NCU-IISR: Using a Pre-trained Language Model and Logistic Regression Model for BioASQ Task 8b Phase B

Jen-Chieh Han[0000-0003-0998-4539] and Richard Tzong-Han Tsai*[0000-0003-0513-107X]

Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan
joyhan@cc.ncu.edu.tw
thtsai@csie.ncu.edu.tw

Abstract. Recent successes in pre-trained language models, such as BERT, RoBERTa, and XLNet, have yielded state-of-the-art results in the natural language processing field. BioASQ is a question answering (QA) benchmark with a public and competitive leaderboard that spurs advancement in large-scale pre-trained language models for biomedical QA. In this paper, we introduce our system for the BioASQ Task 8b Phase B. We employed a pre-trained biomedical language model, BioBERT, to generate "exact" answers for the questions, and a logistic regression model with our sentence embedding to construct the top-n sentences/snippets as a prediction for "ideal" answers. On the final test batch, our best configuration achieved the highest ROUGE-2 and ROUGE-SU4 F1 scores among all participants in the 8th BioASQ QA task (Task 8b, Phase B).

Keywords: Biomedical Question Answering · Pre-trained Language Model · Logistic Regression

1 Introduction

Since 2018, BioASQ[1] (Tsatsaronis et al., 2015) has organized eight challenges on biomedical semantic indexing and question answering. This year, the challenges include three main tasks: Task 8a, Task 8b, and Task MESINESP8. We only participated in Task 8b Phase B (QA task), in which participants are given a biomedical question and list of question-relevant articles/snippets as input, and should return either an exact answer or an ideal answer. The task was provided 3,243 training questions that included 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 con-

* Corresponding author

1 http://bioasq.org/

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.

CEUR-WS.org/Vol-2696/paper_72.pdf

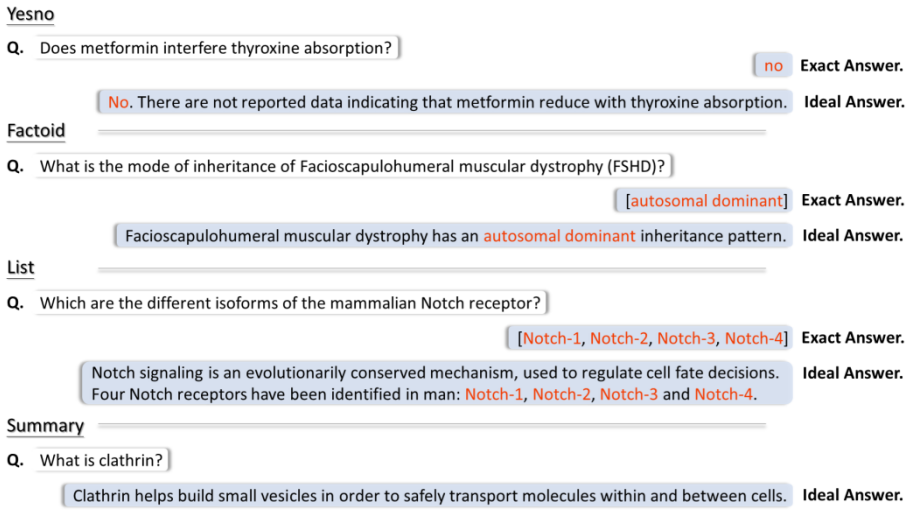

Fig. 1. The QA examples of the BioASQ Task 8b Phase B (QA task).

structed by a team of biomedical experts from around Europe; the questions were categorized into four types: yes/no, factoid, list, and summary. Three types of questions required exact answers: yes/no, factoid, and list. For all four types of questions, participants needed to submit ideal answers. Each participant was allowed to submit a maximum of five results in Task 8b.

Some QA examples are illustrated in **Fig. 1**. Each BioASQ QA instance gives a question and several relevant snippets of PubMed abstracts, including the ID of the full PubMed article. Thus, we formulated the task as query-based multi-document a. extraction for exact answers and b. summarization for ideal answers. In this paper, we employed a pre-trained language model released by BioBERT (Lee et al., 2020), which model achieved the highest performance last year. However, BioBERT was not previously used for generating ideal answers. BioBERT is well-constructed for different natural language processing (NLP) tasks like relation classification and identifying the answer phrase of a question by the given paragraph. BERT uses a masking mechanism to train its language model, thus makes the model learn meanings in different situations. Many biomedical task results show that its language model outperforms traditional word presentation. Therefore, we further applied BioBERT's [CLS] embeddings as input to a logistic regression model for predicting ideal answers.

The sections are organized as follows. Section 2 briefly reviews recent works on QA. The details of our two methods are described separately in Section 3 and 4. Section 5 describes our configurations submitted to the BioASQ challenge. Section 6 gives a summary of our system's performance in the BioASQ QA task.

2 Related Work

In most QA tasks, such as SQuAD[2] (Rajpurkar, Zhang, Lopyrev, & Liang, 2016), SQuAD 2.0 (Rajpurkar, Jia, & Liang, 2018), and PubMedQA (Jin, Dhingra, Liu, Cohen, & Lu, 2019), only exact answers are provided for questions. Exact answers almost always appear in the context of the given relevant articles/snippets; thus, these tasks are usually formulated as a sequence to sequence problem. Recently, it was found that significant improvements can be had in many natural language processing (NLP) tasks by using pre-trained contextual representations (Peters et al., 2018) rather than simple word vectors.

For instance, Google developed Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2018) to solve the problem of shallow bidirectionality. BERT uses a masked language model (MLM) for the pre-training objective, which MLM randomly masks some tokens from the unlabeled input and then predicts the original vocabulary ID of the masked word based on its context. Because MLM jointly concatenates the left and right context as representation, it can pre-train a deep bidirectional Transformer. In BERT's framework, two steps (pre-training and fine-tuning) have the same architectures but different output layers. During fine-tuning, different down-stream tasks initialize models with the same pre-trained model parameters, and all parameters are fine-tuned using labeled data from each task. BERT is the first fine-tuning-based representation model, and its result outperforms prior models on sentence-level and token-level NLP tasks.

Many significant sentence-level classification tasks come from the General Language Understanding Evaluation (GLUE[3]) benchmark (Wang et al., 2018). To help machines understand language just like humans, GLUE provides nine diverse sentence understanding tasks; one example is inputting a pair of sentences, for which the system must predict a relationship with one sentence as the premise and the other as the hypothesis. Where most token-level natural language understanding (NLU) models are designed to carry out a specific task using specific domain data, GLUE is an auxiliary dataset for exploring models with an eye to understanding specific linguistic phenomena across different domains; it thus provides a publicly online platform for evaluating and comparing models.

On the other hand, the two major QA tasks, the Stanford Question Answering Dataset (SQuAD) and SQuAD 2.0, are both token-level tasks. Each instance of the SQuAD gives a question and a passage from Wikipedia, for which the goal is to find the answer text span (start and end position in tokens) in the passage. The SQuAD 2.0 task extends the original SQuAD problem definition by allowing there to be no short answer in the provided paragraph. Each task has an official leaderboard.

Because these NLP tasks have public leaderboards, they are highly competitive and make for rapid expansion in pre-trained models. BERT provided a good start, after which improved models came out such as RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), ALBERT (Lan et al., 2019), and ELECTRA (Clark, Luong, Le, &

[2] https://rajpurkar.github.io/SQuAD-explorer/
[3] https://gluebenchmark.com/leaderboard

Manning, 2020). These models also achieved state-of-the-art results upon being released. The model Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT), based on Google's BERT code, is a language representation model specific to the biomedical domain, pre-trained on large-scale biomedical corpora (1 million articles from PubMed[4] or 270 thousand from PubMed Central[5]). Taking advantage of being able to apply almost the same architecture across tasks, BioBERT largely outperforms previous models and is state-of-the-art in a variety of biomedical text mining tasks.

The BioASQ QA task allows participants to only participate in some batches and to return either only exact answers or ideal answers. The ideal answer includes prominent supportive information, whereas the exact answer only returns yes or no for yes/no questions, entity names for factoid questions, or lists of entity names for list questions; ideal answers can thus be seen as the full definition of exact answers. Ideal answers are usually written by biomedical experts and presented in a short text that answers the question. Because most ideal answers cannot be directly mapped to the given relevant articles/snippets, predicting appropriate ideal answers is more complicated than predicting exact answers.

3 Similarity Between a Snippet and a Question

Although the BioASQ QA task provides biomedical questions and relevant snippets of PubMed abstracts, in actuality, ideal answers do not appear verbatim in the relevant snippets. The goal of our method was to select the most relevant snippet for each question in the BioASQ QA instances. To determine the similarity between a question and a snippet, we directly calculated relevance scores using cosine similarity. Cosine similarity is one of the most common text similarity metrics, thus is widely utilized in NLP tasks. Therefore, we first had to transform questions and snippets into vectors. In general, previous works map words to corresponding vectors by taking word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) embeddings trained on a relevant corpus or else adopt existing word embeddings such as GloVe (Pennington, Socher, & Manning, 2014), and Wiki-PubMed-PMC (Habibi, Weber, Neves, Wiegandt, & Leser, 2017).

A lot of word2vec embeddings and TF-IDF vectors were referred to by Diego Mollá's features (Mollá & Jones, 2019), and we considered that it can be improved. In other words, TF-IDF regarding some common words (such as articles and conjunctions) as trivial terms so as to more readily identify the major words of sentences, these methods are unable to represent polysemic words. Notably, on the GLUE leaderboard, methods using word2vec embeddings (Skip-gram and CBOW) rank much lower than those using the ensemble mode of ELMo, such as BERT. BERT provides contextual embeddings that can solve the problem of polysemy, so deals well with many different tasks. Therefore, we simplified the procedure of extracting features from BioBERT and only took the pre-trained embeddings of sentences.

[4] https://pubmed.ncbi.nlm.nih.gov/

[5] https://www.ncbi.nlm.nih.gov/pmc/

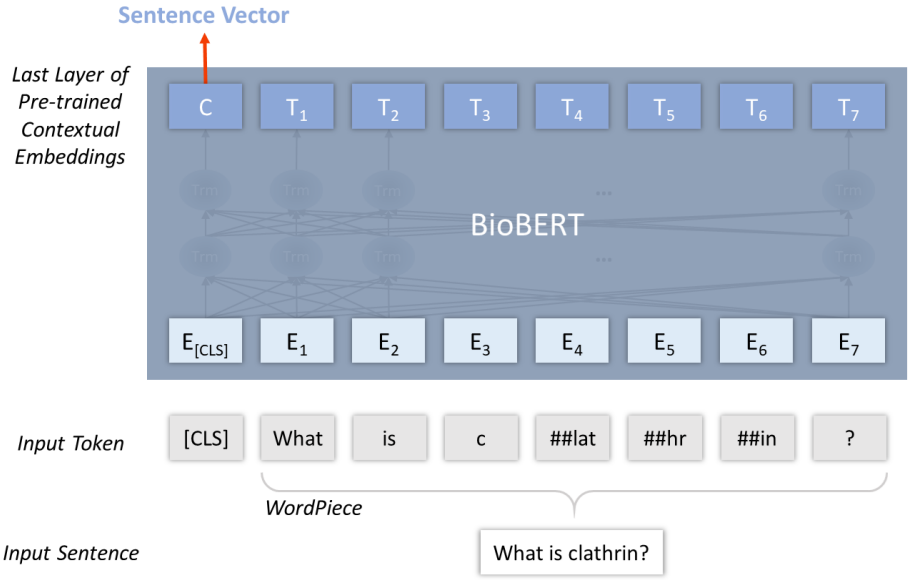

Fig. 2. Illustration of how a single sentence input obtains the pre-trained contextual embeddings (such as ELMo) in the last layer of the pre-trained BioBERT model without fine-tuning.

In our method, before separately obtaining the embeddings of a question and a snippet, each sentence was first pre-processed into word pieces with WordPiece tokenization. Then, inputting all word pieces of the sentence to BioBERT, we extracted the features from the last layer of BioBERT. In BERT, the [CLS] token was inserted into input tokens, and its embeddings could be considered as the sentence vector (the features). The step of extracting pre-trained contextual embeddings from BioBERT is diagrammed in **Fig. 2**.

Finally, we used the embeddings (vectors) of a question and snippet pair to calculate their cosine similarity score. Because each question of a BioASQ QA instance typically has more than one snippet, we re-ranked the snippets in order of their similarity scores and took the top 1 snippet as our prediction of the answer (NCU-IISR_2), as that snippet was considered the most relevant to answering the question.

4 Logistic Regression of Sentences

Our approach was inspired by the framework of the logistic regression model proposed by Diego Mollá. The method follows the two steps of his summarization process: Step 1, split the input text (snippets) into candidate sentences, and score each candidate sentence. Step 2, return the top-*n* sentences with the highest scores. As stated above, we used the pre-trained language model “BioBERT” to replace their features with word embeddings.

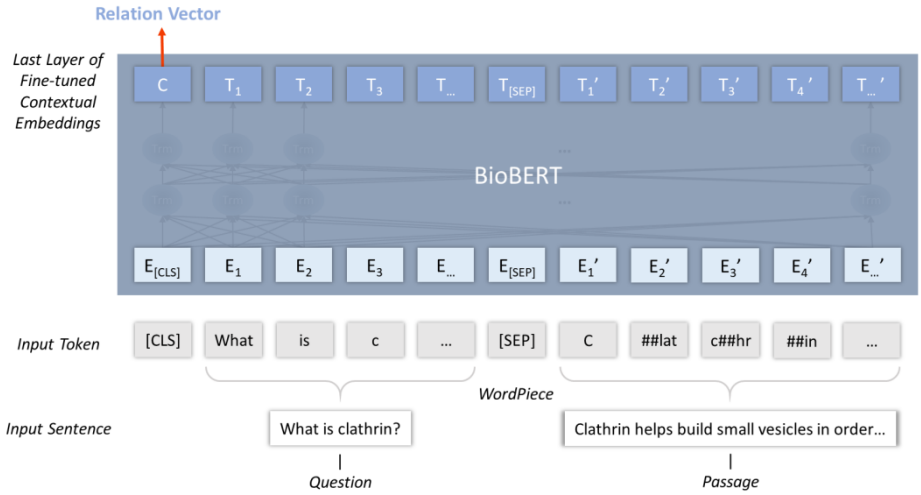

Fig. 3. Illustration showing how a pair of input sentences (a question and a passage) obtains the contextual embeddings in the last layer of the fine-tuned BioBERT.

We first used NLTK's sentence tokenizer to divide snippets into sentences and calculated ROUGE-SU4 F1 scores (Lin, 2004) between each sentence and the associated question, thereby generating positive and negative instances that became the training set for our logistic regression model. After pre-processing, our logistic regression model was slightly different from the cosine similarity method. First, we input a candidate sentence and a question at the same time and used the fine-tuned BioBERT model for fitting the task. Second, we appended a dense layer with ReLU activation after the output layer of BioBERT, and we used mean squared error as the loss function. We took default settings from BERT trained on SQuAD. We also used [CLS] embeddings as the feature from which to predict the ROUGE-SU4 F1 scores of the test data. In our case, [CLS] embeddings represented the relation between a candidate sentence and a question. **Fig. 3** illustrates the modified BioBERT architecture used here. Lastly, we used the prediction values to re-rank the candidate sentences for each question and selected only the top *n* sentences as our system output (NCU-IISR_3).

Due to time limitations, we did not finish aspects of the logistic regression model such as fine-tuning the model with all instances, expanding the range of snippets to the full abstract, and comparing activation or loss functions to find a better one. These can be future work and updates addressed in the next challenge.

5 Submission

To obtain exact answers, we used the BioASQ-BioBert model (Yoon, Lee, Kim, Jeong, & Kang, 2019). This model included two pre-trained weights: one fine-tuned on SQuAD for "yes/no" questions, and the other on SQuAD 2.0 for "factoid and list" questions. We then separately fine-tuned again the yes/no, factoid, and list questions of the

BioASQ QA task. Because BERT performs well on SQuAD, we considered that this method fits suitably into BioASQ's exact answers. We used the open-source code of the BERT and BioBERT pre-trained language model to find the paragraph-sized answer (NCU-IISR_1) additionally in ideal answers. For each training instance, the input is the full PubMed abstracts, and the answer is the snippet.

Our submitted configurations are summarized in **Table 1**. Because our submissions for batch 3 had some errors, **Table 1** only shows the results of batches 1, 2, 4, and 5. In our internal experiments with the "NCU-IISR_3" configuration, we observed that most predictions had lengths as long as ideal answers in the training set. Therefore, we simply selected the top 1 sentence as the ideal answer in all types.

Table 1. Descriptions of our three systems.

System Name	System Description	Section	Participating Batch
NCU-IISR_1	**Exact answers:** Used BioASQ-BioBert. **Ideal answers:** Referred to the SQuAD training in BERT, used snippets from full PubMed abstracts as instances and fine-turned on BioBERT.	-	1, 2, 4, 5
NCU-IISR_2	**Ideal answers:** Used cosine similarity to select the top 1 snippet.	3	5
NCU-IISR_3	**Ideal answers:** Used predicted ROUGE-SU4 scores to select the top n sentences of snippets, where n is equal to 1.	4	5

Model performances in predicting exact answers are shown in **Table 2**. Irrespective of the question type, most of our results outperformed the median scores. In particular, we won second place on the factoid questions at batch 2 and found that "NCU-IISR_1" generally performed higher in the factoid category than on the other two question types.

Model performances in predicting ideal answers are shown in **Table 3**. With ideal answers, two evaluation metrics are used: ROUGE and human evaluation. Roughly speaking, ROUGE counts the *n*-gram overlap between an automatically constructed summary and a set of human-written (gold) summaries, with a higher ROUGE score 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.

Table 2. Results of each test batch (except 3) for exact answers in the BioASQ QA task. Total Systems counts the number of participants for each batch in the given category. For example, in batch 2, we took second place in factoid questions out of 24 submitted systems. There were more systems submitted in later batches. Best Score indicates the best result across all participants, and Median Score the median result.

	Yes/no		**Factoid**		**List**	
Batch	System Name	Macro F1	System Name	MRR	System Name	F-Measure
	Best Score	0.8663	Best Score	0.4688	Best Score	0.4315
1	Ours	0.7243	Ours	-	Ours	-
	Median Score	0.6032	Median Score	0.3156	Median Score	0.3152
Total Systems	*13*		*24*		*23*	
	Best Score	0.9259	Best Score	0.3533	Best Score	0.4735
2	Ours	0.7037	Ours	0.3293 (#2)	Ours	0.2667
	Median Score	0.7000	Median Score	0.2330	Median Score	0.3755
Total Systems	*15*		*24*		*24*	
	Best Score	0.8452	Best Score	0.6284	Best Score	0.4571
4	Ours	0.7204	Ours	0.5735	Ours	0.3905
	Median Score	0.6848	Median Score	0.5211	Median Score	0.3355
Total Systems	*31*		*38*		*37*	
	Best Score	0.8528	Best Score	0.6354	Best Score	0.5618
5	Ours	0.7351	Ours	0.5859	Ours	0.3652
	Median Score	0.7430	Median Score	0.5383	Median Score	0.3652
Total Systems	*32*		*40*		*37*	

The human evaluation results (manual scores) have not yet been reported by the organizers. All ideal answers to the systems will also be evaluated by biomedical experts. For each ideal answer, the experts give a score ranging from 1-5 on each of four terms: information recall (the answer reports all necessary information), information precision (no irrelevant information is reported), information repetition (the answer does not repeat information multiple times, e.g. when sentences extracted from different articles convey the same information), and readability (the answer is easily readable and fluent).

A sample of ideal answers will be evaluated by more than one expert in order to measure the inter-annotator agreement.

Table 3. Results (ROUGE-2 and ROUGE-SU4 F1 scores) of each test batch (except 3) for ideal answers in the BioASQ QA task. Total Systems counts the number of participants in each batch. In batch 5, our system "NCU-IISR_3" took first place out of 28 submitted systems in both scores.

System Name	Batch 1	Batch 2	Batch 4	Batch 5
		ROUGE-2 F1		
Best Score	0.3660	0.3451	0.3087	0.3468 (#2)
NCU-IISR_1	0.1955	0.1675	0.1773	0.2009
NCU-IISR_2	-	-	-	0.2904
NCU-IISR_3	-	-	-	**0.3668**
Median Score	0.1567	0.20765	0.26245	0.3246
		ROUGE-SU4 F1		
Best Score	0.3556	0.3376	0.3001	0.3316 (#2)
NCU-IISR_1	0.1980	0.1652	0.1724	0.1889
NCU-IISR_2	-	-	-	0.2823
NCU-IISR_3	-	-	-	**0.3548**
Median Score	0.1574	0.2058	0.25825	0.31435
Total Systems	*19*	*26*	*26*	*28*

Automatic evaluations in the BioASQ also provide a Recall metric, which shows how many tokens from the prediction appear in the gold answer. For ideal answers, our recall values were lower than the median. The ROUGE-2 and ROUGE-SU4 Recall values for our best system "NCU-IISR_3" are given in **Table 4**. As mentioned earlier, we only returned the top 1 sentence from the logistic regression model, thus we definitely lost some sentences that would have added to ideal answers. In contrast, Diego Mollá's work concatenated the top-*n* sentences when answering questions. If we compile answers from more sentences, we may solve the problem of poor Recall scores. This also can be a direction for improvement in the future.

Table 4. Recall scores (ROUGE-2 and ROUGE-SU4) of ideal answers from test batch 5 in the BioASQ QA task, including the number one to three highest scores and the median score. Our Recall scores were around 0.28 lower than the #1 system.

System Name	Batch 5	
	ROUGE-2 Recall	*ROUGE-SU4 Recall*
#1	0.6646	0.6603
#2	0.6627	0.6587
#3	0.6431	0.6399
NCU-IISR_3	0.3867	0.3805
Median Score	0.4620	0.4650
Total Systems	*28*	

6 Conclusions & Future Work

In the 8th BioASQ QA task, we employed BioBERT to deal with both exact answers and ideal answers. In generating exact answers, we used BioASQ-BioBert to find the offset (including the start and end positions) of the answer within the given passage (snippets). Our performance was almost always above the median for yes/no, factoid, and list question types. However, when it comes to ideal answers, the BioASQ-BioBert method does not readily recognize the most relevant text. In order to maintain the completeness of ideal answers, we selected the most relevant snippet or sentences rather than taking snippet offsets, which may focus on the wrong position and yield imperfect answers.

Our results show that in arriving at ideal answers, using the logistic regression model to select sentences performs better than using cosine similarity to choose a snippet. One reason for this improvement might be that a large number of snippets are too lengthy for ideal answers, thus resulting in lower performance. In other words, snippet answers that consist of only trivial information receive lower ROUGE scores. Our method of selecting sentences achieved the best ROUGE-2 and ROUGE-SU4 F1 scores among all participants, but we also note that our Recall scores were much lower than others. This suggests that our potential improvement with the regression method was unable to convert more possible sentences.

In future work, we may try to solve this problem by referring to other methods and merging in their models. On the other hand, as mentioned previously, we left some work unfinished in the regression experiment. Thus, future directions include completely fine-tuning the model with all instances, 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 BioBERT and make an effort to combine it with different approaches.

Acknowledgments

Appreciating Po-Ting Lai for giving us suggestions during the challenge and revising the paper.

References

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., . . . Polychronopoulos, D. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, *16*(1), 138.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). PubMedQA: A Dataset for Biomedical Research Question Answering. *arXiv preprint arXiv:1909.06146*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). *Xlnet: Generalized autoregressive pretraining for language understanding.* Paper presented at the Advances in neural information processing systems.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Paper presented at the Advances in neural information processing systems.

Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation.* Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics, 33*(14), i37-i48.

Mollá, D., & Jones, C. (2019). *Classification betters regression in query-based multi-document summarisation techniques for question answering.* Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Lin, C.-Y. (2004, jul). *ROUGE: A Package for Automatic Evaluation of Summaries.* Paper presented at the Text Summarization Branches Out, Barcelona, Spain.

Yoon, W., Lee, J., Kim, D., Jeong, M., & Kang, J. (2019). Pre-trained Language Model for Biomedical Question Answering. *arXiv preprint arXiv:1909.08229*.