Assignment 1 - Advanced Statistical NLP (CSE 291-3)

Kaushik Ravindran
A59012879
Computer Science And Engineering
University Of California, San Diego
kravindran@ucsd.edu

Yash Khandelwal
A59012226
Computer Science And Engineering
University Of California, San Diego
ykhandelwal@ucsd.edu

Abstract

COVID-19 has affected human beings all over the world in a significant manner. Updating and learning about any kind of information on this is the need of the hour for most of us.This can be achieved by a creating a mechanism that would answer to peoples queries. This can be accomplished by creating a QA model. We make use of COVID-QA dataset along with a wide range of NLP QA models to build our COVID QA system and evaluate our models.

1 Introduction

Impact of COVID-19 is not unknown to the world. This is a very new but infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). So extracting any kind of information on this disease is one of the primary requirements for most people. Extracting such information is a very challenging and non trivial problem. There is a demand to serve the purpose of extracting information to any kinds of queries. Question answering (QA) is a branch within natural language processing (NLP) and information retrieval fields, which is concerned with building systems that automatically answers questions posed by humans in a natural language. A Question Answering (QA) system that would answer any query raised would solve this need. Studies in Artificial Intelligence (AI) have matured the task of field specific Question-Answering (QA). To build any end-to-end deep learning based QA system, dataset with a large volume is needed. This kind of QA would help researchers, developers, editors/associate editors of a Journal and also to answer queries for the common-man.

Lot of misinformation and false facts are being spread about COVID-19 which can negatively affect the society in multiple ways. There is a huge chance that people would believe such falsified information. Providing accurate and verified information from medical experts is therefore, a challenge which needs to be solved. Scrapping data from approved sources and building a system that can credibly answer natural language questions from humans requires a lot of processing and thus, a trivial solution does not exist.

2 Related Work

The paper by (Möller et al., 2020) first presented dataset for QA on COVID named as COVID-QA. They made use of the COVID-19 scientific articles. They offered biomedical experts annotated 2019 question-answer pairs related to COVID-19 by considering 147 scientific articles.

They selected RoBERTa architecture (Liu et al., 2019a) and fine-tuned it on the SQuAD dataset as a baseline model.They continued to train their baseline model using their COVID-QA annotations in 5-fold cross validation manner.

The paper makes use of two metrics to evaluate their model namely Exact Match and F1 scores. Compared to baseline score of 21.84 in terms of Exact Match, COVID QA model performed better with a score of 25.90. Similarly in terms of F1 scores, COVID QA model outperforms Baseline with a score of 59.53 as compared to 49.43 of baseline.

From the results it was found that the additional training on this domain-specific data leads to significant gains in performance.The overall scores are pretty low compared to SQuAD. It is assumed that the low scores relate to more complex question/answer pairs on much longer documents and the lack of multiple annotations per question.

3 Method

In this section, we will outline the characteristics of the dataset, the models to be used for addressing the QA task, and the evaluation metrics.

3.1 Dataset

COVID-QA is a QA dataset which was first released in the paper (Möller et al., 2020). It is a SQuAD (Rajpurkar et al., 2016) style annotated dataset which consists of 2019 question answer pairs related to the pandemic. This data was pulled from 147 scientific articles, and was annotated by experts in the biomedical field.

3.2 Models

We will use various pre-trained SOTA transformer models which were discussed during the span of the course. We will perform question-answering NLP task by fine-tuning these models to the COVID-QA dataset. The pre-trained models that we have shortlisted for this task span a variety of transformer architectures which gained popularity over the years after BERT (Vaswani et al., 2017) was first proposed. The models are as below:

- BERT (base, large) - pre-trained on SQuAD 2.0 (Vaswani et al., 2017)
- RoBERTa (base, large) - pre-trained on SQuAD 2.0 (Liu et al., 2019b)
- DistilBERT (base) - pre-trained on SQuAD 2.0 (Sanh et al., 2019)
- ALBERT (base, XXL) - pre-trained on SQuAD 2.0 (Sanh et al., 2019)
- ELECTRA (base) - pre-trained on SQuAD 2.0 (Clark et al., 2020)
- Longformer (base, large) - pre-trained on SQuAD 2.0 (Beltagy et al., 2020)
- BigBird (base) - pre-trained on Natural Questions (Zaheer et al., 2020)

3.3 Evaluation

We will use the standard evaluation metrics for QA NLP task as listed below:

- Exact Match
- F1 Score

4 Conclusion

COVID-19 has impacted the entire world in a significant manner. With no end in sight, misinformation is one of the other plagues that we as a society have to deal with. In such an unprecedented and fragile time, a reliable source for truth is imperative, and this source needs to be able to answer the questions with high availability, correctness, and accuracy.

Our project is an effort to fine tune various models on the COVID-QA dataset, some of which were proposed after the dataset was first publicly released. In the next part of the project, we will explore the performance of various models on this dataset, and try to clearly examine the disparity in performance. We hope to see an increase in performance when using models that attend to longer sequences. In addition to this, we will also be reporting the average response time of the models.

Environmental impact of training these models is a concern which was raised by (Strubell et al., 2019), and became an important part of NLP literature thereafter. It is therefore, another avenue that we would be looking at.

In conclusion, we would like to report the models that perform the best, are the quickest, and are the most efficient for training.

Acknowledgments

We are really grateful to Professor Ndapa Nakashole, Computer Science Department, University of California, San Diego and the Yutong Shao (TA) for their continuous support, encouragement and willingness to help us throughout this project.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question

answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *CoRR*, abs/1906.02243.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *CoRR*, abs/2007.14062.