Building Question and Answering Systems for COVID-19 Medical Research Papers

Author
Pablo Marino pablo-n-m@hotmail.com
Catherine Wang cwang42@outlook.com
Vishay Vanjani vishay.vanjani@gmail.com

Abstract

In the light of the COVID-19 Pandemic, numerous medical research papers have been published in the past fewer months presenting the latest medical advancements in the epidemiology. This paper represents a way to build a model, called COVID-BERT, that can understand the content of Coronavirus papers and perform reading comprehension to answers open questions. The model was built using the concept of transfer learning and uses SciBERT(Beltagy et al., 2019) as its base model. Several experiments have been done to create and improve the model, including generating COVID-19 related vocabulary, training a new medical domain tokenizer, pre-training on COVID-19 corpus, and fine-tuning on both SQuAD2.0 and BioASQ. The results of the model have been evaluated in a quantitative and qualitative way. Future work could be done to improve the system by creating humans in the loop for user feedback.

1 Introduction

Recent advances in language modeling have led to substantial gains in the field of natural language processing, with state-of-the-art models such as BERT(Jacob Devlin, 2019), RoBERTa(Yinhan Liu, 2019), XLNet(Zhilin Yang, 2019), and ALBERT(Zhenzhong Lan, 2019), among many others.

Transfer learning is one of the predominant idea in Deep Learning, it assume that general-purpose deep neural network would be able to apply it's previous gained knowledge to similar downstream problems. Based on this archetype, large pre-trained models have emerged. Those models have been trained with a large amount of unlabeled general-purpose corpus to build the best-distributed representation of word vectors, then fine-tuned on specific NLP tasks.

SciBERT(Beltagy et al., 2019) is one step ahead of BERT since it has been pre-trained on scientific papers to build up the vocabulary and scientific embeddings. This gives a considerable improvement on a few NLP tasks in the biomedical domain (including NER, PICO, REL, etc.).

We hypothesize that a good QnA system for Covid-19 dataset(AI2, 2020) needs to be pre-trained on large amount of biomedical scientific text and fine-tuned on medical domain QnA dataset. In this paper, we pre-train SciBERT on covid-19 corpus and then fine-tune it on the BioAsQ(Tsatsaronis et al., 2015) dataset. In addition we compare this model with original SciBERT and BioBERT(Lee et al., 2019) models fine-tuned on BioASQ dataset. Finally, we apply our model to Covid-19 dataset and use it answer the common questions relating to the covid pandemic.

2 Related Work

Global pandemics like Covid-19 highlight the need to use AI for extracting knowledge from the exponentially growing amount of text at scale. An automated QnA system can help spread awareness by answering common questions on transmission and symptoms of covid-19.Bert(Jacob Devlin, 2019) showed that Nlp models can be used to extract information from large-scale, general-domain texts and reach human-level accuracy on most nlp tasks. However, results in (Lee et al., 2019) and (Beltagy et al., 2019) showed that BERT does not perform well on domain specific tasks like QnA on medical texts. BioBert was the first domain specific BERT based model trained on biomedical corpus. Their model was pre-trained on 17 billion tokens as opposed to 3.3 billion tokens for original Bert. The BioBert team chose to use the same vocabulary as BERT. SciBERT was trained on around 3.17 billion tokens from

1.14 papers (20% from computer science and 80% from biomedical domain). SciBERT authors constructed their own vocabulary using the most frequently occuring words in their Scientific corpus.

3 Data

Finding the precise and accurate data source will lay a solid foundation for building the system. There are two main types of data for this paper with different objectives. The first one is for pretraining the language model. The second one is for fine-tuning on question and answer tasks.

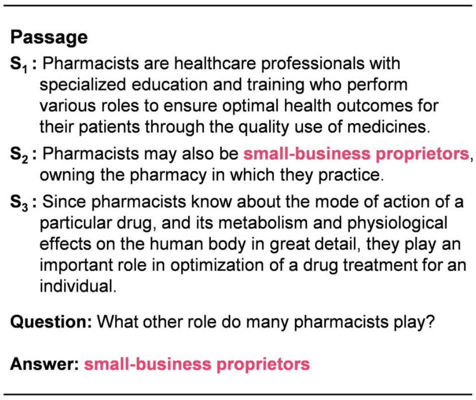
To fit the purpose of the first type, the data source needs to contain a substantial amount of domain-related data in a flat-file format. The sufficient amount will guarantee our training would have enough samples coverage to build new tokenizer and new weighing for the BERT like language model.

The second type data source will formulate the downstream tasks, in this case question and answering, into machine comprehensible format. Then the model will learn from the pattern it provides to generate a new layer of logic.

3.1 SQuAD 2.0

Stanford Question Answering Dataset (SQuAD)(Stanford University, 2018) is a reading comprehension dataset, containing questions and answers pairs from Wikipedia articles posed by crowd-sourcing. The answer to a question is formulated in two ways. The first one, as a segment of the original text. Second is a san of the start and end index from the corresponding passage. Some of the questions might not contain any answers and will be marked as unanswerable.

The version we are using in this project is **SQuAD2.0**. According to (Pranav Rajpurkar, 2018) This dataset combines the original 100,000 questions and answers pairs from SQuAD1.1 and additional 50,000 unanswered questions. The unanswerable questions designed adversarially to look similar to the answerable ones. Thus model learns from this dataset will be robust to distinguish similar questions with different intentions and utterances. An example of SQuAD2.0(Stanford University, 2018). dataset can be found in the Table-1



Passage

S_1 : Pharmacists are healthcare professionals with specialized education and training who perform various roles to ensure optimal health outcomes for their patients through the quality use of medicines.

S_2 : Pharmacists may also be **small-business proprietors**, owning the pharmacy in which they practice.

S_3 : Since pharmacists know about the mode of action of a particular drug, and its metabolism and physiological effects on the human body in great detail, they play an important role in optimization of a drug treatment for an individual.

Question: What other role do many pharmacists play?

Answer: **small-business proprietors**

Figure 1: An example of SQuAD2

3.2 BioASQ 8

BioASQ organizes challenges every year on biomedical semantic indexing and question answering (QnA). The challenges include tasks relevant to hierarchical text classification, machine learning, information retrieval, QnA from texts and structured data, multi-document summarization and many other areas.**Task B** of the challenge involves Semantic QnA and they release the corresponding train and test sets every year. We used the **BioASQ 8** dataset, which has 3,243 questions in train set and 500 questions in test set. Each question has multiple answers, contained in a Pubmed abstract or title. The dataset provides the answer offsets, url links to Pubmed articles and other meta-data. We converted the BioASQ format to **Squad V2** format for ease of processing and generated multiple question answer pairs for each question. As a result we ended up with 28,555 and 1,807 questions-answer pairs for train and test set respectively. We ignored question answer pairs where the context was the article title and where the dataset provided wrong offsets (including these would have made dataset sizes greater than 50,000 and 2,500 for train and test sets respectively). Also, we did not evaluate the different types of questions separately (factoid, summary, list, yesno). We leave these two improvements for future work.

3.3 CORD-19: COVID-19 Open Research Dataset

In response to the COVID-19 pandemic, Kaggle, the biggest data science community, is hosting a competition to develop the AI solution that can help the medical community to answer some of the high priority scientific questions. A coalition of leading research groups and governments have prepared the COVID-19 Open Research Dataset (CORD-19)(AI2, 2020) . And the Kaggle team presents a comprehensive CORD-19 dataset with machine-readable coronavirus literature collection from various resources, including bioRxiv, medRxiv, arXiv, and others. The corpus is updated regularly and, at the time of this report, has more than 59,000 peer-reviewed publications.

This project uses the CORD-19(AI2, 2020) as ground truth to re-train BERT for a new language model that could comprehend COVID-19 related questions.

The original CORD-19 data is in PDF and PMC format, in order to make it compatible for BERT training, the data need to be pre-processed. According to (Jacob Devlin, 2019), BERT uses a "masked language model" (MLM) pre-training objective. To train the model, the input corpus need to randomly masks some tokens and let the model predict the original vocabulary based on the context. Hence the team apply the following process to prepare the training input.

- Convert PDF and PMC documents into raw text.
- Uncase the raw text and remove special characters.
- Sentences the raw text and create a sliding window of 512 tokens per line.
- Add special token [MASK] for random masks, [SEP] for sentence separator, [PAD] for padding the sentences. [CLS] for classification and [UNK] for out of vocabulary unknown token.

4 Model

Several pre-train models were chosen to co-create the COVID-BERT. The strategy was built on the concept of transfer learning and language representation learning. Hence we selected a BERT like model that trained on medical documents called Scibert(Beltagy et al., 2019) as the base model and ran a full network pre-training on the CORD-19 dataset(AI2, 2020). After pre-training, the intermediate model will be further tuned on SQuAD2.0(Stanford University, 2018) and bioASQ(Tsatsaronis et al., 2015) to learn the downstream questions and answers tasks.

4.1 Baseline Models

For Base lining we chose a range of BERT based models, since they have proven to be very effective at question answering and share the same architecture as the model we're presenting in this paper. even though they all share the same architecture, were pre-trained and fine tuned using different datasets and that's what causes the difference in performance.

- uncased Base BERT : base uncased BERT model. pretrained on Wikipedia articles and fine tuned on SQUAD 2
- uncased base SciBERT: BERT based model pretrained on full scientific papers from semanticscholar.org.and finetuned on SQUAD 2
- base BioBERT: BERT model pretrained on biomedical articles and fine tuned on SQUAD

Adding to the models listed above we also further fine tuned SciBERT and BioBERT on BioASQ 8, and used them both for a qualitative evaluation and comparison against COVID-BERT.

We obtained quantitative metrics on BioASQ 8 and SQUAD 2 for all our baseline models and compared them with the ones obtained from our COVID-BERT model.

4.2 Pre-training Language Models

Masked Language Model and Next Sentence predictions are two main objects in BERT(Jacob Devlin, 2019). In MLM, 10-12% of input tokens are replaced with [MASK], and around 1.5-2% are replaced with random vocabulary from the original corpus. NSP's objective is to train the model to remember the relationship of two given sequence. For example, given a sentence A, what is the most like sentence B directly follow A. However based on the ALBERT(Zhenzhong Lan, 2019) and XLNet(Zhilin Yang, 2019) paper, NSP's contribution to the model performance is questionable. Therefore, in this pre-training experiment, *['cls.predictions.bias', 'cls.predictions.transform.dense.weight', 'cls.predictions.transform.dense.bias']* weightings are not used.

Another strategy that applied in the pre-training process is determining the batch size and learning rate. Training with large mini-batches can increase the optimization speed, and tuning the learning speed will guarantee the model to converge quicker.

Due to the computational and time restrictions, the final COVID-BERT model was trained using the following parameters,

- `architectures = "BertForMaskedLM"`
- `vocab_size=31116`
- `max_sequence_len = 512`
- `train_batch_size = 16`
- `learning_rate = 1e-4`

Simple rationals behind the hyperparameter tuning are larger mini-batches increase the perplexity of MLM tasks, likewise, it is easier to parallelize via distributed data-parallel training. Another technique also allows large scale parallelization is gradient accumulation. This method has also been implemented int the pretraining process.

For the final training we used weight from Bert For Masked LM

4.3 Fine-tuned Models

The COVID-BERT models were first fine-tuned on the **Squad v2** and then on **BioASQ** training sets. We made minor modifications to ”run squad.py” provided in hugging face library. We used [batch size=8, learning rate=3e-5, epochs=3.0, max seq length=384, doc stride=128, max answer length=128] . Note the doc stride hyper parameter ensures that the context is split into smaller chunks if the length of the sequence (question + context + meta tokens=3 ([sep],[sep] and [cls]) is greater than max sequence length (384), for example when the size of the sequence is greater than 384 the script breaks the context into smaller chunks using sliding window and sets ”is impossible” to false, if the answer lies within that chunk and true otherwise.Also note that we increased the value of max answer length from 30 (default) to 128 because BioASQ (Tsatsaronis et al., 2015)answers are on average bigger than Squad answers. We fine-tune Scibert pre-trained on covid-dataset, original Scibert(Beltagy et al., 2019) and BioBert(Lee et al., 2019).

5 Metrics

The team will be using the universally accepted machine comprehension and QA evaluation metrics to quantify the model performance and will provide comparison with the original SciBERT and BioBERT. That evaluation will include the following two scores:

- Exact Match (EM) score, which presents the number of answers that match the ground truth answers exactly (with the same start and end index of the answer span)
- F1 score, which captures the harmonic mean of precision and recall, of the sequence predicted by the model w.r.t the ground truth answer

For the task of evaluating model answers to covid-19 related questions from covid-19 dataset, the team will self evaluate the extracted answers using a manual review process. Note the BioASQ challenge uses different metrics for its leader-board. Due to time constraints, we chose not to use those metrics.

6 General Approach

As mentioned before in the section - **4 Model**. The design principles of building a workflow for COVID-BERT are pre-train on the domain-specific corpus and then fine-tune on downstream QA tasks.

In the section - **3.3 CORD-19**, we discussed the approach to use CORD-19 to build the pre-training corpus. Then the pre-training corpus is used to do two tasks: 1) training a new tokenizer and creating a new vocabulary for COVID-BERT, 2)use new tokenizer, vocabulary together with the preprocessed the training data (derived from CORD-19 data) to pre-train original BERT model(Jacob Devlin, 2019).

For the pre-training and fine-tuning process, the team uses the Hugging Face Transformer API (Face, 2018).

In the pre-training step, we initialize COVID-BERT on using weights from both *’allenai/scibert_scivocab_uncased’* and *’google/bert_uncased_L-12_H-768_A-12’* . The initial result from both models shows that using the google base BERT model yields a higher F1 on SQuAD 2.0 evaluation. The reason could be **weighing from Large BERT(L=12 , H=768**

) was more general in terms of understanding natural English comparing to medical specific SciBERT(Beltagy et al., 2019). Since the only difference in this step is initial weighing, we decided to use Large BERT weighing to re-train our language model for the practical reason.

Then, we choose to build a new vocabulary from combining SciBERT(Beltagy et al., 2019) 'vocab' and COVD-19(AI2, 2020) 'vocab' to make our tokenization more robust.

To sum up, the final version COVID-BERT was initialize using weights from LARGE BERT, then combine new medical domain vocabulary and new tokenizer, to pre-trained on COVD-19 corpora (57,000 research paper, 4.3 GB training set). To show the effectiveness and robustness of our approach in biomedical COVID-19 text mining. The hyperparameters we choose for the final model was discussed in the secion - **4.2 Pre-training Language Models** .

The final stage is to teach COVID-BERT how to perform machine comprehension tasks. As discussed in the section -**3 Data**, there are two primary fine-tune sets we choose, **3.1 SQuAD 2.0** and **3.2 BioASQ 8**. Stanford creates the general-purpose QA set SQuAD 2.0, by fine-tuning on this set, the COVID-BERT will learn how to answer queries in natural language. Then we further fine-tune the COVID-BERT on the second QA set BioASQ 8 to understand the complexity of the question and answer patterns in fine-grained medical-related questions.

7 Results

7.1 Qualitative Result

The following questions, each with a respective context taken from a medical research paper abstract, was manually evaluated by our team to draw conclusions as of whether our COVID-BERT model shows an improvement in performance against our baseline models:

1. What are COVID-19 symptoms?
2. What is know about COVID-19 transmission?
3. What is the incubation time?
4. What is the COVID-19 persistence of infectious on surfaces of different materials?

Models used on the evaluation:

- COVID-BERT
- SciBERT finetuned on bioASQ
- BioBERT finetuned on bioASQ

Gold answers for above questions(selected by us, based on our common sense):

1. "There have been several reports noting anosmia and ageusia as possible symptoms of COVID-19"
2. "high temperature and high humidity significantly reduce the transmission of COVID-19"
3. "transmissions have been described with incubation times between 2-10 days"
4. "What is the COVID-19 persistence of infectious on surfaces of different materials?"

Answers produced by the models:
For Question 1:

- COVID-BERT: "anosmia or ageusia as symptoms ofCOVID-19"
- SciBERT: "There have been several reports noting anosmia and ageusia"
- BioBERT: "There have been several reports noting anosmia and ageusia"

For Question 2:

- COVID-BERT: "transmission of COVID-19."
- SciBERT: "It indicates that the arrivalof summer and rainy season in the northern hemisphere"
- BioBERT: "effectivelyreduce the transmission of the COVID-19."

For Question 3:

- COVID-BERT: "transmissions have been described with incubation times between 210 days,"
- SciBERT: "transmissions have been described with incubation times between 2-10 days,"
- BioBERT: "incubation times between 2-10 days,"

For Question 4:

- COVID-BERT: ”severe respiratorytract infections in humans.”
- SciBERT: ”minute.”
- BioBERT: ”The”

To read the full contexts please refer to this jupyter notebook: https://github.com/pablonm3/cs224u/blob/master/QA_hugging_face_baseline.ipynb

We found not enough qualitative evidence suggesting that our model outperforms the baselines.

7.2 Quantitative Results

Quantitative results are generated by evaluating on bioASQ Metrics can be found in Table-1.

We found no evidence that suggests that our COVID-BERT model is better than the baseline models at the bioASQ task.

8 Discussion and Analysis

As Table 1 shows, fine-tuning on a domain specific dataset improved the metrics. The results show that constructing a new vocabulary along with pre-training on large amount of domain specific text helps. Covid-BERT (the model that we pre-trained) did not perform as well as expected. There are several possible reasons for this which we plan to explore in future work

- Not enough data - We need to pre-train on more than 57,000 research papers to see better results
- Tuning hyperParameters - We did not get a chance to fine-tune the hyper-parameters for pre-training
- Issues with hugging face api - The pre-train scripts provided by hugging face library need to be optimized for use with SciBert

The best model that we fine-tuned gets a F1 score of 65.434% which is a lot lower than the performance of Bert based models in general domain. We think this is because QnA in medical domain has much longer answers and we fine-tuned the model with much shorter answer length (max ans length = 128 and max sequence length = 384). In our qualitative analysis we found the same issues the model found the start in some cases and the end in others but was not able to capture the whole answer. In future work we plan to use longer max answer length and max sequence length for fine-tuning

9 Conclusion and Future Work

In this report, we create a methodology to train a new model COVID-BERT. This model is based on the concept of transfer learning and leveraging state-of-the-art language model BERT(Jacob Devlin, 2019) and SciBERT(Beltagy et al., 2019) to initialize, then use various open-source data set to pre-train and fine-tune.

Based on the quantitative and qualitative results we collect, the team can see the improvement in answering COVID-19 related questions and comprehend the content. In addition, the COVID-BERT performs relatively well on general QA tasks, evaluated in SQuAD 2.0(Stanford University, 2018).

To further improve on COVID-BERT, more granular hyperparameters tuning could be done with more substantial computational resources. Also, a fine-grained vocabulary that can be further trained to generalize between natural English and medical/COVID specific language. Apart from building a more robust model, creating a brand new QA training and evaluation for COVID-19 would also help to feed more precise patterns into the existing model.

The team would also want to try to extend the standalone model into a full system with an interactive front end. Thus end users can be engaged in the process of testing and evaluating the model result by asking questions to the system and reviewing the answers generated by COVID-BERT.

A Appendices

Link to github repo

B Supplemental Material

Code Repo https://github.com/pablonm3/cs224u

References

MSR Georgetown NIH The White House AI2, CZI. 2020. Covid-19 open research dataset challenge (cord-19).

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

Model	F1	EM
BioBert (finetuned on Squad2 and BioAsq)	63.719	40.322
Scibert (finetuned on Squad2 and BioAsq)	65.434	41.824
CovidBert (finetuned on Squad2 and BioAsq)	63.995	40.65

Table 1: Model Result Comparison

Hugging Face. 2018. Hugging face transformers.

Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computation and Language (cs.CL)*, arXiv:1806.03822. Version 2.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Percy Liang Pranav Rajpurkar, Robin Jia. 2018. Know what you don't know: Unanswerable questions for squad. *ACL 2018*, arXiv:1806.03822.

Crowdworkers Stanford University. 2018. The stanford question answering dataset, squad2.0.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer Veselin Stoyanov Yinhan Liu, Myle Ott. 2019. Roberta: A robustly optimized bert pretraining approach. *Computation and Language (cs.CL)*, arXiv:1907.11692.

Sebastian Goodman Kevin Gimpel Piyush Sharma Radu Soricut Zhenzhong Lan, Mingda Chen. 2019. Albert: A lite bert for self-supervised learning of language representations. *Computation and Language (cs.CL)*, arXiv:1909.11942.

Yiming Yang Jaime Carbonell Ruslan Salakhutdinov Quoc V. Le Zhilin Yang, Zihang Dai. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Computation and Language (cs.CL); Machine Learning (cs.LG)*, arXiv:1906.08237.