Poly-encoder based Task Adaptation for COVID-19 Question Answering System

Seolhwa Lee  
Johns Hopkins University  
Korea University  
Baltimore, MD  
Seoul, South Korea  
slee540@jhu.edu

João Sedoc  
Johns Hopkins University  
Baltimore, MD  
jsedoc@jhu.edu

Abstract

In this unprecedented era of COVID-19, there is a huge information need for truthful answers to questions relating to all aspects of the Coronavirus. In this paper, we adapt the Poly-Encoder (Humeau et al., 2020) informational retrieval from FAQs. We show that after fine-tuning the Poly-encoder we achieve higher recall at the same level of precision. We compare several models on the task of answer or question selection on evaluation sets and demonstrate superior performance. We make our code publicly available for other researchers to use.

1 Introduction

At the beginning of 2020, the COVID-19 pandemic spread quickly across the world. The information regarding COVID-19 needs to be communicated by verified sources such as Center for Disease Control and Prevention (CDC) and the World Health Organization (WHO). Furthermore, COVID-19 articles are increasing rapidly and also there is information everywhere, where some of this is evidence-based and some is misinformation. In order to help the society respond to COVID-19, people need to acquire these information easily. A conversational interface provides an accessible Question Answering (QA) system using trusted sources.

Research of COVID-19 related work focused on the symptoms (Zhu et al., 2020) and medical domain QA.1 However, here is the recent efforts in natural language processing (NLP) techniques (Li et al., 2020; Sri Manasa Venigalla et al., 2020). For the COVID-19 of NLP territory, we adapt QA system for delivering current COVID-19 connected information. There are mainly three challenges in research COVID-19 QA system using NLP techniques.

The first challenge is that setup the COVID-19 QA task on our purpose. Performance in such task have to be estimated into two aspects: prediction speed and also prediction quality, as scoring many candidates can be remarkably lazy. Recently, the deep pre-trained language models (Devlin et al., 2018) used by fine-tuning to achieve state-of-the-art benchmarks on a QA task (Zhang et al., 2020). But we approach to appropriate our task, in this work, adapt the Poly-encoder (Humeau et al., 2020) for the COVID-19 QA task.

Another challenge is that return trusted information to natural user questions for the COVID-19 QA system. In hence, we have collected two-thousands of verified questions and answers from trusted online sources (e.g., WHO, CDC). The answers are verified by professionals at the Johns Hopkins University Bloomberg School of Public Health.

The third challenge is the evaluation for the COVID-19 QA system. In order to conduct research on this important problem, experts at the JHU Bloomberg School of Public Health have used our human annotation interface for similarity between user questions and those coming from the QA knowledge base. This allow us to evaluate our QA system.

Our contributions are four-fold:

• we build the trusted information dataset for the COVID-19 QA system.

• We adapt COVID-19 QA task on our purpose

• We consistently refine our system through the feedback between experts and system interactively

• We propose novel evaluate sets for COVID-19 QA system and get the results using these datasets


2 COVID-19 QA task

We consider the COVID-19 QA task of answer sentence selection from a set of pre-selected sentences (Severyn et al., 2013). That is to say, we approach finding similar questions from the user query, which are already answered such as “Frequently Asked Questions (FAQs)” (Wang et al., 2009). In fact, COVID-19 QA task we are seeking in the QA context should include two points, which are the response speed considering real-time system and worthwhile quality for delivering the accurate information.

2.1 Poly-encoder

The Poly-encoder depend on large pretrained transformer models with the identical architecture and dimension as BERT-base (Devlin et al., 2018), which also has 12 attention heads, 12 layers, and a hidden size of 768. In conjunction with reflecting the BERT pre-training weights, the Poly-encoder used pre-training schemes from scratching using the same architecture as BERT-base. We use trained setup model based on the Reddit that extracted as 174 million examples of INPUT, LABEL (Mazaré et al., 2018). We decide to use this pretrained model since Reddit is a dataset more tailored to dialogue.

The Poly-encoder accepts two separate transformers for the label and context, which is encoded into vectors like $y_{\text{cand}} = \text{red}(T_1(\text{cand}))$, $y_{\text{ctxt}} = \text{red}(T_2(\text{ctxt}))$, where $T_1$ and $T_2$ are two transformers that have been pre-trained in (Humeau et al., 2020). That’s more, the candidate is encoded into a single vector $y_{\text{cand}}$. Accordingly, the Poly-encoder methodology can be carried out accepting encoded responses from the precomputed cache.

But, the input context is generally significant longer than a candidate. Therefore, the input context is represented with $m$ vectors $(y_{\text{ctxt}}^1, ..., y_{\text{ctxt}}^m)$, where $m$ will affect the inference speed. These $m$ global features use as the input representation, which learn $m$ context codes $(c_1, ..., c_m)$, where $c_i$ extracts representation $y_{\text{ctxt}}^i$ by attending over all the outputs of the previous layer. $y_{\text{ctxt}}^i$ can be obtained as below:

$$y_{\text{ctxt}}^i = \sum_j w_{ji}^i h_j$$

where $(w_{1j}^i, ..., w_{Nj}^i) = \text{softmax}(c_i \cdot h_1, ..., c_i \cdot h_N)$. The randomly initialized $m$ is context codes, and learnt during fine-tuning.

Consequently, $y_{\text{cand}}$, used for attending over $m$ global context feature.

$$y_{\text{ctxt}} = \sum_i w_i y_{\text{ctxt}}^i$$

where $(w_1, ..., w_m) = \text{softmax}(y_{\text{cand}}, y_{\text{ctxt}}, ..., y_{\text{cand}}, y_{\text{ctxt}}^m)$. In the end, $y_{\text{cand}} \cdot y_{\text{ctxt}}$ is the final score of the candidate label. The context candidate attention is only operated at the top layer and as $m < N$, where $N$ is the number tokens. This is the reason extremely faster than previous models.

As a consequence of adjustment in our task, we use the Poly-encoder to maintain the ability for pre-computing representation of each candidate, as it provides a mechanism for attending over the context utilizing the label candidate. Furthermore, the Poly-encoder allows for rapid real-time inference in a construction setup, specifically, they give an improved trade off between speed and accuracy.

2.2 Fine-tuning and Task Adaptation

In this section, we address the method to adapt our task on Poly-encoder. To build our own task, we use the open source software platform the ParlAI (Miller et al., 2017), which is specialized on dialogue research. Our aim is to build our own task and share our experience to other researchers. Hence, our task that goes into the repository should build towards that common goal, rather than been seeing solely as a piece of independent research.

First, we build the COVID-19 task follow the WikiQA task. WikiQA dataset is publicly available set of answer and question pairs, collected and annotated for research on open domain question and answering (Yang et al., 2015). Our task also have question and answer candidate pairs, therefore, it is adjust on our objective. And then, we fine-tune the Poly-encoder for the task with focus on improving accuracy and relevance of answers. Finally, we implement chat interface to receive the feedback between experts and our system response interactively. This feedback used for refining our system to provide proper information.

https://covid-19-infobot.org/chat/
Train Val Test Total
# of ques. 541 67 68 676
# of sent. 3,824 486 427 4,737

Table 1: Statistics of the COVID-19 QA dataset (Q-A) (English). We use schema_{v0.2, 04-30 dump version} dataset.

Test
# of ques. 254
# of avg.word 7.11

Table 2: Statistics of the COVID-19 QA dataset (Q-Q).

3 Experimental Setup

We use fine-tuned Poly-encoder architecture of Reddit-pre-trained transformer for our fine-tuning. We set the batch size to be 64. The model was trained for 20 epochs. The parameter setting details available at our repository mentioned Section 2.2.

3.1 COVID-19 QA Dataset

Our COVID-19 QA dataset 5 (Q-A) consists of 2,300 questions and answers news article and FAQs from over 32 trusted online source (e.g., CDC, WHO, CNN). We scraped each questions and answers as extract the relevant metadata 6, which are updated daily. This efforts conduct by JHU-COVID-QA team 7.

COVID-19 QA dataset incorporates multilingual such as english, german, polish, italian and spanish. We only use english data for the training. We randomly split the data to training (80%), validation (10%) and testing (10%) sets. Also, we use the dataset in case of ‘hasAnswer’ is true. Table 1 shows COVID-19 QA dataset statistics. We use the ParlAI data loader for providing dataset download automatically and preprocess the data to adjust our task format as [question, [cand1, cand2, ...]]. The dataset for the training available at our code repository. Every questions have the answer pair and we randomly sampled 20 candidates answer including ground truth for the training per each question. We chose the number of candidates followed (Humeau et al., 2020) and further provide the performance for a given number of candidates on Table 3.

3.2 COVID-19 QA Evaluation

Evaluation sets We evaluate more using question to question (Q-Q) set that go beyond general QA system by leveraging experts knowledge, allowing it to go beyond trusted information matching. In this setting, we collaborate with public health experts to extract the relevance score of the candidate question-answer pairs into a scale of 0–100. And then, we use these engaged data for calculating Q-Q score through BERT-based Q-A system 8. In the end, Q-Q dataset can be scored through the similarity between query and question.

Table 2 shows the number of questions is 254 and the number of average word in question is 7.11. We scored each of the question-question pairs through the experts. The ground truth consists of the score is more than or equal 80. We discard the question in case of less than 80. Also, the candidates consist of five different questions.

Evaluation metrics We perform the evaluation for the four metrics. (1) Accuracy is for the system response correctness. (2) MRR (Mean Reciprocal Rank) evaluate the relative ranks of ground truth answers in the candidate of a question. (3) F1 employed for the consider both the correctness of system predictions and the presence of answers with respect to a question. Especially, F1 score has been generally reserved for evaluating retrieval based question answering. (4) BLEU-n is a precision based metric by computing the number of n-grams in the candidate that also present in a reference.

4 Results and Analysis

We conducted experiments through two types of evaluation sets. One is COVID-19 QA dataset (Q-A) for the evaluation, which is designed question-answer pairs consists of randomly candidates answers. To investigate the model’s ability according to the number of candidates, we distinguish the experiment models into @k, indicates the number of answer candidates. We fine-tuned @20 case and evaluate this model using @10 candidates Q-A and Q-Q evaluation dataset. The results are shown in Table 3. The metric score results are not surprise that JHU-COVID-QA@20(ft)@10 outperforms JHU-COVID-QA@20(ft)@20. Moreover, we compare the Poly-encoder model as a baseline, which is Reddit-pre-trained transformer noted at

5https://covid-19-infobot.org/data/
6https://github.com/JHU-COVID-QA/scraping-qas/wiki/Schema-v0.1
7https://covid-19-infobot.org/
8We use sentence-transformer model (Reimers and Gurevych, 2019) with Q-A score annotated data for getting the score in question-question.
### Validation (Q-A)

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F1</th>
<th>BLEU-4</th>
<th>MRR</th>
<th>hits@5</th>
<th>hits@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly@20 (Reddit)</td>
<td>0.3582</td>
<td>0.4497</td>
<td>0.3582</td>
<td>0.5368</td>
<td>0.9701</td>
<td>1.0</td>
</tr>
<tr>
<td>Poly@10 (Reddit)</td>
<td>0.4627</td>
<td>0.5391</td>
<td>0.4629</td>
<td>0.545</td>
<td>0.9701</td>
<td>1.0</td>
</tr>
<tr>
<td>JHU-COVID-QA@20(ft) @20</td>
<td>0.8507</td>
<td>0.8818</td>
<td>0.8507</td>
<td>0.8846</td>
<td>0.9701</td>
<td>1.0</td>
</tr>
<tr>
<td>JHU-COVID-QA@20(ft) @10</td>
<td>0.8955</td>
<td>0.917</td>
<td>0.8955</td>
<td>0.9353</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Test (Q-A)

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F1</th>
<th>BLEU-4</th>
<th>MRR</th>
<th>hits@5</th>
<th>hits@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly@20 (Reddit)</td>
<td>0.3676</td>
<td>0.4613</td>
<td>0.3733</td>
<td>0.545</td>
<td>0.9701</td>
<td>1.0</td>
</tr>
<tr>
<td>Poly@10 (Reddit)</td>
<td>0.4559</td>
<td>0.5413</td>
<td>0.4618</td>
<td>0.6393</td>
<td>0.9559</td>
<td>1.0</td>
</tr>
<tr>
<td>JHU-COVID-QA@20(ft) @20</td>
<td>0.7941</td>
<td>0.8317</td>
<td>0.7941</td>
<td>0.875</td>
<td>0.9853</td>
<td>1.0</td>
</tr>
<tr>
<td>JHU-COVID-QA@20(ft) @10</td>
<td>0.9853</td>
<td>0.9898</td>
<td>0.9853</td>
<td>0.9926</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Q-Q

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F1</th>
<th>BLEU-4</th>
<th>MRR</th>
<th>hits@5</th>
<th>hits@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly (Reddit)</td>
<td>0.7283</td>
<td>0.7945</td>
<td>0.6933</td>
<td>0.5864</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>JHU-COVID-QA@20(ft)</td>
<td>0.7205</td>
<td>0.8126</td>
<td>0.6959</td>
<td>0.5914</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 3: Performance of the JHU-COVID-QA system using COVID-19 QA dataset (Q-A & Q-Q pair). (ft) indicates fine-tuned model. JHU-COVID-QA@20(ft) @10 / 20 indicates evaluate 10 / 20 candidates QA with JHU-COVID-QA@20(ft).

Section 2.1 and 3. The comparison between our JHU-COVID-QA model and Poly-encoder (Reddit) is noteworthy us to important the domain adaptation. These comparison highlighted the Table 3.

Another is question to question (Q-Q) for the evaluation. In this case, we perform JHU-COVID-QA@20 model with Q-Q evaluation. The result shows that the accuracy is 72.05%. We demonstrate only Q-A fine-tuned model with Q-Q evaluation. That’s why the accuracy lower than Q-A evaluation. Furthermore, it need to approach fine-tune with Q-Q data format, which means Q-Q and Q-A are necessary different type of being fine-tuned model for dealing with both. For the future work, we look forward to high performance with Q-Q fine-tuned model in COVID-19 QA system.

## 5 Conclusion and Future Work

In this paper, we showcased how fine-tuning the COVID-19 QA system for task adaptation. We introduce the COVID-19 QA task and applied it to adjust multi-selection model. Experimental results across two different evaluation sets show that our model achieves meaningful results.

Although we mainly focused on the Poly-encoder model, to further our research, we plan to take advantage system combination of BERT based and BM25 model considering in contextual bandits (Foster and Rakhlin, 2020) for the more high quality response on the QA system.

**Acknowledgments**

This research was supported by the MOTIE (Ministry of Trade, Industry, and Energy) in Korea, under the Fostering Global Talents for Innovative Growth Program (P0008749) supervised by the Korea Institute for Advancement of Technology (KIAT).

**References**


Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training


