

# Transferability of Natural Language Inference to Biomedical Question Answering

Minbyul Jeong<sup>1</sup>[0000-0002-1346-730X]\*, Mujeen Sung<sup>1</sup>[0000-0002-7978-8114] \*,  
 Gangwoo Kim<sup>1</sup>[0000-0003-4581-0384], Donghyeon Kim<sup>2</sup>[0000-0002-8224-8354],  
 Wonjin Yoon<sup>1</sup>[0000-0002-6435-548X], Jaehyo Yoo<sup>1</sup>[0000-0002-3600-6362], and  
 Jaewoo Kang<sup>1</sup>[0000-0001-6798-9106]

<sup>1</sup> Department of Computer Science and Engineering, Korea University, Seoul, Korea

<sup>2</sup> AIR Lab, Hyundai Motor Company, Seoul, Korea

{minbyuljeong, mujeensung, gangwoo\_kim, wjyoon, jaehyoyoo,  
 kangj}@korea.ac.kr, donghyeon.kim@hyundai.com

**Abstract.** Biomedical question answering (QA) is a challenging task due to the scarcity of data and the requirement of domain expertise. Pre-trained language models have been used to address these issues. Recently, learning relationships between sentence pairs has been proved to improve performance in general QA. In this paper, we focus on applying BioBERT to transfer the knowledge of natural language inference (NLI) to biomedical QA. We observe that BioBERT trained on the NLI dataset obtains better performance on Yes/No (+5.59%), Factoid (+0.53%), List type (+13.58%) questions compared to performance obtained in a previous challenge (BioASQ 7B Phase B). We present a sequential transfer learning method that significantly performed well in the 8<sup>th</sup> BioASQ Challenge (Phase B). In sequential transfer learning, the order in which tasks are fine-tuned is important. We measure an unanswerable rate of the extractive QA setting when the formats of factoid and list type questions are converted to the format of the Stanford Question Answering Dataset (SQuAD).

**Keywords:** Transfer Learning · Domain Adaptation · Natural Language Inference · Biomedical Question Answering

## 1 Introduction

Biomedical question answering (QA) is a challenging task due to the limited amount of data and the requirement of domain expertise. To address these issues, pre-trained language models [13, 26] are used and further fine-tuned on a target task [2, 4, 7, 19, 20, 31, 32, 36]. Although the pre-trained language models improve performance on the target tasks, the models are still short of the upper-bound

---

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.

\* equal contribution

performance in biomedical QA. Sequential transfer learning is based on transfer learning and it is used to further improve biomedical QA performance [2, 20, 36]. For example, fine-tuning on both the SQuAD dataset [28] and the BioASQ dataset [33] results in higher performance than fine-tuning on only the BioASQ dataset. In the general QA domain, learning relationships between sentence pairs first is effective in sequential transfer learning [4, 11, 27, 34, 35]. Thus, in this paper, we fine-tune the task of NLI [1, 10] to improve performance in biomedical QA. We find that performance improves when the objective function of the fine-tuned task becomes similar to the function of the downstream task. We also find that applying the NLI task to the biomedical QA task addresses *task discrepancy*. Task discrepancy refers to the several differences in the distribution of context length, objective function, and domain shift between various fine-tuned tasks.

Specifically, we focus on reducing the discrepancy of context length distribution between NLI and biomedical QA to improve sequential transfer learning performance on the target task. To reduce the discrepancy, we only unify the distributions of context length of the fine-tuned tasks. We reduce the SQuAD context to a single sentence containing the ground truth answer spans [23]. Fine-tuning on a unified distribution reduces the time to train and perform inference on the BioASQ dataset by 52.95% and 25%, respectively. Finally, we measure an unanswerable rate of the extractive QA setting when the format of the BioASQ dataset is converted to the format of the SQuAD dataset.

Our contributions are as follows:

- (i) We show that fine-tuning on an NLI dataset is effective in Yes/No, Factoid, and List type questions in BioASQ dataset.
- (ii) We demonstrate that unifying the distributions of context length between fine-tuned tasks improves the sequential transfer learning performance of biomedical QA.
- (iii) In the Factoid and List type questions, we measure an unanswerable rate of the extractive QA setting, when the format of the BioASQ dataset is converted to that of the SQuAD dataset.

## 2 Related Works

**Transfer Learning** Transfer learning, also known as domain adaptation, refers to the situation of knowledge learned in a previous task that transferred to a subsequent task. In various fields including image processing or natural language processing (NLP), many studies have shown the effectiveness of transfer learning based on deep neural networks [15, 22, 24, 31, 37]. More recently, especially in NLP, pre-trained language models such as ELMo [26] and BERT [13] have been used for transfer learning [4, 11, 13, 18, 19, 21, 25]. In the biomedical domain, unsupervised pre-training has been used for biomedical contextualized representations [9, 16, 20]. BioBERT [20] was fine-tuned on biomedical corpora (e.g., PubMed and PubMed Central) using BERT, and BioBERT can be employed for various tasks in the biomedical or clinical domain [8, 9, 16, 17, 25, 36].

**Transferability of Natural Language Understanding** The authors of [2] transferred the knowledge obtained from the SQuAD dataset to the target BioASQ dataset to address the data scarcity issue. In [20, 36], the authors adopted sequential transfer learning (e.g., *BioBERT-SQuAD-BioASQ*) to improve biomedical QA performance. Meanwhile, multiple NLI datasets have been constructed for the general domain [1, 3, 10, 28, 35] and domain-specific datasets (e.g., biomedical) have recently been introduced [25, 30]. In [4], the authors have found that fine-tuning on the MultiNLI (MNLI) dataset [1] consistently improves performance on target tasks in terms of all the GLUE benchmarks [35]. The authors of [12] have found that applying knowledge from the NLI dataset improves performance on various yes and no type QA tasks in the general domain. Furthermore, the authors of [34] have used various size datasets in question answering, text classification/regression, and sequence labeling tasks. In this paper, we use the MNLI dataset for improving performance in biomedical QA.

### 3 Methods

In this section, we outline our problem setting for the downstream task. Our training details are provided in the Appendix B. We explain our method of learning biomedical entity representations using BioBERT. Then we describe how to perform the sequential transfer learning of BioBERT for each biomedical question type of the BioASQ Challenge. Our method can be used to apply BioBERT, which was used for training NLI dataset, to biomedical QA.

#### 3.1 Problem Setting

We converted the format of the BioASQ dataset to the format of the SQuAD dataset. In detail, training instances in the BioASQ dataset are composed of a question (Q), human-annotated answers (A), and relevant contexts (C) (also called snippets). If the span of answers was not provided by human annotators, we first found exact spans in contexts based on human-annotated answers to factoid and list type questions. In this case, we enumerated all the combinations of Q-C-A triplets only when the answer span exactly matches the context. *Yes* and *No* answers to Yes/No type questions are not suggested in contexts; thus, we fine-tuned a task-specific binary classifier to predict answers.

#### 3.2 Overall Architecture

Input sequence  $\mathbf{X}$  consists of the concatenation of the BERT [CLS] token, Q, and C, with [SEP] tokens in between Q and C. The sequence is denoted as  $\mathbf{X} = \{[\text{CLS}] \parallel \text{Q} \parallel [\text{SEP}] \parallel \text{C} \parallel [\text{SEP}]\}$  where  $\parallel$  refers to the concatenation of tensors. The hidden representation vector of the  $i^{th}$  input token is denoted as  $h_i \in \mathbb{R}^H$  where  $H$  denotes the hidden size. Finally, we fine-tuned the hidden vectors corresponding to each question type, and the vectors were fed into a softmax classifier or binary classifier.

**Yes/No Type** For computing the yes probability  $P^{yes}$ , we projected a linear transformation matrix  $M \in \mathbb{R}^{1 \times H}$  to transform the hidden representation of a [CLS] token  $C \in \mathbb{R}^H$ . In binary classification, the sigmoid function can be used to calculate the yes probability as follows:

$$P^{yes} = \frac{1}{1 + e^{-C \cdot M^\top}} \quad (1)$$

The binary cross entropy loss is utilized between the yes probability  $P^{yes}$  and its corresponding ground truth answer  $a_{yes}$ . Our total loss is computed as below.

$$Loss = -(a_{yes} \log P^{yes} + (1 - a_{yes}) \log(1 - P^{yes})) \quad (2)$$

**Factoid & List type** At hidden representation vectors, the start and end vectors of answer spans were computed in one linear transformation matrix  $M \in \mathbb{R}^{2 \times H}$ . Let us denote the  $i^{th}$  and  $j^{th}$  predicted answer tokens as *start* and *end*, respectively. The probability of  $(P_i^{start}, P_j^{end})$  can be calculated as follows:

$$P_i = P_i^{start} \parallel P_i^{end} = \frac{e^{h_i \cdot M^\top}}{\sum_{t=1}^s e^{h_t \cdot M^\top}}, \quad P_j = P_j^{start} \parallel P_j^{end} = \frac{e^{h_j \cdot M^\top}}{\sum_{t=1}^s e^{h_t \cdot M^\top}} \quad (3)$$

where  $s$  denotes the sequence length of BioBERT and  $\cdot$  is the dot-product. Our objective function is the negative log-likelihood for the predicted answer with the ground truth answer position. Start and end position losses are computed as below:

$$Loss^{start} = -\frac{1}{N} \sum_{n=1}^N \log P_{a_s}^{start, n}, \quad Loss^{end} = -\frac{1}{N} \sum_{n=1}^N \log P_{a_e}^{end, n} \quad (4)$$

where  $N$  denotes the batch size, and  $a_s$  and  $a_e$  are the ground truth answers of the start and end positions of each instance, respectively. Our total loss is the arithmetic mean of  $Loss^{start}$  and  $Loss^{end}$ .

### 3.3 Transferability in domains and tasks

**Yes/No Type** Training a model to classify relationships of sentence pairs can enhance its performance on yes or no type questions in the general domain [12]. Based on this finding, we believe that a classifier could be used for yes and no type questions in biomedical QA. Thus, we fine-tuned BioBERT on the NLI task so that it can be used to answer biomedical yes or no type questions. We used the MNLI dataset because it is widely used and has a sufficient amount of data from various genres. Furthermore, as shown in Table 9 and 10, sequential transfer learning models trained on the MNLI dataset obtained meaningful results. For our learning sequence, we fine-tuned BioBERT on the MNLI dataset which contains the relationships between hypothesis and premise sentences. We

composed a sequential transfer learning method, denoted as *BioBERT-MNLI-BioASQ*. However, using the final layer of the MNLI task instead of the binary classifier to compute  $P^{yes}$  does not improve the performance of BioBERT on the BioASQ dataset. For this reason, we added a simple binary classifier on the top layer of BioBERT. Furthermore, the distributions of context length in the MNLI dataset and the distributions of snippets of Yes/No type questions in the BioASQ dataset are similar. Therefore, we did not unify the context length distributions of yes and no type questions.

**Factoid & List Type** The order of sequential transfer learning is important in bridging the gap between different tasks. Performance improves when the objective function of the fine-tuned task becomes similar to that of the downstream task in Table 5. Thus, we used the learning sequence *BioBERT-MNLI-SQuAD-BioASQ* instead of *BioBERT-SQuAD-MNLI-BioASQ*. To address the discrepancy of context length distribution between the SQuAD dataset and the BioASQ dataset, we slightly modified the original experimental setting. As suggested in [23], we reorganized the context length distributions in the SQuAD dataset which is similar to the MNLI dataset and BioASQ dataset. We developed an extractive QA setting that is scalable to minimal context and that does not use irrelevant sentences in full abstracts [36]. Therefore, we extracted a sentence containing the ground truth answer span and set as a complete paragraph to construct the minimal context. As a result, we reduced the discrepancy of context length distribution by unifying the context length distributions for our sequential transfer learning. Unifying the distributions of context length reduced the time to train and perform inference on factoid and list type questions. Our method achieved comparable results to those of the baseline method.

## 4 Experiments

### 4.1 Datasets

Our datasets are based on the pre-processed datasets provided by [1, 28, 36]. For the extractive QA setting, we converted the BioASQ dataset format (Yes/No, Factoid, and List type questions) to the format of the SQuAD dataset. In [36], the authors suggested three pre-processing strategies, and for our study, we utilized two of the three strategies: *Snippet-as-is* and *Full-Abstract*. However, we added the criterion of having a blank space before and after each biomedical entity. This criterion has shown to improve performance in distinguishing biomedical named entities. The statistics of the pre-processed dataset are listed in Table 8. We have made the pre-processed BioASQ datasets publicly available.<sup>3</sup> In the experimental setting, we removed approximately 5K training instances from the SQuAD dataset because their answer spans do not exactly match the context.

<sup>3</sup> <https://github.com/dmis-lab/bioasq8b>

Reference System	Yes/No (Macro F1)	Factoid (MRR)	List (F1)
Dimitriadis & Tsoumakas [14]	0.5541	-	-
Hosein et al., [7]	-	0.4562	-
Oita et al., [5]	0.4831	-	-
Resta et al., [29]	0.7873	-	-
Telukuntla et al., [6]	0.4486	0.4751	0.2002
Yoon et al., [36]	0.7169	0.5116	0.4061
Ours	<b>0.8432</b>	<b>0.5163</b>	<b>0.5419</b>

**Table 1.** BioASQ 7B (Phase B) Challenge results and our results. We use a dash (-) if the paper does not contain results on each question type. All the scores were averaged when the batch results are reported in each paper. In each column, the best score is in **bold**.

## 4.2 Experimental Results

In Table 1, we compare our results with the best results from last year’s BioASQ Challenge Task 7B (Phase B) [5–7, 14, 29, 36]. From this comparison, we observe that training BioBERT on the MNLI dataset significantly improves its performance on the Yes/No (+5.59%), Factoid (+0.53%), and List (+13.58%) type questions.

Yes/No Type					
# of Tasks	Sequence of Transfer Learning	Evaluation Metric			
		Accuracy	Yes F1	No F1	Macro F1
6B Test	BioBERT-SQuAD-BioASQ	0.8518	0.9004	0.6896	0.7950
	BioBERT-MNLI-BioASQ	<b>0.8857</b>	<b>0.9212</b>	<b>0.7798</b>	<b>0.8505</b>
7B Test	BioBERT-SQuAD-BioASQ	0.8595	0.8990	0.7344	0.8167
	BioBERT-MNLI-BioASQ	<b>0.8945</b>	<b>0.9275</b>	<b>0.7588</b>	<b>0.8432</b>

**Table 2.** Yes/No type question experiments. Evaluation metrics are accuracy (Accuracy), F1 score, and macro F1 score (Macro F1). The F1-score of yes type questions is denoted as Yes F1, and the F1 score of the no type questions is denoted as No F1. In the columns, the best score obtained in each task is in **bold**.

First, the Yes/No type question scores obtained by our method are shown in Table 2. We observed that using the SQuAD dataset for intermediate fine-tuning improves performance [2, 20, 36]. Therefore, we evaluated our proposed method of fine-tuning BioBERT using the sequence *BioBERT-SQuAD-BioASQ*, as done in [20, 36]. BioBERT is trained on the SQuAD dataset for the QA task. Fine-tuning BioBERT with the sequence *BioBERT-MNLI-BioASQ* significantly improves its performance. BioBERT obtains higher macro F1 scores (+5.55%, +2.65%) than the baseline. We believe selecting yes and no type questions in the

Context Length Discrepancy								
# of Tasks	Setting	Sequence of Transfer Learning	Factoid (%)			List (%)		
			SAcc	LAcc	MRR	Prec	Recall	F1
6B Test	Original	BioBERT-SQuAD-BioASQ	39.80	57.82	47.22	45.02	<b>47.69</b>	42.34
		BioBERT-MNLI-SQuAD-BioASQ	38.80	<b>61.34</b>	47.42	46.60	47.01	42.44
	Document	BioBERT-SQuAD-BioASQ	39.71	56.37	45.81	46.81	40.26	39.63
		BioBERT-MNLI-SQuAD-BioASQ	39.71	55.10	45.77	46.26	39.23	38.13
	Snippet	BioBERT-SQuAD-BioASQ	38.23	57.34	46.24	<b>48.24</b>	46.86	<b>42.83</b>
		BioBERT-MNLI-SQuAD-BioASQ	<b>41.41</b>	57.40	<b>48.05</b>	46.01	45.95	42.75
7B Test	Original	BioBERT-SQuAD-BioASQ	41.95	58.30	48.66	61.32	52.83	52.36
		BioBERT-MNLI-SQuAD-BioASQ	42.22	61.06	49.85	<b>61.46</b>	<b>54.62</b>	<b>54.19</b>
	Document	BioBERT-SQuAD-BioASQ	44.46	57.98	50.02	58.30	39.19	43.89
		BioBERT-MNLI-SQuAD-BioASQ	43.34	58.13	49.21	61.01	41.82	45.78
	Snippet	BioBERT-SQuAD-BioASQ	40.79	58.93	48.27	60.08	53.96	53.18
		BioBERT-MNLI-SQuAD-BioASQ	<b>45.10</b>	<b>62.45</b>	<b>51.63</b>	60.92	53.12	53.01

**Table 3.** Context Length Discrepancy Experiments. The metrics used to measure performance on factoid type questions are strict accuracy (SAcc), lenient accuracy (LAcc), and mean reciprocal rank (MRR). The metrics used to evaluate performance on list type questions are precision (Prec), recall (Recall), and macro F1 (F1). ‘Original’ indicates training BioBERT on full documents in SQuAD and snippets in BioASQ. ‘Document’ indicates that BioBERT was trained on full documents in SQuAD and full abstracts in BioASQ. ‘Snippet’ denotes training on a unified distribution of minimal context. All five batch results are averaged. In the columns, the best score obtained in each task is in **bold**.

BioASQ dataset is similar to deciding the relationship between sentence pairs in the MNLI dataset. We also replaced the binary classifier of BioBERT, which is trained on the BioASQ dataset, with the final layer of the MNLI task, but this did not improve performance. Thus, we fine-tuned the binary classifier to select yes and no type questions.

When using the factoid and list type questions in the MNLI dataset, we considered the discrepancy of context length distributions. The obtained results are shown in Table 3. In the original experimental setting, full documents in the SQuAD dataset and snippets in the BioASQ dataset were used for training BioBERT. The performance of our method on the 6B test set did not improve. However, we observed that its performance improves with the size of the training set, as shown by the higher performance on the 7B test set compared with that on the 6B test set.

In the document setting, we used the whole paragraphs and the full abstracts of the SQuAD and BioASQ datasets, respectively. Performance obtained in this setting is lower than that obtained in the original setting due to using longer context rather than snippets in the BioASQ dataset. In other words, rather than using the human annotated corpus (i.e., snippets), the search space in which an answer can be found was expanded to full abstracts. Nevertheless, the

# of Batches	Yes/No		Factoid		List		Macro Avg.
	System Name	Macro F1	System Name	MRR	System Name	F1	
8B batch 1	Ours	<b>0.8663</b>	Ours	<b>0.4438</b>	Ours	<b>0.3718</b>	<b>0.5606</b>
	FudanLabZhu1	0.4518	FudanLabZhu1	0.4557	FudanLabZhu1	0.3408	0.4161
	Umass_czi_4	0.5989	Umass_czi_4	0.3005	Umass_czi_4	0.3448	0.4147
8B batch 2	Ours	<b>0.8928</b>	Ours	<b>0.3533</b>	Ours	<b>0.3798</b>	<b>0.5420</b>
	UoT_multitask_learn	0.7000	UoT_multitask_learn	0.2800	UoT_multitask_learn	0.4108	0.4636
	FudanLabZhu4	0.6303	FudanLabZhu4	0.2900	FudanLabZhu4	0.4678	0.4627
8B batch 3	Umass_czi_4	0.9016	Umass_czi_4	0.3810	Umass_czi_4	0.4522	0.5782
	Ours	<b>0.9028</b>	Ours	<b>0.3601</b>	Ours	<b>0.4520</b>	<b>0.5716</b>
	pa-base	0.8995	pa-base	0.3137	pa-base	0.4585	0.5572
8B batch 4	Ours	<b>0.7636</b>	Ours	<b>0.6078</b>	Ours	<b>0.4037</b>	<b>0.5917</b>
	91-initial-Bio	0.7204	91-initial-Bio	0.5735	91-initial-Bio	0.3905	0.5615
	Features Fusion	0.7097	Features Fusion	0.5745	Features Fusion	0.3625	0.5489
8B batch 5	Ours	<b>0.8518</b>	Ours	<b>0.5677</b>	Ours	<b>0.5582</b>	<b>0.6592</b>
	Parameters retrained	0.7509	Parameters retrained	0.5938	Parameters retrained	0.4004	0.5817
	Features Fusion	0.7509	Features Fusion	0.6115	Features Fusion	0.3810	0.5811

**Table 4.** BioASQ 8B results obtained by the top three systems. The best scores were obtained from the BioASQ leaderboard (<http://participants-area.bioasq.org/results/8b/phaseB/>). We considered a system with different names as one system with the highest scores. We report the macro average scores obtained on all types of questions in the BioASQ dataset. Our systems are in **bold**.

performance of our proposed method on the factoid type questions in the 7B test set improved when BioBERT was fine-tuned on the SQuAD dataset.

For the snippet setting, we unify the distributions of context length in the extractive QA setting. Our method extracts the sentence containing the ground truth answer span, i.e., the minimal context; the performance of our method on the 6B & 7B test sets significantly improved. We recognize that it is hard to prove the generalization of our method because the test sets for the BioASQ dataset are too small and the variance of performance is relatively high. However, we demonstrate our superior performance by reducing the task discrepancy of factoid type questions in 6B & and 7B. Although, we have achieved better performance of list type questions, reducing the discrepancy of context length distribution does not significantly affect. We believe that given the objective function of list type questions, it needs further analyses to demonstrate the generalization of sequential transfer learning with fine-tuning NLI dataset.

## 5 Analysis

**Order of Sequential Transfer Learning** The BioASQ Challenge Task 8B (Phase B) results are shown in Table 4. Each team was allowed to submit up to five systems with different combinations of features. The 8B ground truth answers are not available so we could not use them for manually evaluating our proposed method. Thus, we report the scores from the leaderboard.<sup>4</sup>

<sup>4</sup> <http://participants-area.bioasq.org/results/8b/phaseB/>

Order Importance							
# of Tasks	Sequence of Transfer Learning	Factoid (%)			List (%)		
		SAcc	LAcc	MRR	Prec	Recall	F1
6B Test	BioBERT-SQuAD-BioASQ	39.80	57.82	47.22	45.02	<b>47.69</b>	42.34
	BioBERT-SQuAD-MNLI-BioASQ	<b>41.15</b>	57.95	47.29	46.18	44.56	40.98
	BioBERT-MNLI-SQuAD-BioASQ	38.80	<b>61.34</b>	<b>47.42</b>	<b>46.60</b>	47.01	<b>42.44</b>
7B Test	BioBERT-SQuAD-BioASQ	41.95	58.30	48.66	61.32	52.83	52.36
	BioBERT-SQuAD-MNLI-BioASQ	<b>43.31</b>	58.69	49.24	60.77	50.74	50.72
	BioBERT-MNLI-SQuAD-BioASQ	42.22	<b>61.06</b>	<b>49.85</b>	<b>61.46</b>	<b>54.62</b>	<b>54.19</b>

**Table 5.** Experiments on the importance of the order of sequential transfer learning. The metrics used for measuring performance on factoid-type questions are strict accuracy (SAcc), lenient accuracy (LAcc), and mean reciprocal rank (MRR). The metrics used for evaluating performance on list-type questions are precision (Prec), recall (Recall), and macro F1 (F1). The best score obtained in each task is in **bold**.

In this ablation study, we explore the importance of the order of sequential transfer learning. The results are shown in Table 5. We found that fine-tuning BioBERT on the MNLI dataset improved its performance on factoid type questions. On the other hand, its performance on list type questions improved when the objective function of fine-tuned tasks was similar to that of the BioASQ task. In other words, BioBERT needs to be fine-tuned on the SQuAD dataset after fine-tuning it on the MNLI dataset.

Type	7B Batch1	7B Batch2	7B Batch3	7B Batch4	7B Batch5	7B Total
Factoid	0.359 (14/39)	0.120 (3/25)	0.310 (9/29)	0.118 (4/34)	0.229 (8/35)	0.216 (35/162)
List	0.083 (1/12)	0.235 (4/17)	0.200 (5/25)	0.136 (3/22)	0.500 (6/12)	0.204 (18/88)

**Table 6.** Statistics of the unanswerable rate in the extractive QA setting. The cases where *Ground Truth Answer does not exactly match the context of the Human Annotated Corpus (Snippet)*. The unanswerable rate is related to the upper-bound performance of our proposed method.

**Unanswerable rate of the Extractive QA Setting** So far, the experiments were performed in the extractive QA setting. We manually analyzed differences between the answer span and the context of the human annotated corpus from the BioASQ Challenge Task 7B (Phase B) test set. We used the test set instead of the training set for measuring the unanswerable rate of the extractive QA setting for the following two reasons. First, we wanted to measure the upper-bound performance of our proposed method. Second, the training and test data

of the BioASQ dataset are similar to those of the dataset from the previous year. Table 6 shows the unanswerable rate of all batch results of the 7B test set which contains only factoid and list type questions. We calculated the unanswerable rate of the extractive QA setting using the rule *Ground Truth Answer does not exactly match the context of the Human Annotated Corpus (Snippet)*. The rule applies to the following cases: no exact match, lowercase match, additional phrase added, and different type of blank space between the exact answer and snippet. In Table 7, we randomly sample such cases. Due to the lack of space, we provide more examples of cases at our url <sup>5</sup>. In here, we use the extractive QA setting to measure the upper-bound performance of our method. We hope our analysis is helpful in designing experimental settings.

Limitations of the Supervised Setting	
Type	ID - Question - Context - Answer
Factoid	ID: 5c531d8f7e3cb0e231000017 Question: What causes Bathing suit Ichthyosis(BSI)? Ground Truth Answer: <b>transglutaminase-1 gene (TGM1) mutations</b> Context: Bathing suit ichthyosis (BSI) is an uncommon phenotype classified as a minor variant of autosomal recessive congenital ichthyosis (ARCI). OBJECTIVES: We report a case of BSI in a 3-year-old Tunisian girl with a novel <b>mutation of the transglutaminase 1 gene (TGM1)</b>
	ID: 5c5214207e3cb0e231000003 Question: List potential reasons regarding why potentially important genes are ignored Ground Truth Answer: <b>Identifiable chemical properties, Identifiable physical properties, Identifiable biological properties</b> , <u>Knowledge about homologous genes from model organisms</u> Context: Here, we demonstrate that these differences in attention can be explained, to a large extent, exclusively from a small set of <b>identifiable chemical, physical, and biological properties</b> of genes. Together with <u>knowledge about homologous genes from model organisms</u> , these features allow us to accurately predict the number of publications on individual human genes, the year of their first report, the levels of funding awarded by the National Institutes of Health (NIH), and the development of drugs against disease-associated genes.

**Table 7.** Unanswerable questions of the extractive QA samples used for the BioASQ dataset. We used factoid- and list-type questions from the 7B test set. Context refers to a snippet in the human annotated corpus provided by the organizer of the BioASQ Challenge. No exact matches are in **bold** and exact matches in lowercase are underlined.

## 6 Conclusion

In this work, we used natural language inference (NLI) as a first step in fine-tuning BioBERT for biomedical question answering (QA). Training BioBERT to classify relationships between sentence pairs improved its performance in biomedical QA. We empirically demonstrated that fine-tuning BioBERT on the NLI dataset improved its performance on the BioASQ dataset from the BioASQ Challenge. We unified the distributions of context length to mitigate the discrepancy between NLI and biomedical QA. Furthermore, the order of sequential

<sup>5</sup> <https://github.com/dmis-lab/bioasq8b/tree/master/human-eval>

transfer learning is important when fine-tuning BioBERT. Finally, when converting the format of the BioASQ dataset to the SQuAD format, we measured the unanswerable rate of the extractive QA setting where an answer does not exactly match the human annotated corpus.

## Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF-2020R1A2C3010638, NRF-2016M3A9A7916996) and Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0021).

## References

1. Williams et al., A.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the NAACL: Human Language Technologies, Volume 1 (Long Papers) (2018)
2. Wiese et al., G.: Neural domain adaptation for biomedical question answering. In: Proceedings of the 21st Conference on CoNLL (2017)
3. Levesque et al., H.: The winograd schema challenge. In: Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning (2012)
4. Phang et al., J.: Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. arXiv preprint arXiv:1811.01088 (2018)
5. Oita et al., M.: Semantically corroborating neural attention for biomedical question answering. In: ECML PKDD (2019)
6. Telukuntla et al., S.K.: Uncc biomedical semantic question answering systems. bioasq: Task-7b, phase-b. In: ECML PKDD (2019)
7. Hosein et al., S.: Measuring domain portability and errorpropagation in biomedical qa. arXiv preprint arXiv:1909.09704 (2019)
8. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop (2019)
9. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on EMNLP-IJCNLP
10. Bowman, S., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on EMNLP
11. Chen, S., Hou, Y., Cui, Y., Che, W., Liu, T., Yu, X.: Recall and learn: Fine-tuning deep pretrained language models with less forgetting. arXiv preprint arXiv:2004.12651 (2020)
12. Clark, C., Lee, K., Chang, M.W., Kwiatkowski, T., Collins, M., Toutanova, K.: Boolq: Exploring the surprising difficulty of natural yes/no questions. In: Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (Long and Short Papers) (2019)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the NAACL: Human Language Technologies (2019)

14. Dimitriadis, D., Tsoumakas, G.: Yes/no question answering in bioasq 2019. In: ECML PKDD (2019)
15. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers) (2018)
16. Jin, Q., Dhingra, B., Cohen, W., Lu, X.: Probing biomedical embeddings from language models. In: Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP (2019)
17. Kim, D., Lee, J., So, C.H., Jeon, H., Jeong, M., Choi, Y., Yoon, W., Sung, M., Kang, J.: A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access* (2019)
18. Kim, N., Patel, R., Poliak, A., Xia, P., Wang, A., McCoy, T., Tenney, I., Ross, A., Linzen, T., Van Durme, B., et al.: Probing what different nlp tasks teach machines about function word comprehension. In: Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (\* SEM 2019) (2019)
19. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019)
20. Lee, J., Yoon, W., Kim, S., Kim, D., So, C., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)* (2019)
21. Liu, N.F., Gardner, M., Belinkov, Y., Peters, M.E., Smith, N.A.: Linguistic knowledge and transferability of contextual representations. In: Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1 (Long and Short Papers) (2019)
22. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791* (2015)
23. Min, S., Zhong, V., Socher, R., Xiong, C.: Efficient and robust question answering from minimal context over documents. In: Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers) (2018)
24. Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., Jin, Z.: How transferable are neural networks in nlp applications? In: Proceedings of the 2016 Conference on EMNLP
25. Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task (2019)
26. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the NAACL: Human Language Technologies, Volume 1 (Long Papers) (2018)
27. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019)
28. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on EMNLP
29. Resta, M., Arioli, D., Fagnani, A., Attardi, G.: Transformer models for question answering at bioasq 2019. In: ECML PKDD (2019)
30. Romanov, A., Shivade, C.: Lessons from natural language inference in the clinical domain. In: Proceedings of the 2018 Conference on EMNLP
31. Ruder, S.: Neural transfer learning for natural language processing. Ph.D. thesis (2019)

32. Talmor, A., Berant, J.: Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In: Proceedings of the 57th Annual Meeting of the ACL (2019)
33. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., et al.: An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics (2015)
34. Vu, T., Wang, T., Munkhdalai, T., Sordoni, A., Trischler, A., Mattarella-Micke, A., Maji, S., Iyyer, M.: Exploring and predicting transferability across nlp tasks
35. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: Glue: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (2018)
36. Yoon, W., Lee, J., Kim, D., Jeong, M., Kang, J.: Pre-trained language model for biomedical question answering. arXiv preprint arXiv:1909.08229 (2019)
37. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in NIPS (2014)

## A Dataset Statistic

MNLI		Train		Dev			
Original		392,702		9,815			
SQuAD v1.1		Train		Dev			
Original		87,412		10,570			
Snippet		82,280		9,986			
SQuAD v2.0		Train		Dev			
Original		130,319		11,873			
BioASQ		6B		7B		8B	
Type	Data Strategy	Train	Test	Train	Test	Train	Test
Yes/No	Snippet-as-is	9,421	127	10,560	140	11,531	152
Factoid	Full-Abstract	7,911		9,403		10,147	
	Appended-Snippet	5,953	161	7,179	162	7,896	151
	Snippet-as-is	3,512		4,231		4,759	
List	Full-Abstract	14,008		15,719		16,879	
	Appended-Snippet	10,878	81	12,184	88	13,251	75
	Snippet-as-is	6,922		7,865		8,676	

**Table 8.** Statistics of transferred dataset (MNLI & SQuAD) and target dataset (BioASQ).

## B Training Details

We use BioBERT as learning biomedical entity representation. We utilize a single NVIDIA Titan RTX (24GB) GPU to fine-tune the sequence of transfer learning. In MNLI task, we use hyperparameters suggested by Hugging Face.<sup>6</sup> For fine-tuning, we select the batch size as 12, 24 and a learning rate is within range 1e-6 to 9e-6. In post-processing, we use the abbreviation resolution module called Ab3P<sup>7</sup> to remove the same answer appearance with a different form.

## C Various NLI datasets

Model	Yes/No (%)				Factoid (%)			List (%)		
	Accuracy	Yes F1	No F1	Macro F1	SAcc	LAcc	MRR	Prec	Recall	F1
SQuAD	85.18	90.04	68.96	79.50	39.80	57.82	47.22	45.02	47.69	42.34
MNLI	<b>88.57</b>	92.12	<b>77.98</b>	<b>85.05</b>	38.80	<b>61.34</b>	<b>47.42</b>	<b>47.86</b>	46.89	<b>43.33</b>
SNLI	88.51	<b>92.17</b>	77.47	84.82	39.11	58.23	46.96	44.42	<b>48.16</b>	42.20
MedNLI	77.81	85.24	52.32	68.78	<b>40.05</b>	57.66	47.14	45.56	47.31	42.72

**Table 9.** Experiments of various NLI datasets evaluated on BioASQ 6B (Phase B). The experiments are considered as a first step of the sequential transfer learning. The model of Yes/No type are fine-tuned as same as Table 2. The model of Factoid and List type are fine-tuned as same as Table 3. The best score obtained in each task is in **bold**.

<sup>6</sup> <https://github.com/huggingface/transformers/tree/master/examples/text-classification>

<sup>7</sup> <https://github.com/ncbi-nlp/Ab3P>

Model	Yes/No (%)				Factoid (%)			List (%)		
	Accuracy	Yes F1	No F1	Macro F1	SAcc	LAcc	MRR	Prec	Recall	F1
SQuAD	85.95	89.90	73.44	81.67	41.95	58.30	48.66	61.32	52.83	52.36
MNLI	<b>89.45</b>	<b>92.75</b>	<b>75.88</b>	<b>84.32</b>	<b>42.22</b>	<b>61.06</b>	<b>49.85</b>	<b>61.46</b>	<b>54.62</b>	<b>54.19</b>
SNLI	85.40	90.11	66.95	78.53	41.84	60.03	49.31	56.20	48.07	47.70
MedNLI	78.67	85.38	49.20	67.29	41.45	60.55	49.05	58.40	48.17	48.25

**Table 10.** Experiments of various NLI datasets evaluated on BioASQ 7B (Phase B). The experiments are considered as a first step of the sequential transfer learning. The model of Yes/No type are fine-tuned as same as Table 2. The model of Factoid and List type are fine-tuned as same as Table 3. The best score obtained in each task is in **bold**.