CS230

Developing Answers to Scientific Questions with BERT

Category: Natural Language Processing

Siheng He
SCPD
Stanford University
siheng@stanford.edu

Zahra Bakhtiari
SCPD
Stanford University
zahra22@stanford.edu

Abstract

In this paper we show that BERT model fine-tuned on SQuAD for Question Answering (QA) tasks can be successfully extended to help address emerging COVID-19 questions. By further fine-tuning on COVID data and utilizing BERT with biomedical text understanding capabilities, we are able to improve Exact Match by 49% and F1 score by 30% for COVID QA tasks. As more COVID or biomedical QA datasets are accumulating, the model will continue to advance to find the right answer for more scientific questions.

1 Introduction

In response to the COVID-19 pandemic, academia communities have published a skyrocketing number of scholarly articles recently [1]. In the mean time, there are emerging requests to develop text and data mining tools that can help the medical community develop answers to high priority scientific questions related to COVID-19. We propose a modular system for natural language processing (NLP) tasks like Reading Comprehension (RC) and Question Answering (QA). The input will be COVID-19 related queries/questions issued by users and the outcome will be a list of relevant snippets, which are selected, ranked and highlighted from the peer reviewed medical literature by our model. The system can be further integrated into the AI-powered robotic scientist [2] to build a fully autonomous robotic researcher.

2 Related work

Question answering is one of the main NLP tasks for assessing the reading comprehension capabilities of AI systems. QA tasks rely on the model capability to understand the natural language. A standard, high-quality benchmark dataset for evaluation of QA algorithms is the Stanford Question Answering Dataset (SQuAD), which consists of 100k+ questions on a set of Wikipedia articles, where the answer to each question is a text snippet from corresponding passages [3]. SQuAD2.0 takes a step further by combining the 100k questions with 50k+ unanswerable questions that look similar to answerable ones. For QA to answer hard questions, an unsupervised question decomposition can break a hard question into a series of sub-questions [4].

CS230: Deep Learning, Fall 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

3 Dataset and Features

The main dataset is COVID-QA[1], a QA dataset consisting of 2,019 COVID-19 related question/answer pairs. QA pairs are based on 147 scientific papers from the COVID-19 Open Research Dataset (CORD-19)[2], which is released by Allen Institute for AI and freely available. In addition, SQuAD 1.1 dataset will be used to fine-tune the BERT model for QA tasks. SQuAD is a reading comprehension dataset, consisting of 100,000+ question-answer pairs posed by crowdworkers on a set of 500+ Wikipedia articles. All QA datasets are stored in JSON format.

4 Methods

Bidirectional Encoder Representations from Transformers (BERT) [5] and ELMo [6] is a state-of-the-art contextualized word representation model for learning word representations from a large amount of unannotated text. One objective is to predict randomly masked words in a sequence. We will explore BERT to train a model and learn the word representations along the way. An existing implementation, BioBERT QA model [7], which is initialized with weights from BERT and then pre-trained on biomedical domain corpora, may be used a a baseline. The word representations (especially COVID-19 related terms) can be further optimized by fine-tuning previous model on the new CORD-19 data. The prediction probability of the QA model can be leveraged as the confidence score, which will serve as the input for ranking answers. In addition, we may try to utilize pre-trained UniLM language model [8] to summarize the top 3 snippets that are most relevant to the query.

5 Experiments/Results/Discussion

5.1 Experiments

We start with pre-trained BERT-Base Uncased model[3] as the baseline model. BERT-Base has 12 hidden layers, 768 hidden units, 12 heads and a total of 110M parameters. `Uncased` means that the text has been lowercased before WordPiece tokenization, e.g., `COVID` becomes `covid`. The uncased model also strips out any accent markers. BERT itself is trained to Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) together, with the goal of minimizing the combined loss function of the two strategies. The pre-trained model is further fine-tuned on SQuAD training data to better handle QA tasks.

In the fine-tuning process for QA tasks (Figure 1), BERT learns two extra vectors that mark the beginning and the end of the answer span. Let y be the one-vector which marks the start/end position of an answer within a word sequence of length l, $\hat{y}$ be the predicted probability (softmax activation output) of being start/end position, the loss function for start/end position is

$$L_{\text{pos}} = -\frac{1}{l}\sum^{l} y_{\text{pos}} * \log \hat{y}_{\text{pos}} \qquad \text{where pos} \in \{\text{start}, \text{end}\}.$$

Fine-tuned BERT will minimize the total loss:

$$L_{\text{total}} = \frac{L_{\text{start}} + L_{\text{end}}}{2}$$

COVID-QA is split into train/dev/test with 60%/20%/20% ratio. COVID-QA training data is also concatenated with SQuAD training data.

In Experiment 1, which is also the baseline study, the BERT model is initially fine-tuned with hyperparameters suggested in the original BERT repository[4]. The hyperparameters are further refined within the vicinity of their initial values, based on their performance on COVID-QA dev set. They are summarized in Table 1.

[1] https://github.com/deepset-ai/COVID-QA

[2] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[3] https://github.com/google-research/bert

[4] https://github.com/google-research/bert#squad-11

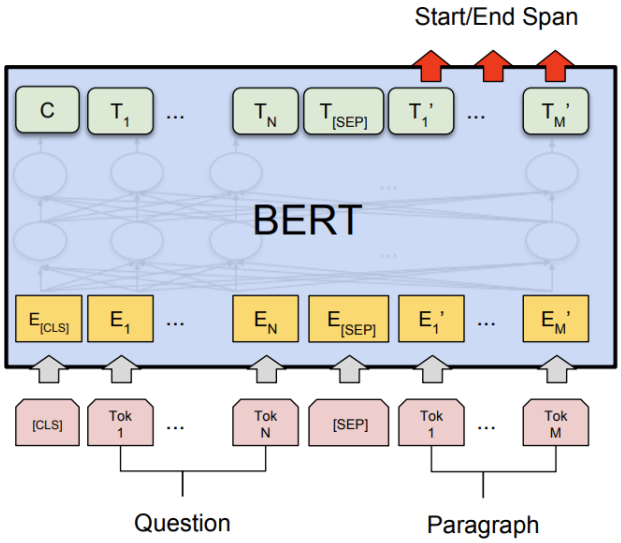

Figure 1: Illustrations of fine-tuning BERT for Question Answering. Adapted from [5].

Model	Learning Rate	Mini-Batch Size	Max Seq Length	Stride	Training Epochs
BERT-Base Uncased	3e-5	16	256	128	2∼3

Table 1: Selected hyperperameters

In Experiment 2, the BERT model is fine-tuned on the SQuAD + COVID-QA training data.

In Experiment 3, the starting model is BioBERT and it is fine-tuned on the SQuAD + COVID-QA training data.

5.2 Results

We adopt two evaluation metrics from SQuAD paper [3] as the main quantitative metrics:

- EM (Exact match)
- (Macro-averaged) F1 score

to compare the model output with the labeled question/answer pairs. Table 2 summarizes evaluations on the COVID-QA test set for each model. Some representative examples of exactly matched answers, partially matched answers and missed answers are shown in Table 3.

	BERT + SQuAD	BERT + SQuAD + COVID-QA	BioBERT + SQuAD + COVID-QA
EM	24.01	30.69	35.89
F1	45.40	55.99	58.87

Table 2: Experiment results.

5.3 Discussion

Overall, the best model (BioBERT + SQuAD + COVID-QA) shows significant increase in EM (35.89 vs 25.90) and comparable F1 score (58.87 vs 59.53) over [9].

BERT-Large fine-tuned on SQuAD traing set is able to achieve the state-of-the-art performance on SQuAD v1.1 (EM: 84.3, F1: 90.8). BERT-Base fine-tuned on SQuAD is already able to answer COVID related questions, but the performance is far away from the performance on SQuAD. The main reason for the disparity is the mismatch of data distributions. More specifically, scientific (especially biomedical) texts may have a very different data distribution than general purpose texts

Passage of Interest	Question	Evaluation
The SARS-CoV-2 virus is a betacoronavirus, like MERS-CoV and SARS-CoV. All three of these viruses have their origins in bats.	What type of virus is SARS-CoV-2?	Exact match
The models suggested that without a vaccine, school closures would be unlikely to affect the pandemic, an estimated 35,000 to 60,000 ventilators would be needed, up to an estimated 7.3 billion surgical masks or respirators would be required, and perhaps most important, if vaccine development did not start before the virus was introduced, it was unlikely that a significant number of hospitalizations and deaths could be averted due to the time it takes to develop, test, manufacture, and distribute a vaccine.	How many surgical masks or respirators have past studies projected will be required for a pandemic in the United States?	Partial match
Dendritic cell-specific ICAM-grabbing non-integrin-related (DC-SIGNR, also known as CD209L or liver/lymph node-specific ICAM-grabbing non-integrin (L-SIGN)) can interact with a plethora of pathogens including HIV-1 and is expressed in placental capillary endothelial cells. DC-SIGNR is organized ...	What is DC-GENR and where is it expressed?	Partial match
As of 24:00 on March 11, 2020, the National Health Commission (NHC) had received reports of 80,793 confirmed cases and 3,169 deaths on the Chinese mainland. There remain 14,831 confirmed cases ... In China, healthcare workers account for 1,716 confirmed cases of COVID-19, including six deaths.	What were the number of cases in mainland china as of March 11th?	Mismatch

Table 3: Question and answer pairs from the COVID-QA dataset. Expert labelled answers are marked with underline and predicted answers are highlighted in red. Parts of the paragraphs may be truncated for clarity.

(Wiki articles for SQuAD), including technical terms, different semantic segmentation, etc. By further fine-tuning on COVID-QA training set, the performance improves with a relatively small margin. Note that SQuAD consists of more than 100,000 QA pairs while COVID-QA training set consists of $\sim$1,200 QA pairs, the combined training set is still heavily biased towards general purpose texts and hence the model does not learn enough about biomedical text. However, the SQuAD dataset is still valuable in fine-tuning the model. As a supplementary experiment, we fine-tuned BERT on COVID-QA only and both EM and F1 metrics are close to zero (results not included in Table 2). In order to alleviate the data mismatch issue, we utilize BioBERT, which is a domain-specific language representation model pre-trained on large-scale biomedical corpora (PubMed abstracts and PMC full-text articles). Indeed BioBERT + SQuAD + COVID-QA shows further improvements on metrics.

Error analysis shows that $\sim$30% of partial matches are actually due to errors in the labels and mismatch between questions and answers in the COVID-QA dataset. For example, in the third example in Table 3, the first part of the question is asking about DC-GENR, while the labeled answer is about DC-SIGNR and DC-GENR does not appear at all in the whole paragraph, therefore it is probably right for the model to skip the part of the question.

6 Conclusion/Future Work

We believe that our BioBERT fine-tuned on SQuAD and COVID-QA data serves a good starting point for further improvements on answering questions regarding COVID-19, or even applicable to broader biomedical questions. Therefore, we plan to further enhance the answer quality of our model and to tackle data mismatch issues, via constructing larger biomedical specific question-answer pairs or converting existing biomedical QA dataset like BioASQ[10] for our particular use cases.

7 Contributions

Both team members contributed equally to the project. Siheng He proposed the project idea, built the baseline model, improved the final model, and finished all reports. Zahra Bakhtiari processed data set to build train and dev set, improved the baseline and final model on AWS GPUs, built final slides and created the video.

References

[1] Rebecca C Jones, Jasper C Ho, Hannah Kearney, Meghan Glibbery, Daniel L Levin, John Kim, Sara Markovic, Jillian Howden, Maya Amar, and Mark A Crowther. Evaluating trends in covid-19 research activity in early 2020: The creation and utilization of a novel open-access database. *Cureus*, 12(8), 2020.

[2] Benjamin Burger, Phillip M Maffettone, Vladimir V Gusev, Catherine M Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M Alston, Buyi Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I Cooper. A mobile robotic chemist. *Nature*, 583(7815):237–241, 2020.

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

[4] Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. Unsupervised question decomposition for question answering. *arXiv preprint arXiv:2002.09758*, 2020.

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

[6] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[7] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[8] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075, 2019.

[9] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Lawrence Livermore. Covid-qa: A question & answering dataset for covid-19. *ACL 2020 Workshop on Natural Language Processing for COVID-19*, 2020.

[10] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138, 2015.