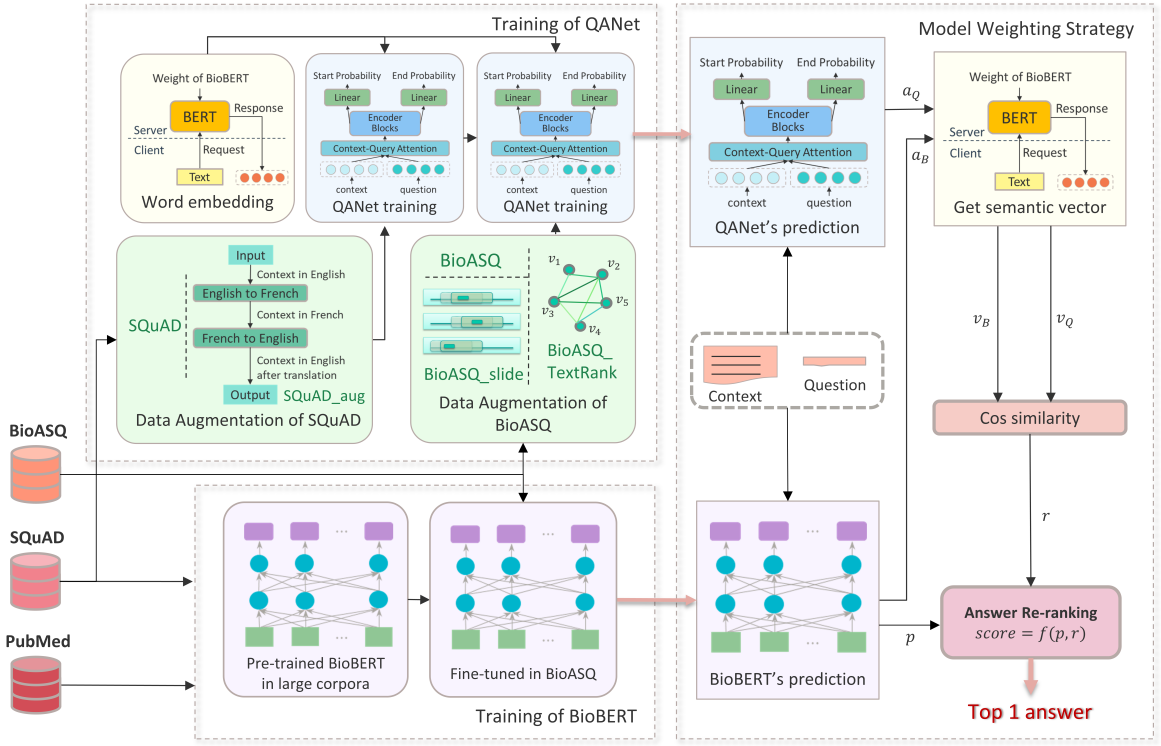


Fig. 1. Model structure based on Dual Model Weighting Strategy and Data Augmentation. After completing the training of QANet and BioBERT, dual model weighting strategy is adopted to predict the final answer to the question.

different embeddings and data augmentation strategies are used during the training process. Fig 1 shows the full structure of our model, including three main components. At first, the two models of QANet and BioBERT are trained separately. Here, three different data augmentation strategies are adopted on the SQuAD and BioASQ datasets during the model training, including round-trip translation, slide window and TextRank for summarization. And then in order to make the final prediction, model weighting strategy is adopted based on the given answers from the above two models comprehensively.

3.1 Training of BioBERT

BioBERT [24] model is adopted 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.

3.2 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. The training process is shown in Fig 2.

3.2.1 *Word embedding*

QANet model uses GloVe [25] 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" [26] 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. For the given sentence, the tokenized tool spaCy is used to get the output and the length will not be changed in the following step. And then the word embedding is achieved by GloVe and the tool "bert-as-service". Both GloVe and BioBERT 768 embedding will be used as the word-level embedding in our following experiments.

3.2.2 *Data Augmentation*

- Round-trip Translation Method

Round-trip translation method is used to enrich SQuAD dataset [27], choosing French as the bridge language to create more training samples. The original text in English is translated into French which will be translated into English in turn as the context of the new sample illustrated in Fig 3. The publicly available codebase[1] provided by Gehring [28]

1. https://github.com/facebookresearch/fairseq

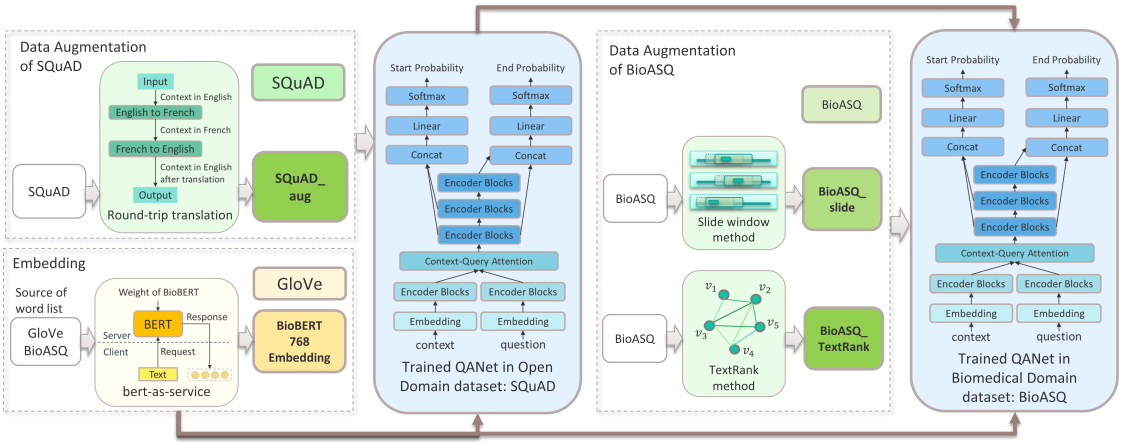


Fig. 2. Training process of QANet based on different word embeddings and data augmentation strategies, including round-trip translation, slide window and TextRank methods.

is utilized. To find the answer to the original question in the 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 the new context.

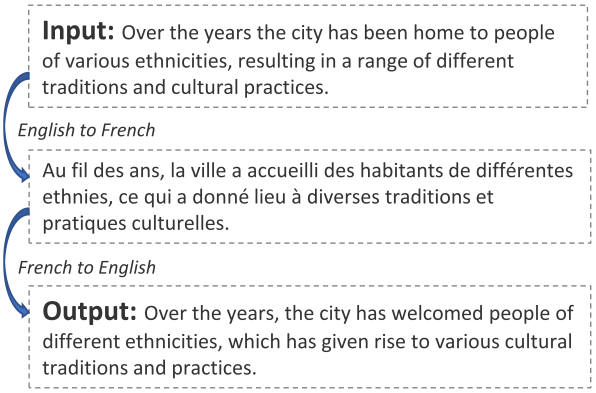


Fig. 3. Round-trip Translation Method with French as the Bridge Language.

- Slide Window Method

QANet model cannot deal with long context, which will be discarded when it 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, slide window method, which is illustrated in Fig 4, is used to shorten the context, and it will expand the number of training samples, so as to make full use of the training data. This procedure is divided into two phases: answer positioning and new context generation.

Answer Positioning: The exact index of the answer span is not given in BioASQ dataset, while the answer span may appear many times in the context of training samples. Most of these sentences containing the answer span are irrelevant to the question. It will bring interference for model training. To find the correct answer sentence, "bert-as-service" with

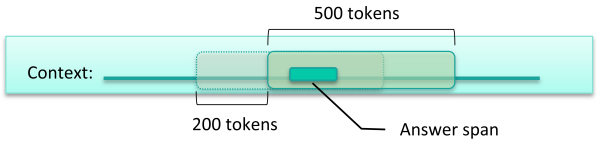


Fig. 4. Slide Window Method for shortening the context.

BioBERT pre-training parameters is used to get the semantic vectors of question sentence and each sentence in the context. The sentence that contains the answer span and has the highest cosine similarity with the question sentence will be regarded as the final answer sentence, and also the position of the answer span is recorded.

New Context Generation: The sample whose context length is less than the length limit will be used for training directly. For the sample whose context length exceeds the length limit, slide window method is used to extract the context segments with the window size of 500 tokens, and the stride of 200 tokens. According to the position of the answer span obtained in the previous step, if the current segment contains the answer, a new sample with the segment as the context will be generated. Otherwise, the current segment will be discarded and the next segment will be investigated iteratively.

- TextRank for Summarization

In order to shorten the length of the context in BioASQ dataset and make full use of training data, inspired by the task of text summarization, TextRank method [29] is used to inspire the model to extract important information from the long context, as shown in Fig 5.

It constructs a network by using the adjacent relationship between sentences, using PageRank method to calculate the weight of each sentence iteratively. Finally, the sentence with the highest weight is selected to form a summary. For the sample whose text length exceeds the limit, the summary will be generated using TextRank according to the summary ratio of 0.9 until the text length is less than 500 tokens. The sample whose context length is less than the limit will be kept directly.

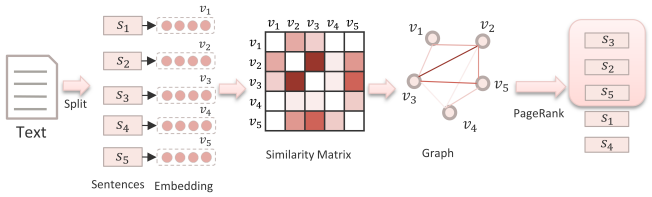


Fig. 5. TextRank method for Summarization, extracting important information from the long context.

TABLE 1
Example of the adversarial sample of BioASQ generated by ADDSENT methods.

Original sample 1	**Context:** Multimodal approaches seem to be the treatment of choice in fibromyalgia. ..... Pregabalin is, therefore, a valuable option in the first-line treatment of patients with fibromyalgia. **Question:** Which drug is considered as the first line treatment of fibromyalgia? **Answer:** pregabalin
New Context in Adversarial Sample 1	Multimodal approaches seem to be the treatment of choice in fibromyalgia. ..... Pregabalin is, therefore, a valuable option in the first-line treatment of patients with fibromyalgia. **The drug of hamster is considered as the last line treatment of migraines.**
Original sample 2	**Context:** The major phytoalexin in alfalfa is the isoflavonoid (-)-medicarpin (or 6aR, 11aR)-medicarpin. ..... Medicarpin, the major phytoalexin in alfalfa, is synthesized by way of the isoflavonoid branch of phenylpropanoid metabolism. **Question:** Which is the major phytoalexin in alfalfa (medicago sativa l.)? **Answer:** medicarpin
New Context in Adversarial Sample 2	The major phytoalexin in alfalfa is the isoflavonoid (-)-medicarpin (or 6aR, 11aR)-medicarpin. ..... Medicarpin, the major phytoalexin in alfalfa, is synthesized by way of the isoflavonoid branch of phenylpropanoid metabolism. **Hamster is the minor phytoalexin in alfalfa (medicago sativa l.).**

3.2.3 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 with rich labeled data for training is needed to refine the predicted results of BioBERT. The process of model weighting strategy is shown in Fig 6.

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 gives the unique prediction answer a_Q. The tool "bert-as-service" with the parameter of BioBERT 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_i and r_i to calculate new score for every candidate answer predicted by BioBERT. They will be used to re-rank ks answers and the top one is regarded as the final answer.

$$answer = a_{B_j}, j = argmax\left(score_j\right) \tag{1}$$

It is noticed that the higher p_i 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_i 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_i is concentrated in the range of [0.7,1]. On the other hand, the value of p_i 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_i and r_i at the same time, we consider the combination of the product item and linear addition, the new score formula of the ith candidate answer predicted by BioBERT is as follows:

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

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_i value and also gives the adjustment by the value r_i appropriately. The item r_i^n is used to increase the influence of r_i and further fine-tune the result. The value of α is used to adjust and balance the proportion of the two items.

3.2.4 Adversarial Training

Current question answering models often tend to rely on the shallow level textual information and lack deep understanding of semantics. Some researchers have introduced adversarial samples by inserting interference fragments to the context, in order to test the robustness of the model. The results show that the performance of most models degrades significantly after adding adversarial samples into training set and test set. It can be seen that there is still a gap between the existing model and human understanding ability. In addition, how to generate effective adversarial samples and improve the robustness of the model by adversarial training becomes the focus of current research.

To further investigate the robustness of our model weighting strategy and data augmentation in biomedical question answering, we use ADDSENT method [30] to generate adversarial samples on BioASQ dataset and these samples are used for training, then the performance are evaluated on the original test batches. ADDSENT method generates adversarial samples by inserting interference sentence into the context, as shown in Table 1. The process is described as follows: firstly, apply semantics-altering perturbations to the question, replace nouns and adjectives with antonyms and change named entities and numbers to the nearest word in GloVe word vector space with the same part of speech; secondly, create a fake answer that has the same "type" as the original answer; finally, combine the altered question and fake answer into declarative form, using a set of roughly 50 manually-defined rules over Stanford CoreNLP constituency parses.

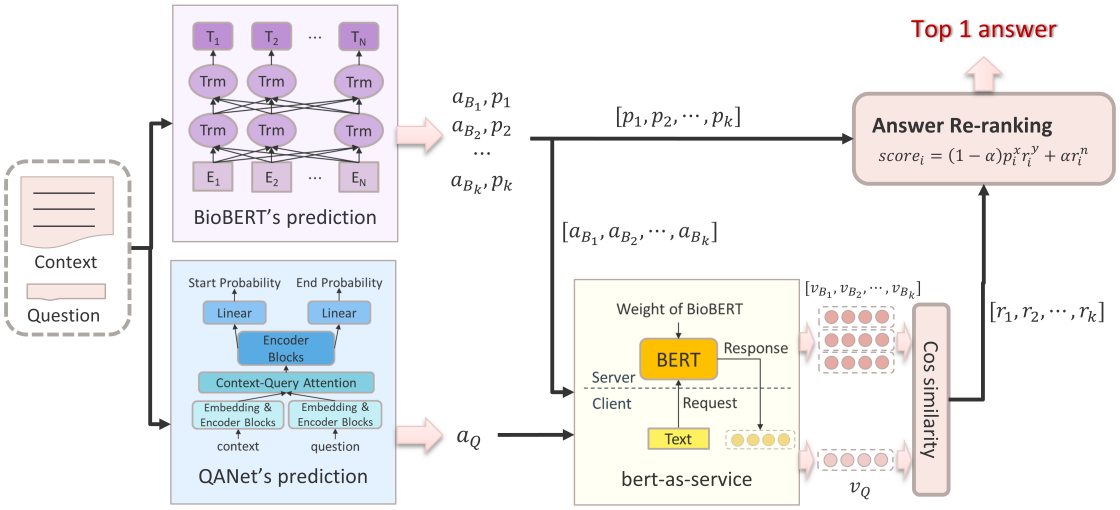


Fig. 6. Model Weighting Strategy based on two models of QANet and BioBERT, using the predictions given by these models to get the final answer.

4 Experiment

We conduct the experiments on BioASQ dataset to verify the effect of data augmentation and model weighting strategy. The comparison with other related models is demonstrated. Further, we give the detail analysis on the performance impact by weighting strategy and data augmentation.

4.1 Datasets

The statistics of the datasets used in our experiment are shown in Table 2.

TABLE 2
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

SQuAD (Stanford Question Answering Dataset) is a large-scale span-based machine reading comprehension dataset, containing 87599 question-answer pairs on 18896 contexts for training. After data augmentation by round-trip translation, it contains 173642 question-answer pairs.

BioASQ is a challenge providing training data for biomedical semantic indexing(Task A) and question answering task(Task B), and Task B takes place in two phases: Phase A (annotate questions, retrieve relevant articles, snippets, triples) and Phase B (find and report 'exact' and 'ideal' answers). The training sets of BioASQ 2018 (6B phase B), 2019 (7B phase B) and 2020 (8B phase B) 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 evaluated 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.

4.2 Experiment settings

Our experiments are carried out on the machine whose CPU is Intel(R) Core(TM) i7-7800X CPU @ 3.50GHz, GPU is GeForce GTX 1080 Ti 11G, memory is Apacer 2666MHz 8G×8, using Tensorflow-gpu(version 1.10) as the deep learning framework.

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

4.3 Results

4.3.1 *Performance Impacted by Model Weighting Strategy*

The model weighting strategy is used to achieve the final predicted answer based on BioBERT and QANet. The selection of parameters has a great influence on the performance. Therefore, we conduct k-fold cross validation experiment on the original datasets, and for the combination of different data augmentation strategies, 12 groups of hyperparameters which perform well are selected and applied

in the following model weighting process. Experiments on different parameter settings are carried out on BioASQ 6B, 7B and 8B datasets, and the results are shown in Table 3.

TABLE 3
EM and F1 score on BioASQ after using model weighting strategy with different parameters. Here, QANet uses BioBERT 768 embedding, and different data augmentation methods are used in the experiments.

Dataset	Parameters				Results	
	x	y	z	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

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 with two single models are shown in Table 4. It can be seen that EM and F1 score of our model weighting strategy are generally improved than the original single model. Especially, EM and F1 score are significantly improved by more than 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.

4.3.2 Performance Impacted by Data Augmentation Strategy

Round-trip translation method: In order to explore the influence of Round-trip translation method for data augmentation, we design experiments on BioASQ 8B for comparison. The QANet model is trained with the original SQuAD dataset and the augmentation SQuAD dataset (denoted as "SQuAD_aug") respectively. In addition, after training on the original BioASQ dataset, the model weighting strategy is used and its performance is evaluated. It can be seen in Table 5 that with the two groups of parameters settings, the results of model weighting strategy are both improved by using the SQuAD_aug dataset, which further proves that the Round-trip translation method plays a certain role in the final answer prediction.

Slide window method: The experiments are conducted to explore the influence of using the slide window method. Table 6 and 7 show the results of QANet and our model weighting strategy before and after using data augmentation of the slide window method on BioASQ 6B. Table 6 shows the results of using SQuAD dataset for QANet training, and Table 7 shows the results of using SQuAD_aug dataset for QANet training. The original number of training samples which QANet model can actually deal with is 544 due to the length limit, while there are 901 samples in BioASQ 6B training set in total. After using the slide window method, the exact number of available training samples reaches 1715. The experimental results show that using the slide window method can also bring the performance improvement. Here, QANet uses BioBERT 768 embedding, and BioASQ_slide represents the BioASQ dataset after data augmentation using the slide window method. The parameters used in the model weighting strategy are $x = 2, y = 1, n = 2, \alpha = 0.1$. In addition, the results of BioBERT are 53.777 of EM score and 70.928 of F1 score.

TextRank method: Two groups of experiments are designed to verify the effect of using TextRank method for data augmentation on BioASQ 6B and 7B. For BioASQ 6B dataset, the number of training samples which QANet

TABLE 4
EM and F1 score of BioBERT, QANet and model weighting strategy on BioASQ. Here, QANet uses BioBERT 768 embedding, and three experiments use different data augmentation methods and model weighting parameters which are chosen from Table 3.

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**

TABLE 5
The results of QANet model and our model weighting strategy before and after data augmentation on BioASQ 8B. Here, QANet uses BioBERT 768 embedding.

Datasets	QANet Results		Parameters				Dual Model Weighting Results	
	EM	F1	x	y	z	α	EM	F1
SQuAD+BioASQ	34.118	48.484	2	1	2	0.1	39.895	58.968
			2	1	2	0.15	39.154	58.010
SQuAD_aug+BioASQ	36.835	50.280	2	1	2	0.1	**42.038**	**60.847**
			2	1	2	0.15	41.297	59.340

TABLE 6
The results of QANet and our model weighting strategy before and after using data augmentation of slide window method on SQuAD and BioASQ 6B.

Datasets	QANet Results		Dual Model Weighting Results	
	EM	F1	EM	F1
SQuAD+ BioASQ	**38.951**	**55.183**	56.089	73.208
SQuAD+ BioASQ_slide	37.889	55.035	**57.756**	**73.995**
Improvement			**↑2.94%**	**↑1.08%**

TABLE 7
The results of BioBERT, QANet and our model weighting strategy before and after using data augmentation of slide window method on SQuAD_aug and BioASQ 6B.

Datasets	QANet Results		Dual Model Weighting Results	
	EM	F1	EM	F1
SQuAD_aug +BioASQ	40.006	55.291	53.777	70.324
SQuAD_aug +BioASQ_slide	**40.316**	**56.624**	**56.923**	**72.162**
Improvement			**↑5.85%**	**↑2.61%**

model can actually deal with is 544, only 60% of the original training dataset. After using TextRank to summarize the contexts in the original samples, QANet model can deal with 801 samples, and it reaches 89% of the total training dataset. The SQuAD_aug dataset is used in the experiments, and BioBERT 768 embedding is used as the word embedding by QANet model on BioASQ 6B, while the experiment on BioASQ 7B uses GloVe word embedding. The parameters used during model weighting strategy are $x = 2, y = 1, n = 2, \alpha = 0.05$. The experimental results are shown in Table 8. It indicates that TextRank method also brings the performance improvement with the model weighting strategy.

Fig 7 shows the results of our model weighting strategy using the above different data augmentation strategies on BioASQ 8B.

It can be seen that the introduction of these data augmentation strategies improves the performance by varying degrees. Especially, the slide window method has more obvious effect than the TextRank method. In our opinion,

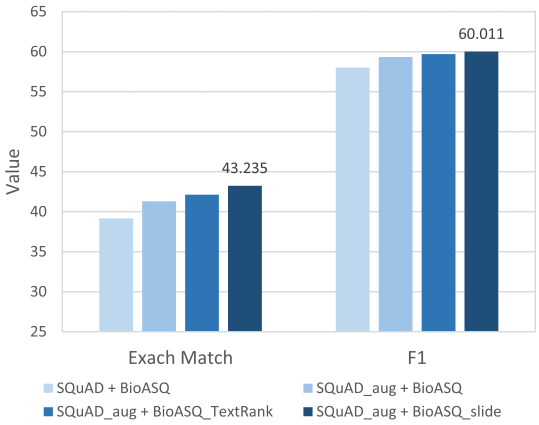


Fig. 7. The EM and F1 score of our model weighting strategy on BioASQ 8B, using different data augmentation meth ods for training on QANet model. Here, QANet uses BioBERT 768 embedding. The parameters used during model weighting strategy are $x = 2, y = 1, n = 2, \alpha = 0.15$.

the main reason is that the slide window method retains the order and coherence of the original context, and it can better reflect the contextual information. Further, the slide window method obviously expands the number of training samples, which also has a positive impact on training.

TABLE 9
Results of four baseline models and our model weighting strategy on BioASQ 6B, using the same evaluation measures with BioASQ Challenge.

Model	Results		
	Strict Acc	Lenient Acc	MRR
LabZhu, FDU	0.2141	0.3113	0.2539
AUTH-QA	0.2015	0.402	0.2713
BioASQ_Baseline	0.1296	0.2321	0.1659
OAQA	0.1629	0.2783	0.2116
Dual Model Weighting(ours)	**0.5662**	**0.7544**	**0.6436**

4.3.3 *Performance Compared with other BioASQ participants*

To compare our result with the models which participate in BioASQ challenge, we use the same metrics for evaluation: Strict Acc, Lenient Acc and MRR. Here, Table 9, Table 10 and Table 11 show the average results of 5 test batches in BioASQ

TABLE 8
The results of BioBERT, QANet and our model weighting strategy before and after using TextRank method for data augmentation on BioASQ 6B and 7B.

Test Batch	Embedding	BioBERT Results		Datasets	QANet Results		Dual Model Weighting Results	
		EM	F1		EM	F1	EM	F1
6B	BioBERT 768 embedding	53.777	70.928	SQuAD_aug + BioASQ	**40.006**	55.291	53.777	70.338
				SQuAD_aug + BioASQ_TextRank	37.055	**55.395**	**55.256**	**71.574**
7B	GloVe	49.523	65.247	SQuAD_aug + BioASQ	**34.564**	**50.557**	**49.523**	65.247
				SQuAD_aug + BioASQ_TextRank	31.659	49.799	**49.523**	**65.668**

TABLE 10
Results of four baseline models and our model weighting strategy on BioASQ 7B, using the same evaluation measures with BioASQ Challenge.

Model	Results		
	Strict Acc	Lenient Acc	MRR
KU-DMIS	0.4367	0.6274	0.5116
Google-gold-input	0.3900	0.5775	0.4562
AUTH-QA	0.2363	0.3710	0.2898
LabZhu, FDU	0.1443	0.2472	0.1867
Dual Model Weighting(ours)	**0.4884**	**0.6711**	**0.5579**

TABLE 11
Results of four baseline models and our model weighting strategy on BioASQ 8B, using the same evaluation measures with BioASQ Challenge.

Model	Results		
	Strict Acc	Lenient Acc	MRR
KoreaUniv-DMIS	0.4071	0.5967	0.4817
Umass_czi	0.3758	0.5366	0.4394
LabZhu, FDU	0.4070	0.5967	0.4766
Bio-answerfinder	0.3109	0.4014	0.3476
Dual Model Weighting(ours)	**0.4383**	**0.6917**	**0.5386**

6B, 7B and 8B respectively. The results of other models are from BioASQ official website[2]. The results indicate that our model weighting strategy outperform the models participated in BioASQ challenge significantly on solving factoid questions.

4.3.4 *Performance Impacted by Adversarial Training*

There are about 24% of the samples in BioASQ training set can be used to generate new adversarial sample by ADDSENT method. We add these samples into the original training set to train BioBERT and QANet. As shown in Table 12, in BioASQ 6B, compared with the result before using adversarial samples, the performance degrades and it indicates that adversarial samples generally interfere with the model. At the same time, model weighting strategy still performs better than the single model, which shows our method has better robustness.

Especially, the result in Table 13 shows the performance of the model is generally improved significantly after adding adversarial samples on BioASQ 8B. It indicates that the introduction of adversarial samples can make the model learn richer semantic expression from these samples. It is beneficial to solve more complex question answering problems in 8B dataset. This group of experimental result further proves the effect of adversarial learning. How to generate better adversarial samples, making the model learn deep semantic information and own stronger stability, is a subject that can be studied further.

2. http://participants-area.bioasq.org/results/

TABLE 12
The comparison results of BioBERT, QANet and our model weighting strategy before and after adding adversarial samples for training on BioASQ 6B. Here, QANet adopts BioBERT 768 embedding and uses SQuAD_aug dataset for training, also slide window method is used for data augmentation on BioASQ.

Model	Results before adding Adv.		Results after adding Adv.	
	EM	F1	EM	F1
BioBERT	53.777	70.928	50.168	69.21
QANet	40.316	56.624	40.12	61.65
Dual Model Weighting (Ours)	**56.923**	**72.162**	**51.801**	**71.132**

TABLE 13
The results of BioBERT, QANet and our model weighting strategy before and after adding adversarial samples for training on BioASQ 8B. Here, QANet adopts BioBERT 768 embedding and uses SQuAD dataset for training, also the TextRank method is used for data augmentation on BioASQ.

Model	Results before adding Adv.		Results after adding Adv.	
	EM	F1	EM	F1
BioBERT	41.071	58.956	45.055 (↑9.70%)	62.908 (↑6.70%)
QANet	36.062	49.389	34.611 (↓4.02%)	51.466 (↑4.21%)
Dual Model Weighting (Ours)	**41.409**	**61.042**	**45.855 (↑10.74%)**	**63.477 (↑3.94%)**

4.3.5 *Model Training Analysis*

We try to use multiple open domain datasets for training and the results show that the performance gets worse on biomedical question answering task as shown in Fig 8. SQuAD is a classic open domain QA dataset of span extraction task, and the experiment indicates that the model has achieved good performance by using SQuAD for training. Our method aims to further enrich the SQuAD dataset by using Round-trip Translation(RTT) method for data augmentation.

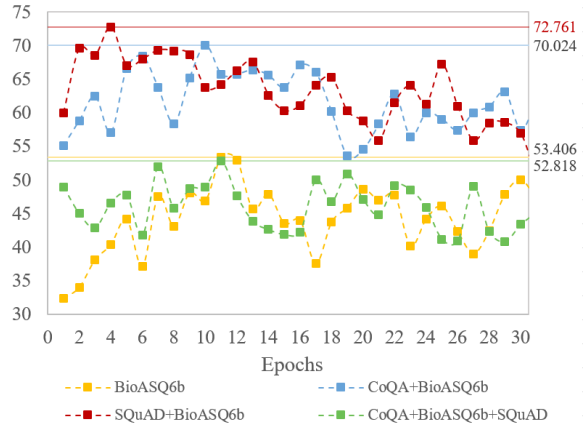
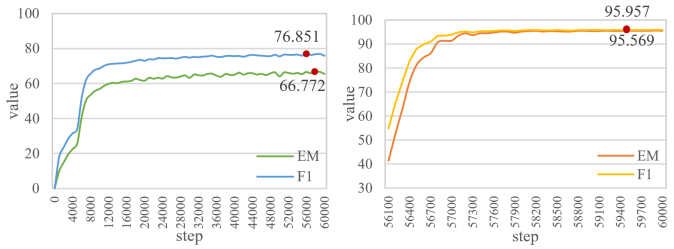


Fig. 8. The F1 value on BioASQ 6B with multiple training sets.

In our model architecture, BioBERT and QANet model are trained separately, and then the model weighting strat-

egy is adopted to get the final answer. The computational complexity mainly focuses on the training process of the above two models.

For BioBERT model, BioASQ dataset is used to fine-tune the model which has already been pre-trained on PubMed and SQuAD. As BioASQ training set has just about 1000 samples, the training process takes less than 5 minutes.



(a) Training on SQuAD dataset (b) Training on BioASQ 6B dataset

Fig. 9. The training process of QANet on the development dataset.

For QANet, it is trained on SQuAD model for 60000 steps firstly. The current model will be saved every 1000 steps, and the best one will be chosen to continue training on BioASQ dataset. We give the training process of QANet on the development dataset. Fig 9(a) and Fig 9(b) represent the training process of the QANet model on SQuAD and BioASQ 6B respectively. When it is trained by 56000 steps on SQuAD dataset, F1 value reaches the highest and so this model is selected to continue training on BioASQ.

5 Conclusion

In this paper, three effective data augmentation strategies are adopted to improve the model performance, and a dual model weighting strategy is proposed to take 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. Experimental results show that our method achieves the best performance compared to the single model on BioASQ 6B, 7B, and 8B datasets. Further, we use ADDSENT method to generate adversarial dataset of BioASQ, and give the adversarial training to the model. The results show the robustness of our method and the possibility of adversarial training to enhance model's understanding ability. 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.

Acknowledgments

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. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv:1611.01603*, 2016.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[9] Y. Seonwoo, J.-H. Kim, J.-W. Ha, and A. Oh, "Context-aware answer extraction in question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020, pp. 2418–2428.

[10] J. Deriu, K. Mlynchyk, P. Schläpfer, A. Rodrigo, D. von Grünigen, N. Kaiser, K. Stockinger, E. Agirre, and M. Cieliebak, "A methodology for creating question answering corpora using inverse data annotation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 897–911.

[11] Q. Cao, H. Trivedi, A. Balasubramanian, and N. Balasubramanian, "DeFormer: Decomposing pre-trained transformers for faster question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 4487–4497.

[12] D. Liu, Y. Gong, J. Fu, Y. Yan, J. Chen, D. Jiang, J. Lv, and N. Duan, "RikiNet: Reading wikipedia pages for natural question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6762–6771.

[13] A. Saxena, A. Tripathi, and P. Talukdar, "Improving multi-hop question answering over knowledge graphs using knowledge base embeddings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 4498–4507.

[14] S. Kundu, Q. Lin, and H. T. Ng, "Learning to identify follow-up questions in conversational question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 959–968.

[15] J. Chen, J. Zhou, Z. Shi, B. Fan, and C. Luo, "Knowledge abstraction matching for medical question answering," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, pp. 342–347.

[16] V. Kommaraju, K. Gunasekaran, K. Li, T. Bansal, A. McCallum, I. Williams, and A.-M. Istrate, "Unsupervised pre-training for biomedical question answering," in *CLEF (Working Notes)*, 2020.

[17] M. Jeong, M. Sung, G. Kim, D. Kim, W. Yoon, J. Yoo, and J. Kang, "Transferability of natural language inference to biomedical question answering," *arXiv preprint arXiv:2007.00217*, 2020.

[18] Y. Yang, N. Jin, K. Lin, M. Guo, and D. Cer, "Neural retrieval for question answering with cross-attention supervised data augmentation," in *Proceedings of the 59th Annual Meeting of the Association*

for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 2021, pp. 263–268.

[19] X. Xu, G. Wang, Y.-B. Kim, and S. Lee, "AugNLG: Few-shot natural language generation using self-trained data augmentation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, 2021, pp. 1183–1195.

[20] Y. Xu, E. Ishii, G. I. Winata, Z. Lin, A. Madotto, Z. Liu, P. Xu, and P. Fung, "CAiRE in DialDoc21: Data augmentation for information seeking dialogue system," in *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, Online, 2021, pp. 46–51.

[21] J. Lichtarge, C. Alberti, S. Kumar, N. Shazeer, N. Parmar, and S. Tong, "Corpora generation for grammatical error correction," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, Minneapolis, Minnesota, 2019, pp. 3291–3301.

[22] D. Kaushik, D. Kiela, Z. C. Lipton, and W.-t. Yih, "On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, 2021, pp. 6618–6633.

[23] J. Lin, J. Zou, and N. Ding, "Using adversarial attacks to reveal the statistical bias in machine reading comprehension models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, 2021, pp. 333–342.

[24] 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.

[25] 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.

[26] H. Xiao, "bert-as-service," https://github.com/hanxiao/bert-as-service, 2018.

[27] 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.

[28] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 2017, p. 1243–1252.

[29] 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.

[30] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 2021–2031.

Yongping Du received the Ph.D. degree in computer science from Fudan University, China, in 2005. She is currently a Professor at Beijing University of Technology. Her research interests include information retrieval, information extraction, and natural language processing.

Jingya Yan received the B.S. degree in Computer Science and Technology from Beijing University of Technology, China, in 2020. She is currently pursuing the M.S. degree in Computer Science and Technology. Her research interests include natural language processing and question answering.

Yuxuan Lu is currently pursuing the B.S. degree in Computer Science and Technology at Beijing University of Technology, China. His research interests include natural language processing and question answering.

Yiliang Zhao received the B.S. degree in Computer Science and Technology from Beijing University of Technology, China, in 2019. He is currently pursuing the M.S. degree in Computer Technology at Beijing University of Technology. His research interests include natural language processing and question answering.

Xingnan Jin received the B.S. degree in Computer Science and Technology from Beijing University of Technology, China, in 2020. She is currently pursuing the M.S. degree in Computer Science and Technology. Her research interests include natural language processing and question answering.