Transfer Learning for Biomedical Question Answering

Arda Akdemir and Tetsuo Shibuya

University of Tokyo, Japan
aakdemir@hgc.jp
tshibuya@hgc.jp

Abstract. Deep Neural Network (DNN) based Machine Learning models achieved remarkable success in many fields of research. Yet, many recent studies show the limitations of these approaches to generalize to unseen examples and to new domains such as the biomedical domain. Besides, supervised-learning based DNN models require a substantial amount of labeled data which is not readily available for many tasks such as the biomedical question answering task. Transfer Learning is shown to mitigate these challenges by transferring information from auxiliary tasks to improve the performance on a source task, and shown to be especially useful for low-resource tasks. These observations and findings motivated us to investigate the effect of transfer learning and multi-task learning on the biomedical question answering task. We proposed a novel multi-task learning model to learn biomedical entities and questions simultaneously. In this work, we explain the three different neural models we used to participate for the BioASQ 8B challenge. Our initial results showed that transferring information from the biomedical entity recognition task brings improvement for the biomedical question answering task.

1 Introduction

Pretrained language models [11, 3] have been frequently leveraged to improve performance on various downstream NLP tasks since their introduction. However, it is shown that the performance of these models, which are trained on general domain corpora, drops significantly when they are tested on a new domain [14]. This performance drop is higher for domains that have significantly different word distributions, such as the biomedical domain. To mitigate this performance drop, a frequently used approach is to pretrain these models on the target domain, which is also called as **domain-adaptation**. Recently, Lee et al. [9] pretrained the BERT [3] language model on the PubMED articles, which is called BioBERT, and achieved state-of-the-art results for several downstream biomedical tasks. This motivated us to use BioBERT as our baseline model in our experiments.

Transfer learning is a general term to describe the learning schemes where the information from a source task is used to improve the performance on a target task. It is shown to be especially useful to improve the performance on low-resource tasks [2]. Ideally, we would like to transfer information from high-resource tasks that have a similar domain with the source task to make the most out of transfer learning. Currently available datasets for biomedical question answering is very limited. Relative to the biomedical question answering datasets, the currently available biomedical entity datasets are large. These findings motivated us to apply transfer learning to improve the performance on the biomedical question answering task. Specifically, we claim that the performance on biomedical question answering can be improved by transferring information from the biomedical entity recognition task. We propose a multi-task learning model that learns both biomedical question answering and entity recognition tasks, which have not been implemented before to the best of our knowledge. Our work can be considered as an extension of the previously proposed BioBERT model. Our model differs from the BioBERT model in two main ways. Unlike the BioBERT model, we propose a single neural architecture to simultaneously learn three question types (factoid, yes/no, list). This allows the model to transfer information between different question types. Next, we propose a multi-task learning model to learn the biomedical entity recognition and question answering tasks. BioBERT uses separate architectures for the two downstream tasks. Thus, the pretrained BioBERT model is fine-tuned from scratch for each task. Unlike BioBERT, our model allows transferring information between these two tasks during the fine-tuning step.

1.1 BioASQ Challenge

BioASQ is a challenge on biomedical semantic indexing and question answering [6]. The challenge aims to advance the state-of-the-art in semantic indexing and question answering, and also establish a reference point for biomedical question answering. More information about the challenge can be obtained from the BioASQ homepage. [1] We participated in the question answering part of the BioASQ 2020 challenge (8B) to test our claim on using transfer learning for biomedical question answering. This paper describes the models we used to make our submissions to the BioASQ 8B challenge. We participated to the challenge with three different neural architectures, and used the BioASQ datasets as our test-bed to compare these proposed models. Our main contributions can be listed as follows:

- We implemented a novel neural architecture that uses a single model to jointly learn three question types in the BioASQ challenge.

[1] http://bioasq.org/

- We proposed a novel multi-task learning model for entity recognition and question answering for the biomedical domain which have not been employed before to the best of our knowledge.
- We analyzed the effect of transferring information from three biomedical entity recognition datasets for the biomedical question answering task.

2 Methodology

In this section we describe each model we used during the BioASQ Task 8b: Biomedical Semantic Question Answering. During the task, we made submissions using five different models, three of which used an identical neural architecture, but the final model is determined using different evaluation methods. We used BioBert-based Question Answering Model [17] as our baseline model, which we refer to as BioBERT_baseline. The second model is an extension of the first model, which jointly learns all question types using a single architecture. We refer to this model as BioBERT_allquestions. We used three variations of this model for our submissions. Finally, we used a novel multi-task learning model that learns biomedical entities and all question types simultaneously. We refer to this model as BioBERT_multitask. For the BioASQ 8B challenge, we only submitted answers for the 'list', 'factoid', and 'yes-no' type questions. 'Summary' type questions require a fundamentally different approach, and was beyond the main scope of this work.

2.1 Pre-processing

The raw input format of the BioASQ dataset needs to be pre-processed into the suitable format expected by the BioBERT model. Following Yoon et al. [17], we used a similar pre-processing scheme to convert the BioASQ questions into the SQUAD Question Answering format. In the BioASQ dataset, multiple gold-answers are provided for most questions. Gold answers are denoted as spans inside the snippets provided for each question. During pre-processing, we treated each gold-label snippet and question pair as separate examples to increase the size of the training set. During all our experiments, we only made use of the gold-label snippets. We did not analyze the effect of appending additional information from external sources such as the links to related documents provided by the BioASQ organizers. Previously, Yoon et al. [17] experimented with various pre-processing methods to bring further improvements. They observed that the benefits of each strategy depend on the question type and the test-batch. For this reason, we fixed the pre-processing method throughout our all experiments to make it clear where the improvements for each proposed model come from. Besides, using only the snippets as input to the neural networks significantly reduces the input size and reduces the overall training time. For factoid and list type questions, each gold-label span is used to create a new Question-Passage pair. An example factoid type question and gold-label spans from the provided

Table 1: An example question and the gold-label spans from the provided snippets from the BioASQ 6B dataset. The gold-label spans are shown in bold.

Question	What is Contrave prescribed for?
Answer 1	Contrave, ... for the potential treatment of **obesity**, is an oral, sustained ...
Answer 2	Contrave is a combination of ... for the treatment of **obesity**, and is used ...
Answer 3	Contrave, a bupropion and ... for the potential treatment of **obesity**.

spans are given in Table 1. The final predictions for the list type questions are handled during the post-processing step, and explained in the relevant section.

Contrary to the previous work that directly adapts the BERT Question Answering Model [3] by modifying the 'is_impossible' field of the SQUAD dataset format for the yes/no type questions, we implemented our own Yes/No component. This enabled us to use the data without adding the 'is_impossible' field, making the dataset format more readable and easier to understand for researchers from the biomedical domain.

2.2 BioBert-based Baseline Model

Pretrained subword contextual embeddings has shown remarkable progress over previous approaches on many downstream Natural Language Processing (NLP) tasks [16, 13, 12]. Specifically the transformer-based BERT model [3] helped achieve state-of-the-art results on many downstream tasks, including question answering.

The performance of models pretrained on general domain corpora (e.g., Wikipedia articles) drops significantly when tested on niche domains such as the biomedical domain. Motivated with this observation, Lee et al. [9] proposed 'BioBERT', BERT architecture pretrained on PubMed articles. The proposed model obtained state-of-the-art results on three different downstream biomedical NLP tasks. Recently, Yoon et al. [17] obtained the best results in the 2019 BioASQ 7B Question Answering Challenge, and achieved state-of-the-art results on all question types (factoid, yes/no, list). In their proposed approach, separate models are trained from scratch for yes/no, and factoid type questions (factoid/list).

For our baseline model, we used this BioBERT-based approach which we refer to as BioBERT_baseline. BERT model is extended with two separate additional neural layers to learn different question types. The overall architectures are given in Figure 1. For the yes/no type questions, the output for the first token ([CLS])) of the final layer of BERT is given as input to a fully connected layer with 2-dim output representing the scores for yes/no scores. This is followed by a softmax layer to convert these scores into probabilities. Given a sequence of n question tokens $Q = q_t : 1 \leq t \leq n$, and m passage tokens $P = p_t : 1 \leq t \leq m$, BioBERT outputs $m+n+2$ fixed-size (L) vectors $V = v_j : 1 \leq j \leq (m + n + 2)$. Next, v_1 is multiplied with an $(L, 2)$ dimensional matrix W to generate scores,

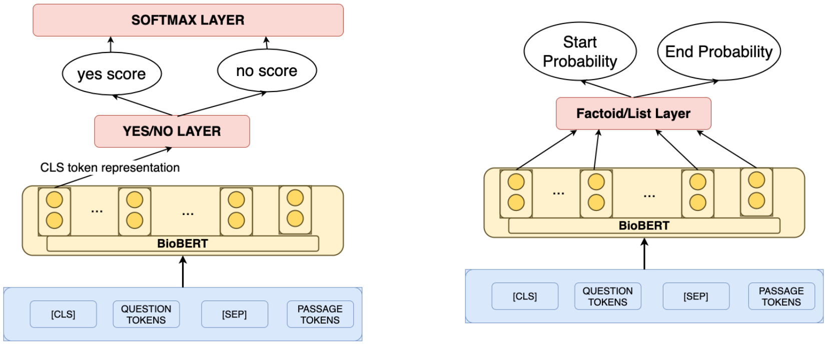
$S = \{s_{yes}, s_{no}\}$, for yes and no answers:

$$V = BioBERT(Q, P)$$
$$S = v_1^T W$$
$$O = Softmax(S)$$

where $O = o_{yes}, o_{no}$ represents the probabilities for each answer, which is the final output for the yes/no type questions. Similarly for the factoid/list type questions, each v_j is multiplied with an $(L, 2)$ dimensional matrix W_2 to generate scores $S_2 = \{s_{start}, s_{end}\}$, which represent the score for the start, end spans for each token p_j inside the input passage P:

$$V = BioBERT(Q, P)$$
$$S_2 = v_j^T W$$

For training, each BioBERT-initialized model in Figure 1 is fine-tuned on the BioASQ-8b for each question type, separately. The main drawback of this previously proposed model is that the common BioBERT layer, which constitutes the majority of the parameters (only a single layer is added for each question type), is fine-tuned separately for each question type. The bottleneck for developing high-performing biomedical question answering systems is the scarcity of the labeled training sets. This approach further limits the training dataset size, and not ideal for low-resource domains like the biomedical domain.



(a) Yes/No model.

(b) Factoid/List model.

Fig. 1: Overall architectures for training separate models for yes/no and factoid/list question types for the BioBERT baseline model [17]. The common BioBERT model layers are finetuned from scratch for each type.

2.3 Joint-Learning Model

The baseline approach does not expose the model to all the examples in the training dataset. This observation motivated us to propose a novel joint-learning model, which uses a single architecture to learn all question types, which we refer to as BioBERT_allquestions. Learning of all question types using a single BERT-based model is not employed before in this domain, to the best of our knowledge. The overview of the proposed joint-learning model is shown in Figure 2. This simple extension to the previously proposed BioBERT-based QA Model [17] allows exposing the model to all the available examples in the training dataset. The common BioBERT layer is trained jointly on all question types. This allows the model to transfer information from other question types for better generalization.

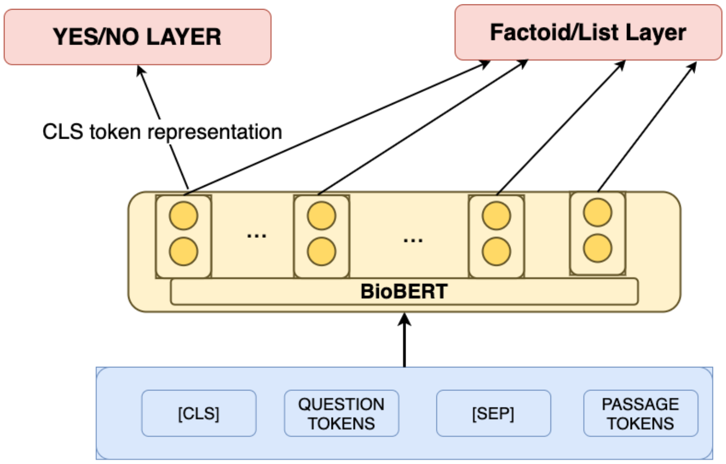


Fig. 2: The proposed joint-learning model for all question types.

An important part of training joint-learning models is the selection of performance metrics. In the conventional single-task machine learning setting, there is usually a single performance metric. The models are evaluated on a development/validation dataset based on this metric, to determine the best performing model during training. In the joint-learning setting, we can evaluate the models based on their performance on each task separately, or we can evaluate them based on their overall performance. For our submissions for the BioASQ 8B challenge we used the following three joint-learning models:

- Overall best-performer
- Best yes/no model
- Best factoid model

To determine the best-performer in each three cases, we used the average results over five test-batches of the Bio-ASQ 6B challenge [6]. All three models are obtained from the same training experiment, and correspond to the checkpoints of the same model instance.

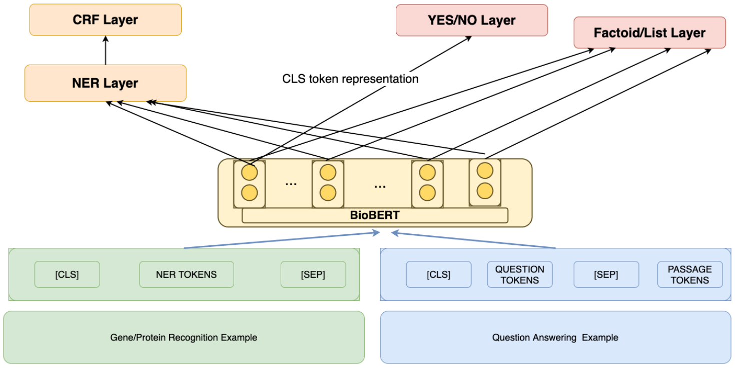


Fig. 3: The proposed multi-task learning model which learns all biomedical question types and the biomedical entities simultaneously.

2.4 Multi-task Learning Model

The multi-task learning model further extends the joint-learner explained in Section 2.3. In this setting, a single neural model is trained for Biomedical Question Answering and Gene/Protein Entity Recognition tasks, simultaneously. The details of the Question Answering component of the model is identical with the joint-learner. In addition, the model contains an entity recognition component consisting of a Fully-Connected layer, followed by a Conditional Random Fields (CRF) layer. CRF-based models are frequently used for the named entity recognition task, to take into account the tag transitions between consecutive tokens [8, 1]. For this reason, we extended the NER-component of the previously proposed BioBERT-based NER model in [9] to include an additional CRF layer. The overall architecture of this proposed multi-task learner is shown in Figure 3. For a sequence of n tokens $t_i : 1 \leq i \leq n$, the NER-component receives the BioBERT representation for each token. The subword token representations are then averaged to get the word-level representations. These word-level representations are fed into the FC-layer to generate the scores for each entity label, for each token. The CRF-layer generates the final score for each label by taking into account the transitions between each label. For the NER component, crf-loss is used. The loss is calculated as the difference between the total score of all possible label-sequences (all possible paths) and the score of the gold-label sequence (gold-label path):

$$
\begin{aligned}
b_j &= BioBERT(t_1, ..., t_i, ...t_n; j) \\
s_j &= FC^{ner}(b_j) \\
\mathbf{S} &= [s_1, ..., s_j, ..., s_n] \\
crf\_loss &= forward\_score(\mathbf{S}, \mathbf{T}) - path\_score(\mathbf{S}, \mathbf{T}, G)
\end{aligned}
$$

where $\mathbf{S}$ denotes the scoring matrix containing scores for each label and word pair, G is the gold-label sequence, and $\mathbf{T}$ is the transition matrix containing transition scores between each label. $forward\_score(\mathbf{S}, \mathbf{T})$ denotes the total score of all paths and $path\_score(\mathbf{S}, \mathbf{T}, G)$ is the score of the gold label sequence. Ideally, we want all probabilities to accumulate on the gold-label path so that these two scores will be identical.

Inference During the inference mode, we used Viterbi decoding [4] to find the highest scoring label sequence for the entity recognition task.

2.5 Post-processing

As explained in the pre-processing section, we divided each question with multiple gold-label snippets into separate inputs. These examples are merged during post-processing to generate a unique answer for each question. For the post-processing step, we followed [17] to combine the predictions to the same question for factoid/list type questions. Majority voting is used to find the highest scoring predictions for each factoid/list type question. For each factoid type question, top N highest scoring predictions are returned where N corresponds to the maximum limit allowed for the BioASQ 8B challenge. For the list type questions, we used 0.50 as the probability threshold, and included all answers that have a higher average probability score.

For the yes/no type questions, we averaged the probability scores for each example belonging to the same question instance to determine the final answer.

3 Experimental Settings

In this section we explain details regarding the experiments we conducted. All experiments are done using a single V100-GPU. For the Question Answering task we used the BioASQ 6B test sets as our validation set, and used the examples in the BioASQ 8B training set, for training. For the entity recognition task, we kept the same train/dev/test split already provided in [9]. It takes around 4-5 epochs on the training set to achieve the highest performance on the question answering validation sets for all models.

3.1 Datasets

The entity recognition component of the final multi-task learning model we used for our submissions is trained on the BC2GM dataset [15]. The dataset contains 20,703 entity mentions in total and annotated using BIO scheme. The first token of each entity is annotated with 'B' and the following tokens are annotated with 'I'. Non-entity tokens are annotated with 'O'.

Table 2: Statistics about the datasets used for the biomedical entity recognition task.

Dataset	Split	# of Entity Tokens	# of Entities
BC2GM	Train	37,301	15,197
	Development	7,498	3,061
	Test	15,101	6,325

In order to evaluate our proposed multi-task learner, we trained the entity recognition component on three different datasets. We used the BC2GM [15], BC4CHEMD [7], and BC5CDR [10] datasets for biomedical entity recognition which contain gene entities, chemical entities and disease mentions respectively. As we had maximum submission limit of five submissions for each test-batch for the BioASQ 8B challenge, we only used the multi-task learning model trained on the BC2GM dataset.

The BioASQ 8B training set contains 3,243 questions in total. We did not make use of the 777 summary type questions, so our overall training set contained 2466 questions. For training our models we used only the snippets already provided by the challenge organizers as the relevant passage for each question. Each snippet and question is treated as a unique (Q, P) pair which is given as input to the question answering component, where Q and P represent 'question' and 'passage', respectively.

Table 3: Number of questions in the BioASQ 8B dataset. Summary type questions are not used in our work.

Dataset	Summary	Factoid	Yes-No	List	Total
BioAsq-8B	777	941	644	881	3,243
BioAsq-8B pre-processed	-	4,916	9,786	9,472	24,174

For evaluating our proposed models, we also used the factoid questions from the BioASQ 6B test set [6]. The test set contains five-batches, and the number of factoid questions for each batch are given in Table 4.

Table 4: Number of factoid questions in each test batch of the BioASQ 6B challenge used for evaluating the multi-task learning model.

Batch	Number of Factoid Questions
batch-1	31
batch-2	21
batch-3	32
batch-4	33
batch-5	44

3.2 Training

In this section, we explain how we trained each of the three models we used to make submissions for the BioASQ 8B challenge. In all three models, we initialized the weights of the BERT component using the BioBERT version 1.1 provided by Lee et al. [9] pretrained on PubMed articles. To have a fair comparison we always used a maximum sequence length of 256, as we observed that going above this value sometimes resulted in memory issues. Table 5 gives a comprehensive list of the hyperparameters we used during our experiments.

Table 5: Important hyperparameters used by the models we proposed.

Name	Final Value	Range
Maximum Sequence Length	256	[64-512]
NER learning rate	0.0015	[0.001-0.1]
QAS learning rate	$5e^{-4}$	[$5e^{-4}$ - e^{-2}]
QAS Adam Epsilon	e^{-8}	[e^{-10} - e^{-5}]
NER Adam Epsilon	e^{-6}	[e^{-10} - e^{-4}]
Optimizer	AdamW	Adam, AdamW, SGD
Weight Decay	0.001	[0 - 0.01]

Baseline model training The baseline model (BioBERT_baseline) is composed of two completely separate neural architectures (one for yes/no and one for factoid/list type questions). In this approach, each architecture is trained separately, only using the corresponding dataset. During pre-processing, list type questions are converted into factoid question format, by treating each answer in the list of answers as a single factoid type answer. After this pre-processing step, the format of the factoid and list type questions become identical, so that the same architecture can be used for training on both types.

Joint-learning model training The joint learner (BioBERT_allquestions) is trained on all question types at once. At each iteration a (Q, P) pair is picked randomly from the whole training set. If the picked example is a 'yes/no' type question the 'Yes/No' component in Figure 2 is used to generate the output of the model. Otherwise, the 'Factoid/List' component is used to generate the 'start' and 'end' scores for each token inside the given passage P. The loss for each input example is backpropagated to update the weights of 1) the question-type specific component, and 2) the common BioBERT component. This way, we allow information transfer between different question types. Considering the relatively small sizes of the biomedical question answering datasets, this allows a better utilization of what is available. Besides, this approach reduces the total number of parameters of the final model almost by half, as the majority of the trainable parameters are the common BioBERT weights. As we have multiple target performance metrics (overall performance and performances on each question type), we continued the training until we could not observe any improvement for any question type on the question answering validation sets.

Multi-task learning model training The multi-task learning model (BioBERT_multitask) is simultaneously trained for the question answering and the entity recognition tasks. At each iteration we flip a random coin to determine the task type (QAS or NER), and use the corresponding component from Figure 3. Similar to the joint-learning model this allows information transfer from the NER dataset examples for the question answering task. The common BioBERT model is updated using examples from both tasks, which allows us to expose the model for a significantly larger amount of sentences from the biomedical domain. In this work, entity recognition task is used as an auxiliary task to help improve the final performance on the target question answering task. For this reason, training is done until we could not observe any improvement on the question answering validation set.

4 Results

In this section we start by giving the results we obtained for evaluating our proposed multi-task learner. We compare BioBERT_multitask, which learns both entity recognition and question answering tasks simultaneously, with the joint-learning model BioBERT_allquestions, which only focuses on the question answering task. The BioASQ 8B data is used to train both models, and the factoid type questions from the BioASQ 6B challenge is used to evaluate them, which contains five different test batches. For training the entity recognition component of the multi-task learning model, we used three different biomedical entity datasets. The results for both models are given in Table 6. Our results showed that learning both tasks simultaneously improved the performance for **all entity datasets** and for **all test batches**. For all three datasets we observed that the multi-task learning model outperformed the model that only learns the question answering task on all five test-batches. These results verified our initial claim on transfering information from entity recognition task to improve the performance on the target question answering task, and motivated us to apply the proposed multi-task learning model on the BioASQ 8B test sets.

Table 6: Analysis of multi-task learning of QA and ER using different ER datasets. Results are given for each test batch of the BioASQ6 Challenge. The official metric used for evaluating models is MRR. Best results for each test-batch are denoted in bold.

		BioASQ-6		
Model	Test batch	SAcc	LAcc	**MRR**
QA Only (Baseline)	batch 1	0.581	0.742	0.646
	batch 2	0.667	0.857	0.74
	batch 3	0.594	0.781	0.682
	batch 4	0.545	0.636	0.586
	batch 5	0.455	0.523	0.485
QA + BC2GM	batch 1	0.581	0.71	0.645
	batch 2	0.714	0.857	**0.786**
	batch 3	0.625	0.781	0.698
	batch 4	0.576	0.636	0.606
	batch 5	0.477	0.523	0.5
QA + BC4CHEMD	batch 1	0.71	0.742	**0.72**
	batch 2	0.714	0.857	0.778
	batch 3	0.656	0.781	0.709
	batch 4	0.606	0.636	**0.616**
	batch 5	0.5	0.523	**0.511**
QA + BC5CDR	batch 1	0.677	0.742	0.7
	batch 2	0.714	0.857	**0.786**
	batch 3	0.656	0.781	**0.719**
	batch 4	0.576	0.636	0.601
	batch 5	0.5	0.523	**0.511**

Next, we give the results obtained on the BioASQ 8B challenge for each model we explained above. For the first test-batch we only made submissions using two models: BioBERT_baseline, BioBERT_allquestions. For the other four test-batches we made five submissions using the three models explained above. To be able to make a clean comparison between the proposed models, we kept the post-processings schemes identical for all our submissions. This is necessary to evaluate our claim on using multi-task learning to improve the performance on biomedical question answering task.

The QAS components of the joint-learning model and the model-task learning model are identical. In order to evaluate our claim on using multi-task learning for question answering, we must compare these models, rather than comparing them with the single-task learning model which uses a different architecture (separate models for each question type). The results show that for the factoid questions, the multi-task learning based model outperformed all three joint-learning models for all four test-batches. This clearly shows that leveraging information obtained about genes and proteins may help improve the final performance on the factoid type questions. The results for list and yes/no type questions are mixed, and the benefits of multi-task learning are unclear for these types.

Table 7: Results for all submitted models on each test batch of BioASQ 8B Question Answering Challenge.

			Factoid			**Yes/No**		**List**		
Model type	**Name**	Test batch	SAcc	LAcc	MRR	Accuracy	Macro-F1	Precision	Recall	F-1
Single-task Learning	BioBERT_baseline	batch 1	0.3125	0.5938	0.4266	0.6800	0.4048	0.3461	0.2933	0.2918
		batch 2	0.1600	0.4400	0.2580	0.7500	0.4286	0.5361	0.4476	0.4306
		batch 3	0.3214	0.3929	0.3512	0.5806	0.3673	0.4861	0.4056	0.4214
		batch 4	0.4118	0.7059	0.5270	0.5000	0.3333	0.3296	0.3810	0.3161
		batch 5	0.5000	0.6875	0.5844	0.6176	0.6007	0.2242	0.2173	0.2179
Joint-learning	BioBERT_allquestions (overall best)	batch 1	0.2813	0.5938	0.4099	0.6800	0.4048	0.3842	0.3200	0.3262
		batch 2	0.1600	0.4800	0.2540	0.7778	0.5355	0.4107	0.3738	0.3712
		batch 3	0.3214	0.3929	0.3423	0.5806	0.3673	0.5972	0.4111	0.4290
		batch 4	0.4412	0.7059	0.5564	0.5000	0.3333	0.4045	0.4623	0.3886
		batch 5	0.4063	0.6875	0.5063	0.5588	0.3585	0.3646	0.3333	0.3347
	BioBERT_allquestions (best factoid)	batch 1	-	-	-	-	-	-	-	-
		batch 2	0.1200	0.4400	0.2413	0.7500	0.4286	0.4827	0.4071	0.3950
		batch 3	0.2857	0.3929	0.3333	0.5806	0.4732	0.5208	0.4056	0.4107
		batch 4	0.4706	0.7059	0.5564	0.5000	0.3333	0.4582	0.4153	0.4005
		batch 5	0.4063	0.6563	0.5026	0.5588	0.3585	0.4965	0.4167	0.4308
	BioBERT_allquestions (best yesno)	batch 1	-	-	-	-	-	-	-	-
		batch 2	0.1600	0.4400	0.2580	0.7500	0.4286	0.5361	0.4476	0.4306
		batch 3	0.3214	0.4286	0.3643	0.5161	0.3404	0.5139	0.3556	0.3721
		batch 4	0.4412	0.7059	0.5564	0.5000	0.3333	0.4045	0.4623	0.3886
		batch 5	0.4063	0.6875	0.5063	0.5588	0.3585	0.3646	0.3333	0.3347
Multi-task learning	BioBERT_multitask	batch 1	-	-	-	-	-	-	-	-
		batch 2	0.2000	0.4000	0.2800	0.8333	0.7000	0.4643	0.4214	0.4108
		batch 3	0.3214	0.4286	0.3643	0.5161	0.3404	0.5139	0.3556	0.3721
		batch 4	0.4706	0.6765	0.5637	0.4615	0.3158	0.3843	0.3226	0.2991
		batch 5	0.4063	0.7188	0.5365	0.5000	0.3333	0.5729	0.4236	0.4667

5 Conclusion

In this paper we described the models we used to make submissions for the BioASQ 8B challenge. We proposed a novel multi-task learning model for biomedical entity recognition and question answering tasks. Our results showed that transferring information from the entity recognition task consistently improved the performance on the factoid type questions of the question answering tasks. On all test-batches of both BioASQ 6B and BioASQ 8B challenges, transferring information brought improvement for factoid questions. We believe that further improvements can be achieved by implementing a more sophisticating information sharing between the two tasks. Analyzing the characteristics of each dataset used, can help us understand why transfer learning improves/degrades the performance for each question type.

So far we have only considered using **domain-adaptive** pretrained models (BioBERT-based). Recent work on pretraining showed that **task-adaptive** pretraining brings additional improvement for low-resource tasks [5]. Our plan is to incorporate **task-adaptive** pretraining for the biomedical question answering task.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: COLING 2018, 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018)
2. Akdemir, A.: Research on task discovery for transfer learning in deep neural networks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 33–41 (2020)
3. 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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
4. Forney, G.D.: The viterbi algorithm. Proceedings of the IEEE **61**(3), 268–278 (1973)
5. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964 (2020)
6. Kakadiaris, I.A., Paliouras, G., Krithara, A. (eds.): Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering. Association for Computational Linguistics, Brussels, Belgium (Nov 2018), `https://www.aclweb.org/anthology/W18-5300`
7. Krallinger, M., Rabal, O., Akhondi, S.A., et al.: Overview of the BioCreative VI chemical-protein interaction Track. In: Proceedings of the sixth BioCreative challenge evaluation workshop. vol. 1, pp. 141–146 (2017)
8. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
9. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (09 2019). https://doi.org/10.1093/bioinformatics/btz682, `https://doi.org/10.1093/bioinformatics/btz682`
10. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database **2016** (2016)
11. 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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237 (2018)
12. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100, 000+ questions for machine comprehension of text. In: EMNLP (2016)
13. Reddy, S., Chen, D., Manning, C.D.: Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics **7**, 249–266 (2019)
14. Ruder, S.: Neural Transfer Learning for Natural Language Processing. Ph.D. thesis, National University of Ireland, Galway (2019)
15. Smith, L., Tanabe, L.K., nee Ando, R.J., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C.M., Ganchev, K., et al.: Overview of biocreative ii gene mention recognition. Genome biology **9**(S2), S2 (2008)

16. Wu, S., Dredze, M.: Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. 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. 833–844 (2019)
17. Yoon, W., Lee, J., Kim, D., Jeong, M., Kang, J.: Pre-trained language model for biomedical question answering. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 727–740. Springer (2019)