MSQ-BIOBERT: AMBIGUITY RESOLUTION TO ENHANCE BIOBERT MEDICAL QUESTION-ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Bidirectional Encoder Representations from Transformers (BERT) and its biomedical variation (BioBERT) achieve impressive results on the SQuAD or medical question-answering (QA) datasets, and so they are widely used for a variety of passage-based QA tasks. However, their performances rapidly deteriorate when encountering passage and context ambiguities. This issue is prevalent and unavoidable in many fields, notably the medical field. To address this issue, we introduce a novel approach called the Multiple Synonymous Questions BioBERT (MSQ-BioBERT), which integrates question augmentation, rather than the typical single question used by traditional BioBERT, to elevate performance. Experiments with both an ambiguous medical dataset and open biomedical datasets demonstrate the significant performance gains of the MSQ-BioBERT approach, showcasing a new method for addressing ambiguity in QA tasks.

1 INTRODUCTION

Question answering systems are highly useful for extracting desired information from narrative passages. Within the field of natural language processing (NLP), many tasks are related to and can be viewed as a form of QA, including information retrieval Kolomiyets & Moens (2011) and reading comprehension Cerdán et al. (2009); Wang et al. (2018). A commonly used model for QA tasks is the BERT framework Devlin et al. (2018). For general QA tasks, BERT can achieve impressive results Devlin et al. (2018); Alberti et al. (2019).

There are domain specific situations, however, where BERT has not perform as well on QA tasks. This has been observed, for example, when there is a degree of ambiguity within the contents. This aspect of BERT leads to deficiencies and poorer performances in certain areas where the content of the passage is complex but high precision is required; a notable example of this is the medical field, where the accuracy of a QA system is highly important for information retrieval, pathology studies, and research. Here, no details can be ignored, and obscure and seemingly unimportant results may very well be important. Compounding this is the fact that medical topics encompass complex language and rare medical vocabulary. Thus, a QA task on medical topics can be challenging. Wen et al. (2020) evaluated clinical QA tasks with BERT and found that improvements were needed in training data preparation in addition to the redesign of fundamental tasks. Jin et al. (2020) created a multiple-choice OpenQA dataset for solving medical problems, MEDQA, but found that even the best current method could only achieve 36.7% test accuracy. One of the significant works for improving medical-field QA task is that Lee et al. (2020) trained the BioBERT by adding medical text corpora (PubMed and PMC) and got a 12.24% MRR improvement on biomedical QA task.

In this work, we propose a method to further improve BioBERT QA model, called the Multiple Synonymous Questions BioBERT (MSQ-BioBERT), that can start from any pre-trained or fine-tuned BioBERT models Lee et al. (2020). The MSQ-BioBERT model obtains the answer scores of different questions by question augmentation using back-translation method, and then uses (1) the relevance of the augmented questions to the original question and (2) the relevance of questions to the corresponding passage to strengthen certain answers. Finally, singular value decomposition (SVD) is used to reduce the answer matrix of multiple questions to a lower-rank space to determine the final answer. We use various metrics to evaluate our model on (i) a symptom-treatment medical QA dataset created by us from Wikipedia articles. (ii) an open and public large-scale biomedical

QA dataset, BioASQ (6B, 7B, 8B and 9B) Tsatsaronis et al. (2015). We make both our code and data publicly available. [1]

The methods and approaches we present offer the following contributions and innovations:

- Our proposed MSQ-BioBERT model is lightweight and it can start from any fine-tuned BioBERT models to enhance the QA task without any additional training data.
- This work is intended to be applied to the challenging ambiguous QA datasets, such as passages with rare vocabulary, long complex sentences, complex grammar, unclear keywords, or interference information, which are very applicable to medical situations.
- We explored a new method of information integration for determining one answer from multiple synonymous questions and ultimately improved BioBERTs' performance on medical QA tasks. We exhaustively analyze the effects of the structure and parameters of our proposed MSQ-BioBERT model on its performance.

2 RELATED WORK

Many NLP tasks are based on Recurrent neural networks (RNNs) Cho et al. (2014); Sundermeyer et al. (2014). In 2017, Vaswani et al. (2017) introduced the Transformer architecture, which eliminates RNNs and instead integrates self-attention mechanisms. Then, the BERT language model Devlin et al. (2018) and its biomedical variation BioBERT Lee et al. (2020) were introduced. A standard BERT/BioBERT utilizes the Transformer, and after fine-tuning, can obtain exceptional results on many NLP tasks including QA tasks.

Many works have been proposed to enhance QA tasks. One direction in particular has been the enhancement of the robustness of the answer. Buck et al. (2017) framed QA task as a Reinforcement Learning task and proposed an agent to learn to reformulate questions to elicit answers. Dong et al. (2017) used paraphrases as a means of capturing knowledge and introduced a framework where questions and their paraphrases serve as input to a neural scoring model and are assigned different weights according to the correctness of the output answers. Yang et al. (2019) implemented data augmentation by using a combination of passages. Min et al. (2020) constructed the AmbigNQ dataset to study the task of resolving ambiguity in answers. Jeong et al. (2020) presented a sequential transfer learning method and applyed BioBERT to transfer the knowledge of natural language inference (NLI) to biomedical QA. Du et al. (2021) proposed a dual model weighting strategy, which combined the advantage of two models, QANet and BioBERT, to improve the model's generalization ability in biomedical QA. Du et al. (2021) also used Round-trip translation method (in our work it is named "back-translation") but the way they use back-translation differs from ours in the following ways: (1) they only choose French as the bridge language and apply back-translation on passages, whereas we choose 56 bridge languages and apply back-translation only on questions, (2) they create more training samples, whereas we do not interfere with training or fine-tuning, but intend to obtain multiple answer scores for each original question-passage pair.

In summary, while the above enhancement methods have been shown to enhance answer fidelity, they do so by increasing the training dataset, introducing new knowledge, or combining models, which is not possible in all applications. In our work, the goal is to enhance the BioBERT without any additional training data and training cost.

Finally, for the question augmentation, there are several other options to achieve it. For example, RNN-based methods Bowman et al. (2015), three levels of augmentation via word embeddings Ma (2019), and easy data augmentations (EDAs) Wei & Zou (2019). However, for our work, the augmented questions should be guaranteed to be very accurate and free of grammatical errors due to our objective of implementing QA on medical application. That is why we adopt the Back Translation method Edunov et al. (2018; 2019).

[1] `https://anonymous.4open.science/r/MSQ_biobert-E91C/README.md`

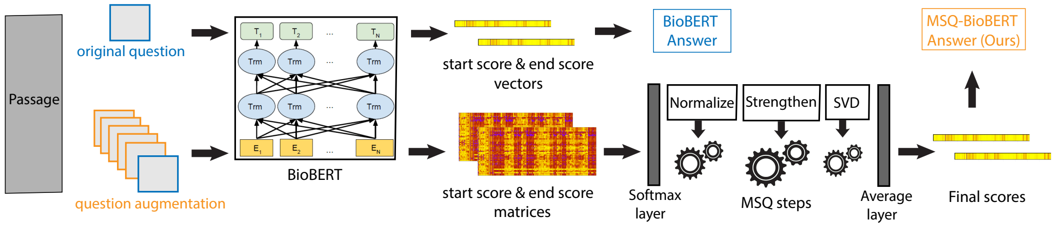

Figure 1: Overview of BioBERT QA and our proposed MSQ-BioBERT QA models. The upper part represents the BioBERT process of generating answers. The bottom part shows our approach.

3 Datasets

3.1 ST-QuAAD

We have constructed a new validation dataset called Symptom-Treatment QA Ambiguous Dataset (ST-QuAAD), in which we selected 37 common symptoms of COVID-19 such as fever, cough, sore throat, etc. Note that the selection of COVID-19 was made simply because of its current prevalence. Then, a web crawler code to obtain corresponding Treatment and Management paragraphs from Wikipedia was written. Finally, paragraphs were separated into 343 reading passages to ensure that each reading passage has only one continuous span answer, i.e., treatment method, of the corresponding symptom. The dataset is objective, yet ambiguous as some of the passages may not contain symptom keywords since they are generated by separating the original paragraphs. In addition, the passages may contain answers with complicated medical words. Furthermore, there are in fact some wrong treatments as the answers to interference in the passages. In summary, ST-QuAAD is coarser and ambiguous and has the same format as Stanford Question Answering Dataset (SQuAD) Rajpurkar et al. (2016). Because our goal is to obtain the treatment methods of symptoms, all question formats take the following form: "what possible therapeutical method is helpful to treat <symptom>?", where <symptom> will be replaced by specific symptoms.

3.2 BioASQ

We also chose the BioASQ factoid test datasets (6B, 7B, 8B and 9B) Tsatsaronis et al. (2015) because their format is similar to that of SQuAD. In these datasets, the passages come from a large numbers of biomedical scientific articles, the questions and the "exact answers" are prepared by biomedical experts. We excluded samples with unanswerable questions from the datasets because their answers did not appear in the given passages. Due to its biomedical relevance (rare medical vocabulary, long complex sentences, etc.), the QA task on BioASQ is also more challenging compared to that on SQuAD.

4 Methods

As shown in Figure1, our proposed MSQ-BioBERT will start with question augmentation so that each passage corresponds to multiple synonymous questions. For each original question-passage pair, MSQ-BioBERT will obtain a Start Score matrix and an End Score matrix instead of score vectors. After a softmax layer, there are three MSQ steps: (1) normalization, (2) strengthening layer that makes the scores of certain question-passage pairs stand out, and (3) SVD approximation, followed by an averaging layer. The final start/end scores obtained from the MSQ method are in the same format as the start/end scores of the single-question QA task. Therefore, MSQ-BioBERT uses the same criteria as BioBERT to find text answers from the final start/end scores.

4.1 Question Augmentation

The Back Translation Edunov et al. (2018; 2019) method was adopted to augment test questions. This approach ensures that the structures of questions are diversified, and that the question grammar and word spellings are correct. We chose 56 commonly used languages including French, Chinese, Hindi, Spanish, etc., as bridge languages. English questions were first translated into these

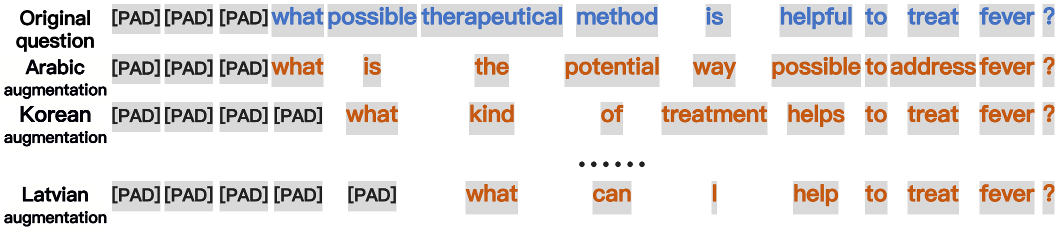
Original question [PAD] [PAD] [PAD] what possible therapeutical method is helpful to treat fever ?
Arabic augmentation [PAD] [PAD] [PAD] what is the potential way possible to address fever ?
Korean augmentation [PAD] [PAD] [PAD] [PAD] what kind of treatment helps to treat fever ?
......
Latvian augmentation [PAD] [PAD] [PAD] [PAD] [PAD] what can I help to treat fever ?

Figure 2: A given original question (q_0) is in blue. The synonymous questions generated by back translation ($q_1, ..., q_n$) are in orange. [PAD] tokens are added before questions to align positions of multiple scores.

languages via Google Translation API, then, the questions were translated back to English and repeated questions were removed. Figure 2 exemplifies this approach and how question augmentation was achieved.

To specify the input and output of our model, we use a single passage denoted by c as an example. Suppose this passage has $n+1$ non-repeating questions: $q_0, q_1, \ldots, q_n$, where q_0 is the original question. The maximum length of each question-passage pair is limited to 384 tokens. The model input for this example is:

$$X_c = \begin{pmatrix} q_0 & q_1 & \ldots & q_n \\ c & c & \ldots & c \end{pmatrix}^T, \tag{1}$$

where the i-th row, (q_i^T, c^T), $i = 0, 1, ..., n$, is a token vector of length 384 representing the pair of the i-th question q_i and the passage c. In our work, all the vectors represented by letters are by default column vectors and T is the symbol for transpose.

4.2 ORIGINAL SCORES AND ITS NORMALIZATION

BioBERT model for QA task will search answers from Start Scores (S_{start}) and End Scores (S_{end}), which are the model's original output vectors. In our proposed MSQ-BioBERT model, both S_{start} and S_{end} are in matrix form. Using S_{start}, the following example illustrates our method and approach; note that the same process is utilized for S_{end} as well. For the input X_c, S_{start} generated by MSQ-BioBERT is given by:

$$S_{start} = \text{MSQ-BioBERT}(X_c)_{start} = (s_0, \quad s_1, \quad \ldots, \quad s_n)^T, \tag{2}$$

where s_i $(i = 0, 1, ..., n)$, a vector of length 384, is the Start Score vector of the (q_i^T, c^T) question-passage pair. An example of S_{start} is shown in Figure 3. After getting S_{start}, we apply softmax function $\sigma(.)$ to transform values of each s_i $(i = 0, 1, ..., n)$ into a probability distribution and then perform min-max normalization for each s_i to obtain the Normalized Start Scores (S_{Nstart}), which have range from 0 to 1. S_{Nstart} is shown in Equation (4).

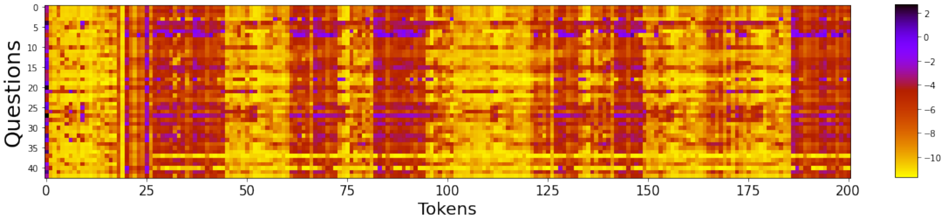

Figure 3: An example of Start Scores (S_{start}) for one passage and n+1 questions. The rows correspond to questions, and the columns correspond to question-passage input tokens (we only show the first 200 tokens); The values of the matrix are scores.

$$s_i' = \frac{\sigma(s_i^T) - min(\sigma(s_i^T))}{max(\sigma(s_i^T)) - min(\sigma(s_i^T))}, \quad i = 0, 1, ..., n. \tag{3}$$

$$S_{Nstart} = (s_0' \quad s_1' \quad \ldots \quad s_n')^T. \tag{4}$$

4.3 Strengthening: Make Questions and Tokens Stand Out

S_{Nstart} includes scores obtained with $n+1$ non-repeating questions, and different questions have different relevance to the passage. For example, the answers obtained from the questions with "reducing symptoms" and "helping to treat symptoms" may vary depending on the focus of the verb. Or, the answers obtained from the questions with "therapeutic method" and "what should I do" may also vary depending on the focus of the subject. An intuitive idea here is to use the number of words shared by the passage and the question to reflect their relevance. Therefore, we define the word-frequency score (WFS) f_i as the number of shared words by the $i\text{-}th$ question and the passage divided by the length of question. Note that the stopwords Bird et al. (2009) that do not provide any useful information will not be counted. In order for the algorithm to always have none-zero WFS, we initialize f_0 to 0.1 for the original question and f_i $(i = 1, 2, ..., n)$ to 0.01 for the augmented questions because the original question is probably the most accurate one. The initial WFS is the score used in case there is no word-overlap at all. We denote the initial word-frequency score for augmented questions as IWFS, (IWFS=0.01 in our model), and we discuss IWFS in Section 5.5. The word-frequency score vector for the single passage with $n+1$ questions is written as:

$$F = \begin{pmatrix} f_0 & f_1 & \dots & f_n \end{pmatrix}^T, \tag{5}$$

On the other hand, we encourage diversity of questions and avoid the dominance of some very similar questions. For each question q_i, we calculate its Levenshtein Similarity (LS) Levenshtein et al. (1966); Navarro (2001) to the original question q_0 (Equation 6):

$$Lev(q_0, q_i) = \begin{cases} |q_0|, & if\ |q_i| = 0 \\ |q_i|, & if\ |q_0| = 0 \\ Lev(t(q_0), t(q_i)), & if\ q_0^1 = q_i^1 \\ 1 + min \begin{Bmatrix} Lev(t(q_0), q_i), \\ Lev(q_0, t(q_i)), \\ Lev(t(q_0), t(q_i)) \end{Bmatrix}, otherwise \end{cases} \quad , \quad i = 0, 1, 2, ..., n \tag{6}$$

where $t(x)$ denotes the string of all but the first character of x, and x^1 denotes the first character of the string x. Then, the Levenshtein Distance Ratio (LDR) between q_0 and q_i is defined as:

$$r_{Lev}(q_0, q_i) = \frac{Lev(q_0, q_i)}{Alignment\ length\ of(q_0, q_i)} \tag{7}$$

if $r_{Lev}(q_0, q_i)$ is greater than a Levenshtein similarity threshold (LST), we exclude the question q_i. In our model, LST = 0.85 and we discuss the LST effect in section 5.4 (note that the original question will always be kept even though $Lev(q_0, q_0) = 1$). We define the LDR score vector for $n+1$ questions as:

$$L = \begin{pmatrix} l_0 & l_1 & \dots & l_n \end{pmatrix}^T, \tag{8}$$

where

$$l_i = \begin{cases} 1, & if\ r_{Lev}(q_0, q_i) < LST\ or\ i = 0 \\ 0, & if\ r_{Lev}(q_0, q_i) \geq LST\ and\ i \neq 0 \end{cases} \quad , \quad i = 0, 1, 2, ..., n \tag{9}$$

Then, we use element-wise multiplication of S_{Nstart} activated by sigmoid function ϕ, WFS, and LDR scores to highlight the important questions. Their products are smoothed again by ϕ. This idea gives us the Highlighted Start Scores (S_{Hstart}) as follows:

$$S_{Hstart} = \phi\left(\phi(S_{Nstart}) \circ \begin{pmatrix} F & F & \dots & F \end{pmatrix}_{(n+1)\times 384} \circ \begin{pmatrix} L & L & \dots & L \end{pmatrix}_{(n+1)\times 384}\right), \tag{10}$$

where $\circ$ is the Hadamard product. A role of S_{Hstart} is to make both important tokens and important questions stand out (See Figure 4).

4.4 SVD approximation: Reduce Score Matrix Rank

S_{Hstart} highlights important tokens of the important questions. Now, we use SVD to reduce S_{Hstart} matrix to a lower-rank matrix (with rank one) in order to obtain a single characteristic of S_{Hstart}. Suppose the rank of the S_{Hstart} matrix is R, and S_{Hstart} has such a singular value decomposition:

$$S_{Hstart} = U\Sigma V^T = \Sigma_{k=1}^{R} \lambda_k u_k v_k^T, \tag{11}$$

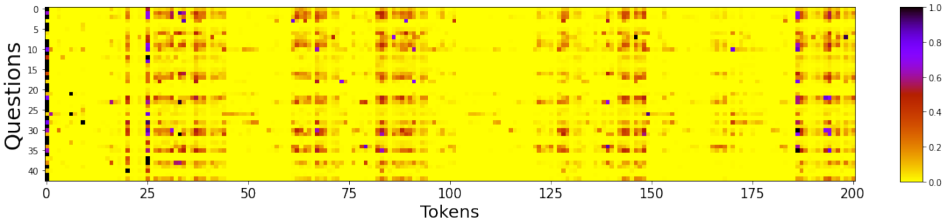

Figure 4: An example of Highlighted Start Scores (S_{Hstart}).

where λ_k is the k-th largest eigenvalue of S_{Hstart}, u_k and v_k are corresponding left singular vector and right singular vector. Then, the Truncated SVD Start Scores (S_{Tstart}) can be obtained by applying truncated SVD on S_{Hstart} matrix:

$$S_{Tstart} = \Sigma_{k=1}^{r} \lambda_k u_k v_k^T . \tag{12}$$

In our model, $r = 1$. We discuss the effect of rank r in Section 5.6. An example of S_{Tstart} can be seen in Figure 5.

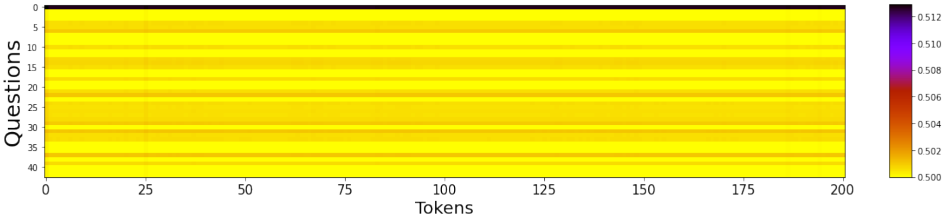

Figure 5: An example of Truncated SVD Start Scores (S_{Tstart}).

4.5 Final scores and answers

At last, the Final Start Score (S_{Fstart}) (shown in Figure 6) is the mean value of each column (token) of S_{Tstart} matrix:

$$S_{Fstart} = \begin{pmatrix} \overline{S}_{Tstart,:,1} & \overline{S}_{Tstart,:,2} & \dots & \overline{S}_{Tstart,:,384} \end{pmatrix} . \tag{13}$$

As mentioned earlier, MSQ approach also yields Final End Scores (S_{Fend}) by the same process. Our MSQ-BioBERT model will generate answers from S_{Fstart} and S_{Fend}. Because S_{Fstart} and S_{Fend} are in the same format as the start scores and end scores of the general single-question QA task, we adopt the default settings of the QA task of the BERT family to generate answer: the maximum length of an answer that can be generated is 30 and the total number of n-best predictions to generate when looking for an answer is 20. The settings are consistent for all models in this paper.

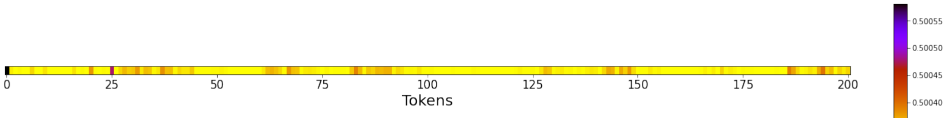

Figure 6: An example of Final Start Score (S_{Fstart}).

5 Experiments and Results

5.1 Example and Metrics

Firstly, we demonstrate actual differences in specific answer results to several distinct questions between BioBERT and our MSQ-BioBERT, using a concrete example (Table 1). Although there are no significant contextual differences among these questions from the view of a human, the BioBERT gives diverse potential answers, some of which are completely incorrect. Nevertheless, MSQ-BioBERT uses this variety of questions to determine a correct answer.

Table 1: An example of answers of BioBERT and MSQ-BioBERT on BioASQ factoid dataset; the ground truth answer is written in blue while incorrect answers are noted in orange. The answers in the gray boxes are the actual model outputs. The transparent answers are NOT original BioBERT outputs, but potential answers, because BioBERT does not have question augmentation.

Passage		
the recount2 resource is composed of over 70 000 uniformly processed human rna seq samples spanning tcga and sra including gtex the processed data can be accessed via the recount2 website and the bioconductor package this workflow explains in detail how to use the package and how to integrate it with other ...		
Question Examples **(original q_0 and augmented $q_1 \sim q_4$)**	**BioBERT answers and potential answers**	**MSQ-BioBERT Final Answer**
q_0 : which workflow in bioconductor has been developed for accessing human rna-seq samples?	the bioconductor package	the recount2
q_1 : what workflow has been developed in bioconductor to access human rna segments?	the recount2	
q_2 : what functional flow in bioconductor has been developed for obtaining human rna-seq samples?	coverage count matrices were computed in recount2	
q_3 : which workflow was created in the bioconductor to access human rna-seq samples?	the bioconductor package	
q_4 : what bioconductor workflow has been developed to obtain human rna-seq samples?	the recount2	

Our work is aimed at the enhancement of BioBERT QA task, so BioBERT QA large model Lee et al. (2020) is viewed as the baseline. In addition to the F1 score and exact match score (EM) used by the original BERT authors, we add three more metrics to evaluate models: Word match score (WM) (see Equation 14), String match score (SM) (see Equation 15), and Levenshtein similarity (LS) Levenshtein et al. (1966) of predicted answers and groundtruth answers.

$$\text{Word match score} = \frac{\text{number of matching words}}{\text{number of groundtruth answer words}} \tag{14}$$

$$\text{String match score} = \frac{\text{length of matching string}}{\text{length of groundtruth answer string}} \tag{15}$$

5.2 MSQ-BioBERT PERFORMANCE ON DATASETS

In addition to BioBERT and our proposed MSQ-BioBERT, we add the BERT model to validate the medical properties of our ST-QuAAD dataset in Table 2. First, we observe that BioBERT outperforms BERT, which is consistent with our expectation since the dataset we constructed is medically relevant. Secondly, it is worth mentioning that after applying our method on BioBERT, the improvement on all the metrics is obvious. Specifically, F1 score increases by 4.41% absolute points and EM score increases by 4.37% absolute points. WM score increases by 5.71% absolute points, which indicates that our method can improve the word overlap between predicted and groundtruth answers. If the answers are viewed as strings, our method also yields answers that are more similar to the groundtruth answers, as the SM and LS scores increase by 4.53% and 4.16%, respectively.

Table 2: Performance of BERT/BioBERT and our MSQ-BioBERT models on ST-QuAAD dataset

Model	F1 score	EM	WM	SM	LS
BERT	70.24 %	53.94%	79.26%	70.19%	75.59%
BioBERT	72.34%	56.56%	77.85%	71.48%	76.23%
MSQ-BioBERT (ours)	**76.75%**	**60.93 %**	**83.56%**	**76.01%**	**80.39%**

Since BioASQ datasets are open biomedical QA datasets, in Table 3, we focus on the enhancements of our approach relative to BioBERT. There are four BioASQ factoid test datasets released from 2018 to date. The passages in BioASQ are derived from snippets of biomedical scientific articles, but rare medical words, long complex sentences, and the lack of keywords in some passages lead to some ambiguity and challenges. We can see that our method can enhance BioBERT comprehensively. F1, EM, WM, SM, and LS scores are improved by 0.48%∼1.72%, 0.51%∼2.61%, 1.74%∼3.17%, 2.07%∼4.22%, and 0.50%∼1.92%, respectively on the four BioASQ datasets. This demonstrates the robustness and applicability of our method.

Table 3: Performance of BioBERT and our MSQ-BioBERT on BioASQ factoid test datasets

Model	Data set	F1 score	EM	WM	SM	LS
BioBERT	BioASQ6	75.72%	59.35%	84.65%	83.89%	78.84%
MSQ-BioBERT (ours)	BioASQ6	**77.31%**	**61.51%**	**86.39%**	**86.06%**	**80.44%**
BioBERT	BioASQ7	80.56%	58.79%	87.89%	91.43%	81.77%
MSQ-BioBERT (ours)	BioASQ7	**82.28%**	**59.30%**	**90.58%**	**94.08%**	**83.69%**
BioBERT	BioASQ8	74.46%	65.52%	79.88%	91.09%	83.58%
MSQ-BioBERT (ours)	BioASQ8	**74.94%**	**66.21%**	**83.05%**	**95.31%**	**85.33%**
BioBERT	BioASQ9	76.87%	56.74%	87.43%	90.89%	81.30%
MSQ-BioBERT (ours)	BioASQ9	**77.81%**	**58.02%**	**89.98%**	**92.96%**	**81.80%**

In the following Section 5.3 to Section 5.6, we analyze the structure and parameters of our MSQ-BioBERT model. We combine the four BioASQ test datasets (6B, 7B, 8B, and 9B) and always use the combined dataset for analysis below.

5.3 Effect of MSQ structure

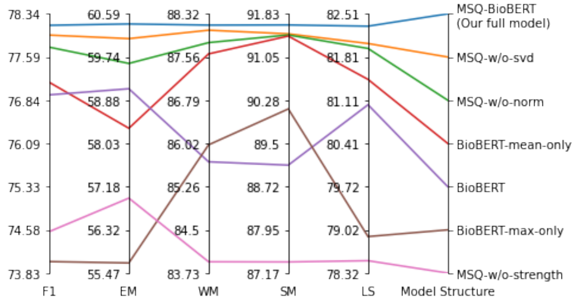

Figure 7: Effectiveness of MSQ-BioBERT structure on BioASQ combined datasets

We evaluate the effectiveness of our proposed MSQ-BioBERT structure. There are three main steps in our proposed MSQ approach: Normalization (Section 4.2), Strengthening (Section 4.3), and SVD (Section 4.4). We eliminate the three steps one by one to obtain partial MSQ-BioBERT models, denoted by MSQ-w/o-norm, MSQ-w/o-strength, and MSQ-w/o-svd, respectively. In addition to BioBERT, we use token-wise mean scores (BioBERT-mean-only) and token-wise maximum scores (BioBERT-max-only) as two other baselines. As shown in Figure7, MSQ-BioBERT performs the best among all models, followed by MSQ-w/o-svd and MSQ-w/o-norm. BioBERT-mean-only is very close to MSQ-BioBERT on SM metric, but far from MSQ-BioBERT on other metrics. We also note that MSQ-w/o-strength performs poorly, which illustrates that the strengthening step is critical in our MSQ-BioBERT model.

5.4 Effect of Levenshtein similarity threshold

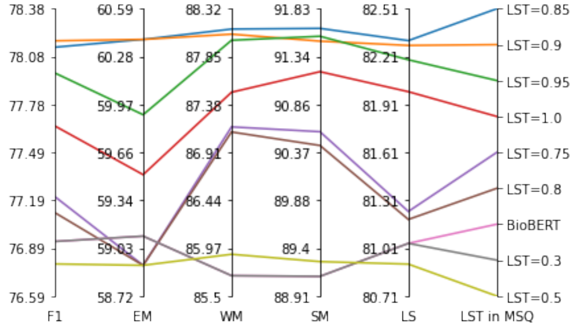

Figure 8: Effectiveness of LST in MSQ-BioBERT

We evaluate the effectiveness of LST in MSQ-BioBERT. Intuitively, LST controls: (1) the number of augmented questions; (2) the degree of variation of the augmented questions from the original question. The larger the LST, the more augmented questions will be selected, but some of them will be more similar to the original question. When LST = 1, all the augmented questions will be selected; when LST = 0, MSQ-BioBERT will degenerate to BioBERT. Our model uses LST = 0.85 by default, which is similar to the effect of LST = 0.9 as seen in the figure 8. We notice that when LST decreases to 0.5, the MSQ-BioBERT performance becomes significantly worse; when LST decreases to 0.3, MSQ-BioBERT has degenerated to BioBERT. This is because none of the augmented questions has a Levenshtein similarity to the original question of less than 0.3.

5.5 Effect of initial word-frequency scores

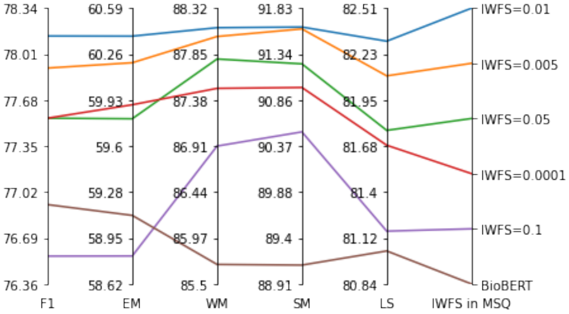

Figure 9: Effectiveness of IWFS in MSQ-BioBERT

We evaluate the effectiveness of IWFS in MSQ-BioBERT. The IWFS reflects our initial confidence in the augmented questions. Since we fix such an initial score of the original question to 0.1, if IWFS = 0.1, the augmented questions will have the same initial weight as the original question; the smaller the IWFS, the lighter the initial importance of the augmented questions. Although some IWFS may be trivial due to the addition of WFS, for some augmented questions that do not share words with the passage (i.e. WFS=0), IWFS will play a critical role. From Figure 9 we can see that when IWFS = 0.1, MSQ-BioBERT has modest performance (lower F1, EM; higher WM, SM, LS compared to BioBERT). However, MSQ-BioBERT will fully outperform BioBERT as long as IWFS decreases by one order of magnitude, for example, IWFS = 0.01, which suggests that the original question should indeed have more initial weight than augmented questions.

5.6 Effect of truncated SVD rank

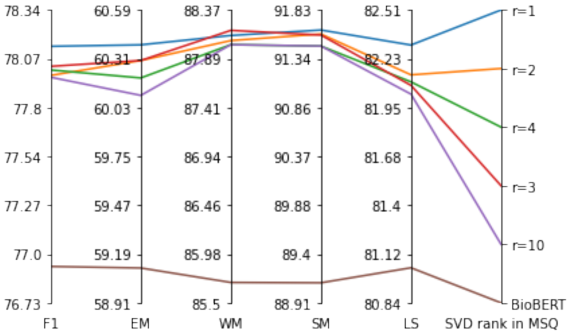

Figure 10: Effectiveness of truncated SVD rank in MSQ-BioBERT

We evaluate the effectiveness of truncated SVD rank r in MSQ-BioBERT. The truncated SVD is no longer an exact decomposition of the score matrix (S_{Hstart} in our work), but rather provides the optimal low-rank score matrix approximation (S_{Tstart} in our work). r characterizes the number of "common feature" of a bunch of question-passage scores. In the proposed MSQ-BioBERT, r is defaulted to 1 because we think we only need to focus on the most important one common feature of the score matrix S_{Hstart}. As a matter of fact, MSQ-BioBERT performs best with the $r = 1$ setting (highest F1, EM, SM, and LS; second highest WM), as shown in Figure 10. Overall, the change in rank r has a limited impact on MSQ-BioBERT. When r is large enough, for instance $r = 10$, the MSQ-BioBERT will become MSQ-w/o-svd model described in Section 5.3, because in this case the estimation error is very small and the approximation matrix S_{Tstart} is almost identical to S_{Hstart}.

6 Summary and Future Work

In this paper, we demonstrate a new method, the MSQ-BioBERT method, that integrates question augmentation to enhance the performance of BioBERT on biomedical QA task. Experiments demonstrate that MSQ method can enhance BioBERT QA models on almost all metrics. In future work, we intend to enhance our model with richer question augmentation and once again examine its robustness for BioBERT QA task that possess passages with ambiguous and challenging contents. It is our hope that work in this field continues to drive enhanced performance in passage-based QA tasks that demand precise and accurate answers such as the field of medicine.

REFERENCES

Chris Alberti, Kenton Lee, and Michael Collins. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*, 2019.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. Ask the right questions: Active question reformulation with reinforcement learning. *arXiv preprint arXiv:1705.07830*, 2017.

Raquel Cerdán, Eduardo Vidal-Abarca, Tomas Martinez, Ramiro Gilabert, and Laura Gil. Impact of question-answering tasks on search processes and reading comprehension. *Learning and Instruction*, 19(1):13–27, 2009.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. Learning to paraphrase for question answering. *arXiv preprint arXiv:1708.06022*, 2017.

Yongping Du, Jingya Yan, Yiliang Zhao, Yuxuan Lu, and Xingnan Jin. Dual model weighting strategy and data augmentation in biomedical question answering. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 659–662. IEEE, 2021.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. On the evaluation of machine translation systems trained with back-translation. *arXiv preprint arXiv:1908.05204*, 2019.

Minbyul Jeong, Mujeen Sung, Gangwoo Kim, Donghyeon Kim, Wonjin Yoon, Jaehyo Yoo, and Jaewoo Kang. Transferability of natural language inference to biomedical question answering. *arXiv preprint arXiv:2007.00217*, 2020.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Oleksandr Kolomiyets and Marie-Francine Moens. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434, 2011.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pp. 707–710. Soviet Union, 1966.

Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020.

Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. Translation modeling with bidirectional recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 14–25, 2014.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Wei Wang, Ming Yan, and Chen Wu. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. *arXiv preprint arXiv:1811.11934*, 2018.

Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pp. 6383–6389, Hong Kong, China, November 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D19-1670`.

Andrew Wen, Mohamed Y Elwazir, Sungrim Moon, and Jungwei Fan. Adapting and evaluating a deep learning language model for clinical why-question answering. *JAMIA open*, 3(1):16–20, 2020.

Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. Data augmentation for bert fine-tuning in open-domain question answering. *arXiv preprint arXiv:1904.06652*, 2019.