Using the Poly-encoder for a COVID-19 Question Answering System

Seolhwa Lee[*]
Korea University
Seoul, South Korea
whiteldark@korea.ac.kr

João Sedoc
New York University
New York, USA
jsedoc@stern.nyu.edu

Abstract

To combat misinformation regarding COVID-19 during this unprecedented pandemic, we propose a conversational agent that answers questions related to COVID-19. We adapt the Poly-encoder (Humeau et al., 2020) model for informational retrieval from FAQs. We show that after fine-tuning, the Poly-encoder can achieve a higher F1 score. 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 as did harmful misinformation (e.g., about how the disease spreads) leading to preventable mistakes in managing its spread. Information regarding COVID-19 must be communicated by verified sources such as the Center for Disease Control and Prevention (CDC) and the World Health Organization (WHO). Furthermore, COVID-19 information is rapidly changing, making it difficult to distinguish evidence-based guidance from misinformation. To help society respond to COVID-19, people need easy access to accurate information about COVID-19. For some people, a conversational interface can provide a quick and accessible interface to trusted sources.

We adapt the Poly-encoder (Humeau et al., 2020) for the COVID-19 QA task. The performance of a QA model is assessed not only by prediction quality but also by prediction speed, as scoring many candidates can be extremely slow. Recently, the fine-tuning of deep pre-trained language models (Devlin et al., 2018) has been shown to achieve state-of-the-art benchmarks on a QA task (Zhang et al., 2020b). The Poly-encoder extends these methods for conversational information retrieval.

Another challenge for the COVID-19 QA system is to return trusted information to user questions. Hence, we use the dataset of Poliak et al. (2020) which has over two-thousand verified questions and answers from trusted online sources (e.g., WHO, CDC). The answers are verified by researchers at the Johns Hopkins Bloomberg School of Public Health (JHSPH).

Finally, a third challenge is how to evaluate the COVID-19 QA system. To research this important problem, experts from JHSPH annotated a dataset of similarity scores between FAQs from trusted sources and user questions from online sources (e.g. Qorona[1], Twitter, and DialogueMD[2]).[3] This similarity set allows us to evaluate our QA system.

We contribute the following:

- We adapt the Poly-encoder for the COVID-19 QA task.
- We evaluate the performance using the question - QA similarity dataset.

2 COVID-19 QA

We formulate our COVID-19 QA task as a problem of how to choose answers from a set of pre-selected sentences (Severyn et al., 2013). Specifically, we want to respond to a user query with a response from Frequently Asked Questions (FAQs) (Wang et al., 2009). The Poly-encoder model can use the QA context to respond both quickly and accurately to user queries.

2.1 Poly-encoder

The Poly-encoder (Humeau et al., 2020) depends on large pre-trained transformer models with identi-

[*] This work was conducted while a visiting student at Johns Hopkins University.

[1] https://github.com/allenai/Qorona

[2] https://github.com/dialoguemd/covidfaq

[3] For a full description of this dataset see Poliak et al. (2020).

cal architecture and dimensions as BERT-base (Devlin et al., 2018), i.e. 12 attention heads, 12 layers, and a hidden size of 768. The Poly-encoder is trained from scratch using a subset of Reddit extracted data composed of 174 million examples of [INPUT, LABEL] (Mazaré et al., 2018). We use this pre-trained model since questions on Reddit are similar to COVID-19 questions (Murray et al., 2020; Zhang et al., 2020a).

The Poly-encoder accepts two separate transformers for the label and context, which encode them into vectors, $y_{cand} = red(T_1(cand))$, $y_{ctxt} = red(T_2(ctxt))$, where T_1 and T_2 are two transformers that have been pre-trained (Humeau et al., 2020). $T(x) = h_1, ..., h_N$ is the output of a Transformer T and N is the number of tokens. $red(\cdot)$ is the mean-pool function which reduces the sequence of vectors into one vector. The candidate is encoded into a single vector y_{cand_i}. The Poly-encoder model precomputes and caches embedded responses.

In general, the context is significantly longer than a candidate; therefore, the input context is represented with m vectors $(y^1_{ctxt} \cdots y^m_{ctxt})$, where m will affect the inference speed. The m global features are used as the input representation, which learns m context codes $(c_1, ..., c_m)$, where c_i extracts representation y^i_{ctxt} by attending over all the outputs of the previous layer. Formally,

$$y^i_{ctxt} = \sum_j w^{c_i}_j h_j,$$

where $(w^{c_i}_1, ..., w^{c_i}_N) = softmax(c_i \cdot h_1, ..., c_i \cdot h_N)$. The m context codes are randomly initialized and learned during fine-tuning.

Consequently, we use y_{cand_i} to attend over m global context feature, such that,

$$y_{ctxt} = \sum_i w_i y^i_{ctxt},$$

where $(w_1, ..., w_m) = softmax(y_{cand_i} \cdot y^1_{ctxt}, ..., y_{cand_i} \cdot y^m_{ctxt})$. In the end, $y_{cand_i} \cdot y_{ctxt}$ is the final score of the candidate label.

We note that the Poly-encoder is faster than previous models which take context into account because the context candidate attention is only over the last layer and the number of context codes is much smaller than the number of tokens (i.e. $m < N$).

We use the Poly-encoder to precompute the 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.

2.2 Fine-tuning and Task Adaptation

We use the open-source software platform the ParlAI (Miller et al., 2017), which is specialized for dialogue research. We aim to not only build our chatbot but also share our results with other researchers via our repository[4] in order to build towards the common goal of providing the public with an accessible source for up-to-date and verified information on COVID-19.

First, we build the COVID-19 model following the same parameters and training approach as the WikiQA task.[5] The WikiQA dataset is a publicly available set of answer and question pairs, collected and annotated for research on open-domain question and answering (Yang et al., 2015). Given the similarity between tasks, we fine-tune the Poly-encoder using the WikiQA recipe for our task with a focus on improving the accuracy and relevance of answers. Finally, we implement a chat interface[6] to receive feedback from experts who interact with our system response. We use this feedback to refine our system to provide proper information.

3 Experimental Setup

We fine-tuned the Poly-encoder with the Reddit-pre-trained transformer using the COVID-19 QA dataset and evaluated it on the Q-A and Q-Q datasets.

3.1 COVID-19 QA Dataset

We used the COVID-19 QA dataset[7] (Q-A) from Poliak et al. (2020). The dataset has roughly 2,200 English questions and answers from FAQs from over 40 trusted online sources (e.g., CDC, WHO, CNN). These websites are scraped and the question and answer pairs are extracted along with the relevant associated metadata.[8] The data is updated daily. This effort was conducted by JHU-COVID-QA team[9] (Poliak et al., 2020).

[4] https://github.com/sseol11/Parlai_ver2/tree/master/parlai/tasks/covid19

[5] https://github.com/facebookresearch/ParlAI/tree/master/parlai/tasks/wikiqa

[6] We use ParlAI demo interface and open-source the demo system (See Appendix 1) for researchers who want to study this topic in our repository outside of the full system https://covid-19-infobot.org/chat/.

[7] https://covid-19-infobot.org/data/

[8] https://github.com/JHU-COVID-QA/scraping-qas/wiki/Schema-v0.1

[9] https://covid-19-infobot.org/

	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).

COVID-19 QA dataset incorporates multilingual such as English, German, Polish, Italian, and Spanish.[10] We only used the English data where 'hasAnswer' is true. We randomly split the data into training (80%), validation (10%), and testing (10%) sets. Table 1 shows COVID-19 QA dataset statistics. We used the ParlAI data loader to automatically download and preprocess the data in the format $[question, [cand_1, cand_2, ...]]$. The dataset used for our experiments is available in our code repository. For every question and answer pair, we randomly sampled 20 candidate answers including ground truth for training.

3.2 COVID-19 QA Evaluation

Evaluation sets We evaluated the model using a question to question (Q-Q) similarity set created by leveraging experts knowledge. We collaborated with public health experts to extract the relevance score of the candidate question-QA pairs into a scale of 1–100. For full details of the dataset reader should refer to Poliak et al. (2020). The Q-Q dataset can be scored through the similarity between query and question. The Q-Q evaluation dataset consists of five candidates question per question with relevance score from question-QA pairs.

We used 254 questions with an average of 7.11 words per question (Table 2). We scored each of the question-question pairs based on experts' assessment. The ground truth has a similarity score ≥ 80.

Evaluation metrics We used four metrics: accuracy, MRR (Mean Reciprocal Rank), F1 score, and BLEU-n. We assessed the accuracy of the model at selecting the single best response. MRR evaluates the relative ranks of ground-truth answers in the candidate of a question. F1 score considers both the precision and recall of the system. Finally, BLEU-n is a precision-based metric that computes the number of n-grams in the candidate that are also present in a reference.

Baseline Our baseline is the Poly-encoder pre-trained on Reddit, which we discuss in Section 2.1. We refer to this baseline as Poly-encoder (Reddit).

3.3 Fine-Tuning

We fine-tuned the model using 20 candidates form the COVID-19 QA dataset and thus refer to the model as JHU-COVID-QA@20(ft). We set the batch size to 64. The model was trained for 20 epochs. The specific parameter set is described in our repository.[4]

4 Results and Analysis

We conducted experiments using two types of evaluation sets. One is COVID-19 QA dataset (Q-A) for the evaluation, which is composed of question-answer pairs and includes randomly selected distractor candidate answers. To evaluate the model's ability relative to the number of candidates, we evaluated models with a different number of candidates in Table 3.

We evaluated our model and the baseline using both 10[11] and 20 candidates in Q-A (Table 3). As expected the metrics all showed worse performance as the candidate set increases; however, the performance of our fine-turned model decreases much less than the baseline. The comparison between our JHU-COVID-QA@20 model and baseline shows the importance of fine-tuning.

Another avenue of evaluation is the question to question (Q-Q) similarity set. In this case, we evaluated the JHU-COVID-QA@20 model with Q-Q set and attained an accuracy of 72.05%. This demonstrated that a model only fine-tuned on the Q-A dataset worked well on Q-Q evaluation; however, we only saw improvement in recall over the baseline. The performance can be improved by also including Q-Q as well as Q-A in fine-tuning. We leave this for future work.

[10] Deepset scraping websites include multiple languages, but roughly 65% of our datasets are in English.

[11] We randomly removed 10 of the 19 distractors candidates from @20.

			Validation (Q-A)		
	Candidates	Accuracy	F1	BLEU-4	MRR
Poly-encoder (Reddit)	20	0.3582	0.4497	0.3582	0.5368
Poly-encoder (Reddit)	10	0.4627	0.5391	0.4629	0.545
JHU-COVID-QA@20(ft) (ours)	20	0.8507	0.8818	0.8507	0.8846
JHU-COVID-QA@20(ft) (ours)	10	0.8955	0.917	0.8955	0.9353
			Test (Q-A)		
	Candidates	Accuracy	F1	BLEU-4	MRR
Poly-encoder (Reddit)	20	0.3676	0.4613	0.3733	0.545
Poly-encoder (Reddit)	10	0.4559	0.5413	0.4618	0.6393
JHU-COVID-QA@20(ft) (ours)	20	0.7941	0.8317	0.7941	0.875
JHU-COVID-QA@20(ft) (ours)	10	0.9853	0.9898	0.9853	0.9926
			Q-Q		
	Candidates	Accuracy	F1	BLEU-4	MRR
Poly-encoder (Reddit)	5	0.7283	0.7945	0.6933	0.5864
JHU-COVID-QA@20(ft) (ours)	5	0.7205	0.8126	0.6959	0.5914

Table 3: Performance of the JHU-COVID-QA system evaluated on COVID-19 QA datasets (Q-A & Q-Q pair). (ft) indicates that the model is fine-tuned.

5 Conclusion and Future Work

In this paper, we showcase the fine-tuning of the Poly-encoder model for the COVID-19 QA task. Experimental results across two different evaluation sets show that our model achieves meaningful improvements over the baseline. We create two evaluations sets from Poliak et al. (2020) and make them publicly available for researchers.

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

Our hope is that by deploying this conversational agent, the public will have another tool to get credible answers to their questions about COVID-19.

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) and IITP grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). This work was supported in part by DARPA KAIROS (FA8750-19-2-0034). The views and conclusions contained in this work are those of the authors and should not be interpreted as representing official policies or endorsements by DARPA or the U.S. Government.

References

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

Dylan J Foster and Alexander Rakhlin. 2020. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Curtis Murray, Lewis Mitchell, Jonathan Tuke, and Mark Mackay. 2020. Symptom extraction from the narratives of personal experiences with covid-19 on reddit. *arXiv preprint arXiv:2005.10454*.

Adam Poliak, Max Fleming, Cash Costello, Kenton W Murray, Mahsa Yarmohammadi, Shivani Pandya, Darius Irani, Milind Agarwal, Udit Sharma, Shuo Sun, et al. 2020. Collecting verified covid-19 question answer pairs.

Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. 2013. Learning semantic textual similarity with structural representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 714–718, Sofia, Bulgaria. Association for Computational Linguistics.

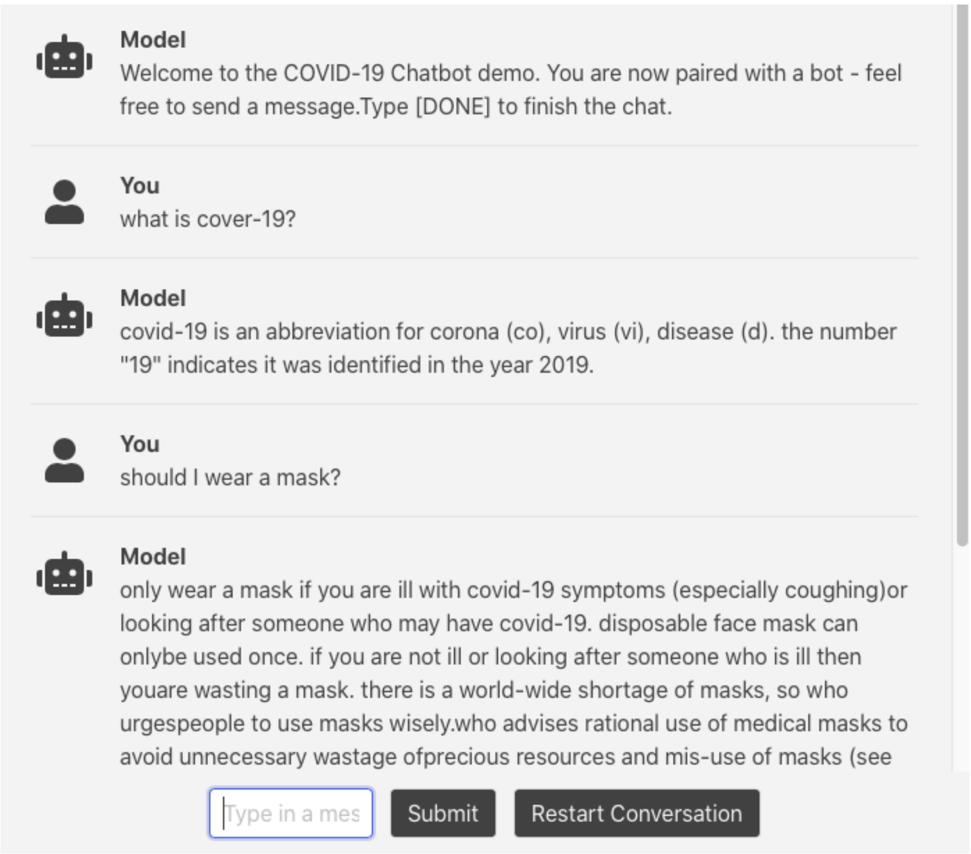
Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based qa services. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 187–194.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.

Jason Shuo Zhang, Brian C Keegan, Qin Lv, and Chenhao Tan. 2020a. A tale of two communities: Characterizing reddit response to covid-19 through/r/china_flu and/r/coronavirus. *arXiv preprint arXiv:2006.04816*.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020b. Retrospective reader for machine reading comprehension. *arXiv preprint arXiv:2001.09694*.

A Appendices



Model
Welcome to the COVID-19 Chatbot demo. You are now paired with a bot - feel free to send a message.Type [DONE] to finish the chat.

You
what is cover-19?

Model
covid-19 is an abbreviation for corona (co), virus (vi), disease (d). the number "19" indicates it was identified in the year 2019.

You
should I wear a mask?

Model
only wear a mask if you are ill with covid-19 symptoms (especially coughing)or looking after someone who may have covid-19. disposable face mask can onlybe used once. if you are not ill or looking after someone who is ill then youare wasting a mask. there is a world-wide shortage of masks, so who urgespeople to use masks wisely.who advises rational use of medical masks to avoid unnecessary wastage ofprecious resources and mis-use of masks (see

Figure 1: Interface of our COVID-19 Chatbot demo. Note that the example created based on April-2020 scraped information. Therefore, it may not up-to-date information.