NCU-IISR/AS-GIS: Results of Various Pre-trained Biomedical Language Models and Linear Regression Model in BioASQ Task 9b Phase B

Yu Zhang [1], Jen-Chieh Han[1] and Richard Tzong-Han Tsai[123]*

[1] *Department of Computer Science and Information Engineering, National Central University, Taiwan*
[2] *IoX Center, National Taiwan University, Taiwan*
[3] *Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan*

Abstract

Transformer has been widely applied in Natural Language Processing (NLP) field, and it also results in an amount of pre-trained language models like BioBERT, SciBERT, NCBI_Bluebert, and PubMedBERT. In this paper, we introduce our system for the BioASQ Task 9b Phase B. We employed various pre-trained biomedical language models, including BioBERT, BioBERT-MNLI, and PubMedBERT, to generate "exact" answers for the questions, and a linear regression model with our sentence embedding to construct the top-n sentences as a prediction for "ideal" answers.

Keywords

Biomedical Question Answering, Pre-trained Language Model, Linear Regression

1. Introduction

Given the rapid growth of people's interest in Artificial Intelligence (AI), and biomedical question-answering has been receiving attention [1-3]. Is AI able to answer a biomedical question, like "Does metformin interfere thyroxine absorption?", correctly? Is AI able to give textual evidence for its answer? To facilitate answering these questions, we participated in BioASQ Task 9b Phase B (QA task), where participants should return either an exact answer or an ideal answer based on the given biomedical question and List of question-relevant articles/snippets. BioASQ Task 9b PhaseB task provided 3743 training questions, including the previous year's test set with gold annotations, plus 500 test questions for evaluation, divided into five batches of 100 questions each. All questions and answers were constructed by a team of biomedical experts from across Europe and were classified into four types: Yes/no, Factoid, List, and Summary. Three types of questions required accurate answers: Yes/no, Factoid and List. For all four types of questions, participants were asked to submit the ideal answers. In Task 9b, each participant was allowed to submit up to five results per batch.

Figure 1 illustrates four examples of QA types for BioASQ Task 9b Phase B (QA task). As shown in **Figure 1**, the BioASQ QA example gives a question and several relevant PubMed abstract fragments as relevant snippets. Therefore, we formulated the task as a query-based multi-document a. extraction for the exact answer and b. summarization for the ideal answer. Last year, we used the BioBERT model combined with logistic regression to achieve the best result in generating ideal answers at batch 5 [4].

In this paper, we employed pre-trained language models to improve our results, including BioBERT [5], BioBERT-MNLI [6], and PubMedBERT [7]. BioBERT-MNLI is a fine-tuned model of BioBERT on the MultiNLI (The Multi-Genre Natural Language Inference) corpus, which is a dataset created for

[1] CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania.
EMAIL: phoenix000.taipei@gmail.com(A.1); joyhan@cc.ncu.edu.tw(A.2); thtsai@csie.ncu.edu.tw(A.3)
* Corresponding author

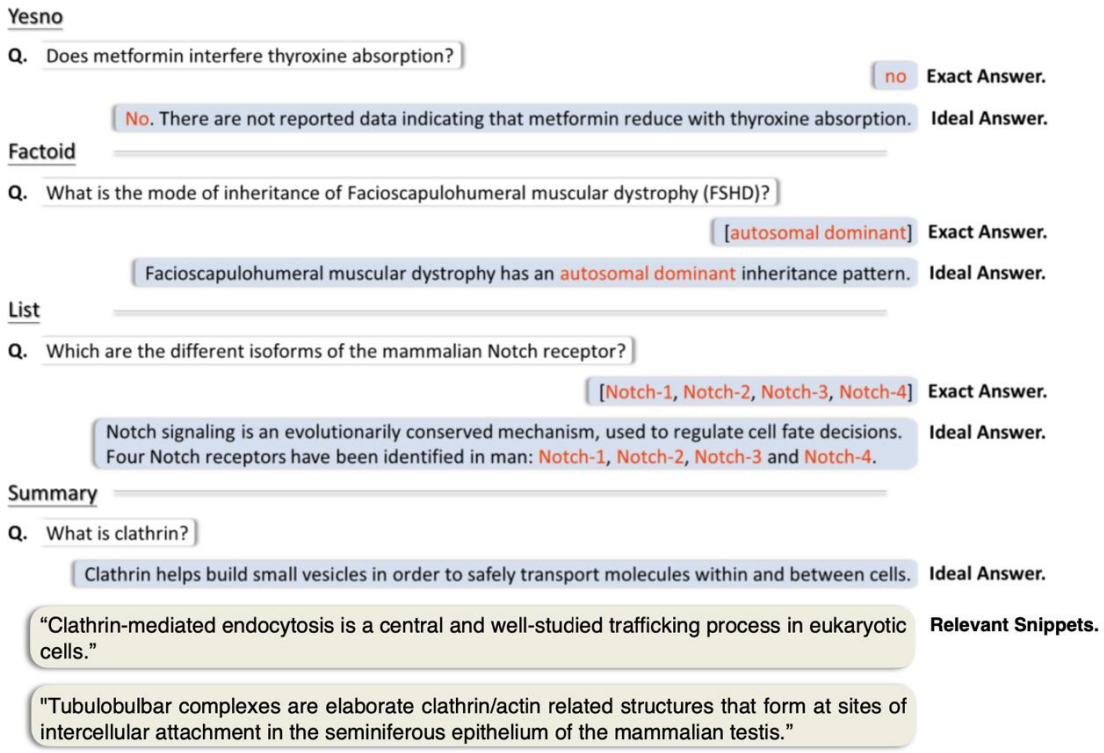

Figure 1. The QA examples of the BioASQ Task 9b Phase B (QA task).

sentence understanding task [8]. BioBERT related models achieved the best performance in extracting exact answers last year [6]. PubMedBERT is the latest BERT model pre-trained on the biomedical corpus, which outperformed BioBERT on the BLURB (Biomedical Language Understanding and Reasoning Benchmark).

We applied Pre-trained models' [CLS] embeddings as input to a linear regression model for predicting ideal answers. The sections are organized as follows. Section 2 briefly reviews recent works on biomedical QA and pretrained model. The details of our methods are described separately in Section 3. Section 4 describes our configurations submitted to the BioASQ 9b Phase B challenge and the results. Section 5 is the discussion and summary of our system's performance in the BioASQ QA task.

2. Related Work

The acquisition of biomedical knowledge is often carried out by reading academic papers. This process is time-consuming and labor-intensive, and has a high professional threshold. Biomedical professionals cannot quickly obtain the required knowledge in a short period of time. The general public is also unable to complete the acquisition of biomedical knowledge in the absence of expert assistance. QA in natural language processing tasks has the potential to solve these problems by providing direct answers to users' questions. This tests the ability of machine learning systems to semantically understand, retrieve, and generate answers from existing text. Many QA models based on deep learning have been developed and even applied in the past [9].

Biomedical QA Task: Biomedical QA tasks require a large amount of annotated corpus to train the model. This is a prerequisite for deep learning. In addition to BioASQ, many QA datasets annotated by biomedical experts have been published recently [1, 2]. The PubMedQA dataset is a research question set, and each question has a reference text from a PubMed abstract and the span of the text providing the answer (yes/maybe/no) to the research question. BioBERT generally outperformed other deep learning methods such as BiLSTM and ESIM on the PubMedQA dataset [1]. Another biomedical QA task that deserves our attention is the COVID-QA. COVID-QA is a SQuAD-like Question Answering dataset consisting of 2,019 question/answer pairs annotated by volunteer biomedical experts on

scientific articles related to COVID-19. This dataset differs from traditional MRC datasets such as SQuAD in that the answers to the questions come from a much longer context [2].

PubMedBERT: Following the successful application of BERT to natural language processing tasks in various fields, more and more specialized pre-trained language models are being developed in the biomedical field, including BioBERT, SciBERT [10], ClinicalBERT [11], BlueBERT [12], PubMedBERT, and so on. Among them, PubMedBERT is the state-of-art model developed by Microsoft. Its pre-training method is different from the existing biomedical language models, PubMedBERT adopts the method of training professional texts (PubMed papers) from scratch, instead of continuing training on the basis of texts in the general domain [7]. It has outperformed BioBERT in many biomedical NER, QA, and Relation Extraction tasks.

Sequential Learning with BioBERT: The pre-trained language model effectively improves the performance of target tasks, while sequential transfer learning based on transfer learning can be used to further improve the performance of biomedical question answering. In the general QA domain, learning relationships between sentence pairs first is effective in sequential transfer learning [13]. BioBERT's research team has also found that this approach can be applied to biomedical QA. They demonstrated that fine-tuning on the language comprehension dataset and the question-answer dataset can improve BioBERT's performance on BioASQ tasks and released new fine-tuned models such as BioBERT-MNLI, BioBERT-MNLI-SQuAD [6].

3. Method

For ideal answer, We basically used the similar method that we used to participate in the BioASQ 8B last year and tried to test this method on different pre-trained language models to boost performance. The goal of our method is to select the most relevant segments for each question in the BioASQ QA instance, and our work was inspired by the logistic regression model framework proposed by Diego Mollá [14]. The approach follows the two steps of his summarization process. In the first step, the input text is segmented into candidate sentences and each candidate sentence is scored. In the 2nd step, the top n sentences with the highest scores are returned. We use a pre-trained language model and replace its features with word embeddings.

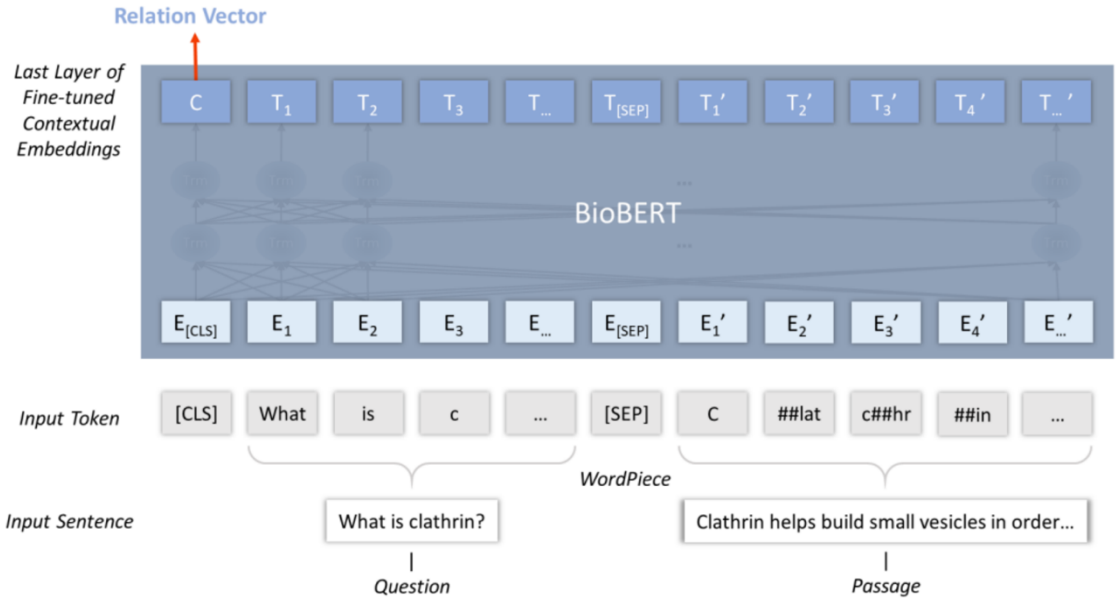

Figure 2. How a candidate answer (sentence) and the corresponding question obtains the contextual embeddings in the last layer of the BERT model (BioBERT, PubMedBERT etc.).

The training steps are as follows:

1. For the snippets and ideal answer from training set released from BioASQ organizers, we used NLTK's sentence tokenizer to divide snippets into sentences.

2. We calculated ROUGE-SU4 F1 scores [15] between each sentence and the associated ideal answer.
3. All the sentences from snippets with different generated scores were considered as candidate answers. The candidate answers, their corresponding questions and scores became the training set for our linear regression model.
4. We input a candidate answer (sentence) and the corresponding question at the same time, using score as the prediction target. The pre-trained BERT language model is used for fitting the task. We used [CLS] embeddings representing the relation between a candidate sentence and a question as the feature and appended a dense layer with ReLU activation after the output layer of BERT model. Mean squared error was used as the loss function. Our script is modified from Google BERT's official TensorFlow code and took default settings from BERT trained on SQuAD [16].

For inference, we used the fine-tuned model from step 3 to predict the scores of the test data and then re-rank the candidate sentences for each question. Because the ideal answer in training set mostly consist of only one sentence, we selected only the top 1 sentence as our system output (ideal answer).

The improvement of the above method is mainly focused on the replacement of the pretrained language models. We used BioBERT-MNLI (NCU-IISR/AS-GIS-2) and PubMedBERT (NCU-IISR/AS-GIS-3) to replace BioBERT (NCU-IISR/AS-GIS-1) in the above method in an attempt to improve the performances. BioBERT-MNLI is a fine-tuned BioBERT model on the MultiNLI dataset. MultiNLI (Multi-Type Natural Language Inference), published by New York University, is a text entailment task that requires determining whether a hypothesis holds given a premise (Premise), or determining whether the hypothesis is contradictory and neutral to the premise. MNLI's main feature is that it is a collection of texts in many different domains. We believe that the task of MultiNLI dataset has a high similarity to the ideal answer selection task. On the one hand, the questions and ideal answers in the BioASQ 9b training set are usually one sentence, and the premises and assumptions in the MultiNLI data set are also one sentence. Therefore, the data lengths are basically the same. On the other hand, the questions and the ideal answers need to maintain a logical entailment relationship. We can analogize the question to the premise and the ideal answer to the hypothesis. Only answers that are logically related should be considered.

PubMedBERT is similar to BioBERT in that both are trained using the PubMed corpus. However, BioBERT adopts a continuous pre-training approach based on BERT. So it uses vocabulary from Wikipedia and the book corpus. PubMedBERT, on the other hand, is pre-trained from scratch using the PubMed text. This means that PubMed is less influenced by general domain texts and focuses on the biomedical research corpus. In addition, to test the applicability of PubMedBERT in BioASQ tasks, we also used PubMedBERT with KU-DMIS's method [6] for the exact answer task (similar to SQuAD). This method converts BioASQ's List, Factoid question and answer data format to a format similar to SQuAD. Then, it uses a fine-tuning method similar to Google's BERT on SQuAD for model training. For the Yes/No problem, it adds a linear regression layer to the BERT model for sequence binary classification. These methods have performed well in past challenges. To simplify the parameters adjustment process, We used Microsoft's open source AutoML system NNI [17] to automatically adjust the parameters of this task. However, in the ideal answer task, we did not perform multiple experiments because of the time limit.

For the hardware, we used an NVIDIA GeForce GTX 1080 GPU for Factoid, List question exact answer tasks. Ideal answer tasks and Yes/no question exact answer tasks were trained using an NVIDIA Tesla T4 GPU provided by Google Colab. Because of the limitation of GPU memory, we reduce the batch size for Factoid, List type question tasks to 4, which may affect the performance of following experiments.

4. Submission

Our submitted configurations are summarized in **Table 1.** We tested the performance of the pre-trained language models by conducting experiments with BioASQ 9b data for each task of the exact answer. Since the results of PubMedBERT are not as good as the BioBERT-related models in the

experiments, we only submit the best KU-DMIS BioBERT-related model results. The BioBERT-MNLI model was used for the Yes/no questions, while the BioBERT-MNLI-SQuAD model was used for both the Factoid and List questions. We should additionally mention that all three systems use the same answer for the exact answer submitted.

Table 1. Descriptions of our three systems

System Name	System Description	Participating Batch
NCU-IISR/AS-GIS-1	**Exact answers:** Using KU-DMIS BioBERT related models. **Ideal answers:** Using **BioBERT** with predicted ROUGE-SU4 scores to select the top 1 sentences of snippets.	4,5
NCU-IISR/AS-GIS-2	**Exact answers:** Using KU-DMIS BioBERT related models. **Ideal answers:** Using **BioBERT-MNLI** with predicted ROUGE-SU4 scores to select the top 1 sentences of snippets.	4,5
NCU-IISR/AS-GIS-3	**Exact answers:** Using KU-DMIS BioBERT related models. **Ideal answers:** Using **PubMedBERT** with predicted ROUGE-SU4 scores to select the top 1 sentences of snippets.	4,5

Table 2. Results of test batch 4,5 for exact answers in the BioASQ QA task. Total Systems counts the number of participants for each batch in the given category. For example, our system ranked third in batch 5 in Yes/no questions. Best Score indicates the best result across all participants, and Median Score the median result.

Batch	**Yes/no**		**Factoid**		**List**	
	System Name	Macro F1	System Name	MRR	System Name	F-Measure
4	Best Score	0.9480	Best Score	0.6929	Best Score	0.7061
	Ours	0.8441	Ours	0.4232	Ours	0.4261
	Median Score	0.4186	Median Score	0.5030	Median Score	0.4960
Total systems	*52*		*41*		*30*	
5	Best Score	0.8246	Best Score	0.5880	Best Score	0.5175
	Ours	0.7738(#3)	Ours	0.5287	Ours	0.3673
	Median Score	0.5522	Median Score	0.4722	Median Score	0.3438
Total systems	*56*		*45*		*38*	

Model performances in predicting exact answers are shown in **Table 2**. Our system performed better than the median system score for all three question types in batch 5. In particular, our system generally performed higher in the Yes/no category than on the other two question types and scored near the best

Macro F1 scores for both batch 4 and batch 5. Among them, we ranked third in the fifth batch in terms of Yes/no type questions.

Table 3. Results (ROUGE-2 and ROUGE-SU4 F1 scores and Recall scores) of test batch 4,5 for ideal answers in the BioASQ QA task. Total Systems counts the number of participants in each batch. In batch 4 and 5, our system "NCU-IISR/AS-GIS-2" took first place out of submitted systems in both F1 scores. However, the Recall Score of our systems are lower than the best score.

System Name	Batch 4	Batch 5
	ROUGE-2 F1	
Best Score	*0.3790(#2)*	*0.3846(#2)*
NCU-IISR/AS-GIS-1	0.3280	0.2839
NCU-IISR/AS-GIS-2	**0.4454(#1)**	**0.3946(#1)**
NCU-IISR/AS-GIS-3	0.2694	0.2817
Median Score	0.3414	0.2629
	ROUGE-SU4 F1	
Best Score	*0.3681(#2)*	*0.3733(#2)*
NCU-IISR/AS-GIS-1	0.3318	0.2846
NCU-IISR/AS-GIS-2	**0.4402(#1)**	**0.3893(#1)**
NCU-IISR/AS-GIS-3	0.2674	0.2666
Median Score	0.3330	0.2573
	ROUGE-2 Recall	
Best Score	*0.7124*	*0.6056*
NCU-IISR/AS-GIS-1	0.3370	0.2962
NCU-IISR/AS-GIS-2	0.4505	0.4072
NCU-IISR/AS-GIS-3	0.2830	0.2817
Median Score	0.4505	0.2863
	ROUGE-SU4 Recall	
Best Score	*0.7107*	*0.6077*
NCU-IISR/AS-GIS-1	0.2851	0.3093
NCU-IISR/AS-GIS-2	0.4550	0.4087
NCU-IISR/AS-GIS-3	0.3471	0.2939
Median Score	0.4550	0.2926
Total Systems	*31*	*28*

The performance of the model in predicting the ideal answer is shown in **Table 3**. For ideal answers, BioASQ used two evaluation metrics: ROUGE and human evaluation. Roughly speaking, ROUGE calculates the n-gram overlap between an automatically constructed summary and a set of human-

written (golden) summaries, with higher ROUGE scores being better. Specifically, ROUGE-2 and ROUGE-SU4 were used to evaluate ideal answers. These automatic evaluations are the most widely used versions of ROUGE and have been discovered to correlate well with human judgments when multiple reference summaries are available for each question. The organizers have not yet reported the results of the human evaluation (manual scoring). All ideal system answers will also be evaluated by biomedical experts.

In batch 4 and 5, our system "NCU-IISR/AS-GIS-2" took first place out of submitted systems in ROUGE-2 F1 and ROUGE-SU4 F1. In particular, in batch 4, the "NCU-IISR/AS-GIS-2" system scored 0.0664 (ROUGE-2) and 0.0721 (ROUGE-SU4) higher than the second-ranked system in terms of F1 score. However, the Recall Score of our systems are lower than the best score. This may be related to the fact that we ended up submitting only the top 1 sentence. We considered increasing the number of sentences submitted, but in the end, we did not have time to test it.

Results of internal experiments for exact answers on the BioASQ 9b dataset are shown in **Table 4**. Both Factoid and List type problem experiments were performed using NNI to fine-tune the parameters. Each model was experimented at least 20 times to find the best performance. We conducted this experiment to examine whether PubMedBERT could achieve better results than BioBERT on the BioASQ task. The results are generally in line with our expectations. For Factoid and List type questions, PubMedBERT (especially the Fulltext version) outperformed the basic version of BioBERT. But for Yes/no questions, PubMedBERT was not even as good as the basic version of BioBERT. This result is similar to what we have seen in ideal answer, where the ROUGE-related scores of the system using PubMedBERT "NCU-IISR/AS-GIS-3" are worse than the system of the version using BioBERT "NCU-IISR/AS-GIS-1".

Table 4. Results of internal experiments for exact answers on the BioASQ 9b dataset. Because of the difference in problem types, not all BioBERT-related models have been used in the experiments. Although PubMedBERT-Fulltext outperformed the basic version of BioBERT for Factoid, List type questions, the score was still much lower than the BioBERT-MNLI-SQuAD results.

Pretrained Model Name	Yes/no*	Factoid**	List***
	Macro F1	MRR	F-Measure
BioBERT	0.7659	0.3990	0.3518
BioBERT-MNLI	**0.8671**	-	-
BioBERT-MNLI-SQuAD	-	**0.4509**	**0.3740**
PubMedBERT-Abstract	0.7199	0.4020	0.3470
PubMedBERT-Fulltext	0.6960	0.4248	0.3548

** Because results of Yes/no questions were too disparate, we did not conduct enough experiments to adjust the parameters to achieve the best performance. Except for the BioBERT-MNLI, we run the experiments for only three times for each model.*

***Parameter search space for Factoid type question task: [1e-6 - 5e-5] for learning rate, [4,6] for batch size and [2,3,4] for epoch.*

**** Parameter search space for List type question task: [1e-6 - 1e-5] for learning rate, [4] for batch size and [1,2] for epoch.*

Although PubMedBERT has better results than BioBERT for some tasks, it still has a gap with BioBERT-MNLI and BioBERT-MNLI-SQuAD, which have been fine-tuned with external datasets. Therefore, we did not use the PubMedBERT trained exact answer system for formal submissions. In addition, we also tried to fine-tune PubMedBERT on MultiNLI and SQuAD datasets to get better results. However, we were not able to make any progress until the end of the competition.

5. Discussions and Conclusions

In the 9th BioASQ QA task, we used pre-trained models including BioBERT, BioBERT-MNLI, BioBERT-MNLI-SQuAD, PubMedBERT to generate both the exact and ideal answers. In generating exact answers, we use the KU-DMIS approach to find the offsets (both start and end positions) of the answer within the given passage (snippets). Although PubMedBERT outperforms the basic version of BioBERT in Factoid, List type questions, it still cannot reach the performance of BioBERT-MNLI-SQuAD which has been fine-tuned with external datasets. This result indicates the significant effect of sequential learning using existing datasets.

When it comes to the ideal answer, the most relevant fragment or sentence was selected in order to maintain the integrity of the ideal answer, rather than taking the fragment offset approach, which may focus on the wrong location and produce imperfect answers. Our results combining BioBERT-MNLI with linear regression ranked first for both ROUGE-2 F1 and ROUGE-SU4 F1 scores in batch 4 and 5. Our results show that using the linear regression model to select sentences can yield excellent results in generating ideal answers. At the same time, BioBERT's performance on this task was significantly improved after fine-tuning with the MultiNLI dataset, which means that the sentence entailment relationships contained in the MultiNLI dataset is useful for finding the ideal answer.

However, we also found that the combined PubMedBERT scored worse than the basic version of BioBERT in generating answers for the ideal answer and the exact answer to the Yes/no question. We speculate that this may be related to the difference between pre-trained corpus of PubMedBERT and BioBERT. PubMedBERT was not trained on the general field corpus such as Wikipedia and Books, but pre-trained from scratch on the PubMed research papers corpus. Are differences between the general field corpus and the research paper corpus likely to contribute to the differences in the predictive results of these two tasks? Are there any linguistic elements that are present in general field texts but missing in biomedical research texts? We do not have sufficient evidence to answer these questions now. Future research could further explore the possible reasons for this discrepancy and conduct more experiments.

Directions for improvement for our system also include expanding the range of snippets to include full abstracts, and comparing activation or loss functions to find a better one. In the regression method, we only processed snippet context and did not use the complete PubMed abstracts. Thus, these can be utilized in the future. All told, we hope to keep the base of pre-trained language model and make an effort to combine it with different approaches.

6. Acknowledgments

This study is supported by the Ministry of Science and Technology, Taiwan (No.: MOST 109-2221-E-008-062-MY3).

7. Reference

[1] Jin, Q., et al., *PubMedQA: A dataset for biomedical research question answering.* arXiv preprint arXiv:1909.06146, 2019.

[2] Möller, T., et al. *COVID-QA: A Question Answering Dataset for COVID-19*. in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. 2020.

[3] Tsatsaronis, G., et al., *An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition.* BMC Bioinformatics, 2015. **16**: p. 138.

[4] Han, J.-C. and R.T.-H. Tsai, *NCU-IISR: Using a Pre-trained Language Model and Logistic Regression Model for BioASQ Task 8b Phase B.* 2020.

[5] Lee, J., et al., *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics, 2020. **36**(4): p. 1234-1240.

[6] Jeong, M., et al., *Transferability of natural language inference to biomedical question answering.* arXiv preprint arXiv:2007.00217, 2020.

[7] Gu, Y., et al., *Domain-specific language model pretraining for biomedical natural language processing.* arXiv preprint arXiv:2007.15779, 2020.

[8] Williams, A., N. Nangia, and S.R. Bowman, *A broad-coverage challenge corpus for sentence understanding through inference.* arXiv preprint arXiv:1704.05426, 2017.

[9] Jin, Q., et al., *Biomedical Question Answering: A Comprehensive Review.* arXiv preprint arXiv:2102.05281, 2021.

[10] Beltagy, I., K. Lo, and A. Cohan, *SciBERT: A pretrained language model for scientific text.* arXiv preprint arXiv:1903.10676, 2019.

[11] Huang, K., J. Altosaar, and R. Ranganath, *Clinicalbert: Modeling clinical notes and predicting hospital readmission.* arXiv preprint arXiv:1904.05342, 2019.

[12] Peng, Y., S. Yan, and Z. Lu, *Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets.* arXiv preprint arXiv:1906.05474, 2019.

[13] Clark, C., et al., *BoolQ: Exploring the surprising difficulty of natural yes/no questions.* arXiv preprint arXiv:1905.10044, 2019.

[14] Mollá, D. and C. Jones. *Classification betters regression in query-based multi-document summarisation techniques for question answering*. in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2019. Springer.

[15] Lin, C.-Y. *Rouge: A package for automatic evaluation of summaries*. in *Text summarization branches out*. 2004.

[16] Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.

[17] Microsoft. *Neural Network Intelligence (NNI)*. Available from: https://github.com/microsoft/nni.