Find the Optimal Model for Covid-19 QA

Xiao Li
New York University
xl998@nyu.edu

Yanyan Xu
New York University
yx2193@nyu.edu

Zian Chen
New York University
zc674@nyu.edu

Yichen Liu
New York University
yl7043@nyu.edu

Abstract

With the spread of COVID-19 and intensively accumulated research papers, it is important for researchers and users to efficiently get their questions answered. We would like to find out what are the optimal question and answering (QA) models in this specific domain. Recently, BERT (Devlin et al., 2018) based models achieved the state-of-the-art results in multiple tasks including the QA task. And there are also domain specific BERT models existing, e.g., BioBert (Lee et al., 2019). Hence based on the ALBERT model (Lan et al., 2019), we build COVID-ALBERT, a COVID-19 specific model to explore whether it could help to improve the downstream QA task performance in this area. Finally, we compare the performance of 3 models: ALBERT, BioBert and COVID-ALBERT on our manually-labeled test Covid-19 QA dataset.

1 Introduction

The spread of COVID-19 has seriously influenced human health and social life. As more and more people are concerned about COVID-19, the relevant information is exploding on the Internet. Much of it, however, falls short of sufficient scientific argument, thus posing a significant risk of misleading the general public. Meanwhile, the intensively accumulated research papers about this pandemic has made it difficult and expensive for researchers and other people to identify answers for their queries in this domain. Thus, how to build a successful QA model related to COVID-19 becomes an urgent task in this unprecedented social circumstance.

Recently, BERT (Devlin et al., 2018), based on multi-layer Bidirectional Transformer, has changed traditional language models by predicting randomly masked tokens and sentence continuity and achieved higher performance on multiple downstream tasks including the QA task. ALBERT (Lan et al., 2019), by sharing the same parameters across layers, successfully reduces the model size and improves the performance on different downstream language tasks. There are also some domain specific BERT models, for example, BioBert (Lee et al., 2019) focuses on the biomedical field by training on PubMed abstracts and PMC full-text articles; CLINICBert(Alsentzer et al.) models continue to pretrain BERT and BioBERT on clinical text and discharge summaries to further the models' domain specificity. Hence we are curious to find out whether a pretrained model specifically focused on Covid-19 texts could improve the performance of downstream Covid-19 QA tasks. Therefore, in this work, we pretrain a COVID-ALBERT model and compare its fine-tuning performance with ALBERT and BioBERT on Covid-19 QA task. To fulfill this task, we also manually build a test Covid-19 QA dataset which will be discussed in detail at section 3.4. We also hope that our experiment could shed some light and provide us experience on how to build a successful QA model when facing a new domain of knowledge.

2 Related Work

Various BERT-related Models There are many BERT-related models out there. Some of them changed the design of BERT slightly to achieve better performances on downstream tasks. For example, ALBERT (Lan et al., 2019) uses a very similar backbone architecture as BERT while it also has some design choices different with BERT: ALBERT implements factorized embedding parameters, cross-layer parameter sharing and inter-sentence coherence loss. The first two design choices provide the advantage of fewer parameters than BERT which enables faster training speed and less memory usage. The inter-sentence coherence loss enables ALBERT to change the next-sentence prediction(NSP) task of BERT to a harder version - sentence-order prediction(SOP) which lets ALBERT achieve improvement on several downstream tasks. Our team decided to implement our pre-trained model on ALBERT because of the relatively faster training speed, less memory usage of ALBERT compared with other BERT-based models and the great performance of ALBERT on SQuAD datasets. Meanwhile, we used ALBERT to compare performance with other models on the test COVID-19 QA dataset.

There are also some domain specific BERT models. BioBERT (Lee et al., 2019) model has focused on the biomedical domain. It significantly improves the performance on complex biomedical text mining tasks including biomedical named entity recognition, biomedical relation extraction, and biomedical question answering. SciBERT (Beltagy et al., 2019), on the other hand, is targeting the large corpus of scientific texts. By frozen BERT embeddings, it has achieved state-of-the-art performance on different NLP tasks like long sentence classification and dependency parsing. ClinicalBERT (Huang et al., 2019) is a clinically-relevant BERT model, aiming to tackle unsupervised language modeling tasks, evaluated on a large corpus of clinical texts. It uses the sum of the last four hidden states of encoders as a representation of medical terms and achieved higher performance in interpretable prediction task, readmission prediction task with discharge summaries and early clinical notes compared to bag-of-words model and LSTM.

Pre-trained Language Model for Biomedical Question Answering (Wonjin Yoon, 2019) This paper utilizes BioASQ datasets to build a BioBERT-based QA system within large biomedical corpus for various question types. It achieved state-of-the-art performance on factoid, list and yes/no type questions tasks by first pre-training BioBERT on SQuAD and then fine-tuning on BioASQ datasets. Our work followed a similar methodology while using SQuAD and BioASQ datasets to finetune Albert.

3 Datasets

3.1 CORD-19

The dataset our team will be using is the kaggle CORD-19 dataset. It is part of the COVID-19 Open Research Dataset Challenge(CORD-19). This dataset is updated weekly. It contains an extensive collection of literature that's related to coronavirus, with over 44,000 scholarly articles, of which 29,000 with full text.

3.2 SQuAD(Rajpurkar et al., 2016)

Stanford Question Answering Dataset (SQuAD)is a reading comprehension dataset consisting of 100,000+ questions posed by crowd-workers on a set of Wikipedia articles. Basically each record comprises a question body, a context, an exact answer and the answer's starting position in the context. The answer types include date, number, named entity(i.e., location, person and other entities), adjective phrase, verb phrase and clause.

3.3 BioASQ(Tsatsaronis et al., 2015)

BioASQ is a challenge on large-scale biomedical data semantic indexing and question answering. It mainly consists of PubMed documents. It is the dataset BioBert used for to pretrain the BERT model. In the BioASQ dataset, factoid datasets' format is similar to SQuAD, so can be preprocessed to fit the original Albert fine tuning code. There are over 700 pairs of question answering in the train set and almost 200 pairs in the test set.

3.4 Manually Labeled COVID-19 Test QA Dataset

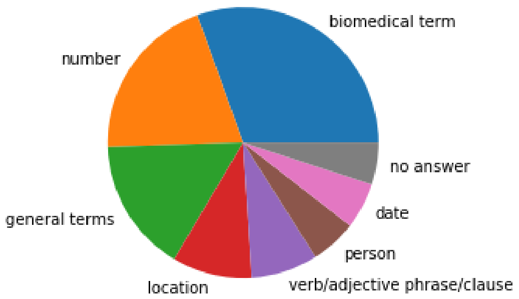
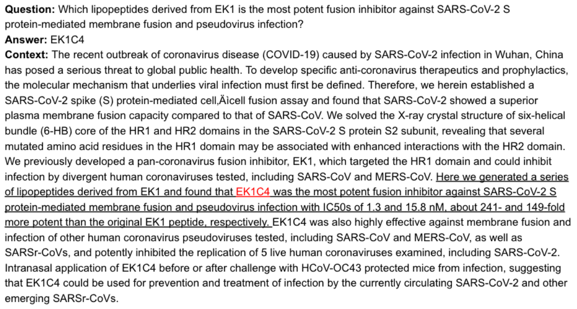


Figure 1: Question type of COVID-19 QA dataset

Since COVID-19 is a very new and small research area, there is no available specific QA dataset to test the performance of our model until now. So we manually create such a dataset following the same rule as SQuAD 2.0. We create 200 COVID-related QA pairs with the corresponding context. All Context are extracted from the abstract. 190 questions have answers, while 10 are unanswerable. The answer type structure is shown in the



Question: Which lipopeptides derived from EK1 is the most potent fusion inhibitor against SARS-CoV-2 S protein-mediated membrane fusion and pseudovirus infection?
Answer: EK1C4
Context: The recent outbreak of coronavirus disease (COVID-19) caused by SARS-CoV-2 infection in Wuhan, China has posed a serious threat to global public health. To develop specific anti-coronavirus therapeutics and prophylactics, the molecular mechanism that underlies viral infection must first be defined. Therefore, we herein established a SARS-CoV-2 spike (S) protein-mediated cell‚Äìcell fusion assay and found that SARS-CoV-2 showed a superior plasma membrane fusion capacity compared to that of SARS-CoV. We solved the X-ray crystal structure of six-helical bundle (6-HB) core of the HR1 and HR2 domains in the SARS-CoV-2 S protein S2 subunit, revealing that several mutated amino acid residues in the HR1 domain may be associated with enhanced interactions with the HR2 domain. We previously developed a pan-coronavirus fusion inhibitor, EK1, which targeted the HR1 domain and could inhibit infection by divergent human coronaviruses tested, including SARS-CoV and MERS-CoV. Here we generated a series of lipopeptides derived from EK1 and found that EK1C4 was the most potent fusion inhibitor against SARS-CoV-2 S protein-mediated membrane fusion and pseudovirus infection with IC50s of 1.3 and 15.8 nM, about 241- and 149-fold more potent than the original EK1 peptide, respectively. EK1C4 was also highly effective against membrane fusion and infection of other human coronavirus pseudoviruses tested, including SARS-CoV and MERS-CoV, as well as SARSr-CoVs, and potently inhibited the replication of 5 live human coronaviruses examined, including SARS-CoV-2. Intranasal application of EK1C4 before or after challenge with HCoV-OC43 protected mice from infection, suggesting that EK1C4 could be used for prevention and treatment of infection by the currently circulating SARS-CoV-2 and other emerging SARSr-CoVs.

Figure 2: An example of question answer pair with corresponding context

Figure 1 . From the pie chart, it is easy to notice that a lot of answers are biomedical terms that only virologists could possibly know. Number type and general terms type account for roughly 1/6 respectively. The vast majority of answers are noun or noun phrases, while 10% are verb phrase, adjective phrases or clause. An example is shown in Figure 2. The answer EK1C4, a rarely seen chemical, could be found in the underlined sentence in context. We wonder if our model can identify such technical entity in a long academic passage correctly.

4 Methods

4.1 Data Preprocessing

Since the CORD-19 dataset is saved in json files, we firstly extract all text sections from these json files and put them in a txt file based on the requirement input format for creating pre-training data in ALBERT. For validation, we modify the huggingface's transformer package to fit our csv-formatted QA testset.

4.2 Pre-training COVID-Albert

Since the corpus of CORD-19 after pre-processing is not as big as the Wikipedia and BookCorpus used by ALBERT and due to the rather limited resources we have (need to queue for NYU prince's gpu or pay to use Google Cloud TPU), we decide not to train our COVID-ALBERT from scratch. Instead, we choose to use ALBERT-base as our starting point and we then use the CORD-19 corpus to continue pre-training. Although in this way we could finish pre-training faster, the drawback is that we have to use the same vocabulary as ALBERT-base. Since we want this model to be very specific on the COVID-19 domain, use the same vocabulary as ALBERT-base could potentially compromise the results of downstream tasks.

Also, because of the resource limitation, we only use half of the entire corpus from the CORD-19 dataset and we reduce the dupe factor which denotes how many times a sentence will be masked differently when doing the Masked Language Modeling (MLM) part of pretraining to 5 instead of the original 40 used by ALBERT when creating pretraining data. These 2 design choices may affect the model's performance as well.

To avoid the potential overfitting, we save multiple checkpoints when doing pretraining, in section 5, we compare models from 2 different checkpoints and show their performances are similar.

4.3 Fine-tuning the Models and Test on Our Custom QA Dataset

For fine-tuning the 3 models we choose, we use 3 different strategies:
1. Fine-tune the models solely on SQuAD.
2. Fine-tune the models firstly on SQuAD and then use BioASQ to do further fine-tuning.
3. Fine-tune the models firstly on BioASQ and then use SQuAD to do further fine-tuning.
After the fine-tuning is done, we then apply the models on our manually labeled COVID-19 QA dataset and use their performances as indicators on how well these models could generalize on larger COVID-19 QA tasks in the future.

5 Results and Analysis

5.1 Model Comparison

The performances for different models and different fine-tuning strategies on our manually labeled test set is shown in Table 1. We could see that the performances of ALBERT and BioBERT are similar and they both significantly outperform our pretrained COVID-ALBERT model. We think the probable reason is that the multiple design choices mentioned in section 4.2 did compromise this model's ability to some degree. We also notice that the strategy of fine-tuning on SQuAD outperforms all other strategies. We think this result indicates that SQuAD is a comprehensive QA dataset and it could be broadly applicable to various kinds of QA tasks to supply as a training dataset. The results in the table also show that fine-tuning on the BioASQ dataset harms the model's performance no matter where we choose to use it, we think this is probably due to that the BioASQ dataset is too specific in its domain which makes it less compatible with the test set we used.

Table 1: Model comparison result

Model	EM	F1
ALBERT-base		
SQuAD	**65.50**	83.60
SQuAD then BioASQ	42.00	64.04
BioASQ then SQuAD	61.00	80.26
BioBERT-base		
SQuAD	65.00	**83.87**
SQuAD then BioASQ	45.00	64.07
BioASQ then SQuAD	63.50	81.99
COVID-ALBERT_58000		
SQuAD	52.00	69.44
SQuAD then BioASQ	26.50	46.64
BioASQ then SQuAD	46.50	66.28
COVID-ALBERT_100000		
SQuAD	52.00	69.00
SQuAD then BioASQ	34.50	51.52
BioASQ then SQuAD	50.00	67.55

It's curious to see that BioBERT does not perform worse than ALBERT while at the same time using BioASQ dataset harms models' performances given the context that both BioBERT model and BioASQ dataset are based on PubMed documents. Hence we are still trying to come up with ideas to figure out what causes this difference. Finally, from the results, we could also point out that all models we use in our experiments don't have resistance to catastrophic forgetting, they all forget what they have learned before quickly after we start to train them on different datasets.

5.2 Answer Type Breakdown Result

We use the prediction of ALBERT model (fine-tuned solely on SQuAD) that reaps the highest EM (exact match) to evaluate how well it could predict answers for different answer types. The result is showed at Table 2. The no answer type obtains the highest EM and F1. This is because Albert predict 15 no answer in total while only 10 questions are unanswerable. General teams, location, date, number, person types are also well predicted with desirable EM score, meaning that the model can predict the general noun or noun phrase well. In contrast, Biomedical terms type only obtains 57.38 EM, significantly less than mean EM 65.5. Due to the generalization of ALBERT and SQuAD, the model has not learned domain-specific word representation and thus shows deficiency in identifying such uncommon biomedical terms. Verb/adjective phrase and clause obtain the lowest EM because they are normally long answers, making the exact match very difficult.

In terms of F1 score, to our surprise, number and person types are even less than the biomedical terms and verb/adjective phrase/clause types. This might result from the fact that there are generally very few tokens in the answer span(1 or 2 in most cases). Either overestimated answer span or underestimated span could lead to huge F1 reduction.

This analysis verifies the necessity of pretraining or fine-tuning on domain-specific corpus or dataset. However, the best way to integrate the domain-specific data into general language representation is still yet to solve.

Table 2: Albert result breakdown by answer type

Answer type	EM	F1
no answer	90.00	90.00
general term	81.25	88.88
location	73.68	88.27
date	72.72	89.00
number	70.00	79.45
person	63.64	80.00
biomedical term	57.38	81.72
verb/adjective phrase/clause	50.00	84.05

6 Discussion and Future Work

Based on the results shown in section 5, we conclude that ALBERT is comprehensive enough for doing starter works at a new domain. We also want to point out that although our pretrained COVID-ALBERT performs the worst compared to ALBERT and BioBERT, it's too early to determine that COVID-19 specific BERT model is not useful at all since the failure of this model could be due to various design choices we make. Hence in the future, when we have enough resources, we want to utilize larger text corpus and larger dupe factor to pretrain another more comprehensive COVID-19 BERT model and replicate our experiments in this work.

We notice that recently there is a work in which the authors provide 124 Covid-19 related QA pairs (Tang et al., 2020). Although their dataset is not used in this project, we would like to use it in the future when we have our newer version of COVID-19 BERT model.

Moreover, as pointed out in the previous section, how to deal with the problem of catastrophic forgetting is another issue we want to solve in the future. We believe that if we find a way to do continual learning, it could be easier to apply ALBERT or other BERT-related models on different downstream tasks by fine-tuning it on multiple datasets in a continuous fashion and thus potentially achieve better results.

7 Collaboration Statement

Xiao Li is mainly responsible for preprocessing part of the data and pretraining the ALBERT model on Kaggle COVID-19 texts. Yanyan Xu mainly worked on COVID19 test QA data preprocessing, model fine-tuning and paper drafting. Zian Chen is responsible for BioASQ data preprocessing and model fine-tuning. Yichen Liu mainly worked on manually creating the Kaggle labeled question and answer pairs as well as model fine tuning. All members participate in the fine-tuning part equally.

8 Reproducibility

All code and mannually labelled QA testset for this report is available at `https://github.com/Heimine/NLU_project`.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. arXiv:1903.10676v3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv:1904.05342.

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

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. arXiv:1901.08746.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv:1606.05250.

Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly bootstrapping a question answering dataset for covid-19. arXiv:2004.11339.

George Tsatsaronis, Georgios Balikasand, 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-Cyrille Ngonga Ngomo, 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*.

Donghyeon Kim Minbyul Jeong Jaewoo Kang Wonjin Yoon, Jinhyuk Lee. 2019. Pre-trained language model for biomedical question answering. arXiv:1909.08229.