Covid-19 Question Answering System

Tan Ching Fhen, Yeo Chai Kiat
School of Computer Science and Engineering

Yan Shi Xing, Sunil Sividas, Josey Mathew
NCS Pte. Ltd. (SCALE@NTU)

***Abstract* –** Language model pre-training has a significant effect on model performance on downstream tasks. Many new pre-training techniques have emerged in recent years such as, replaced token detection, sentence order prediction and knowledge distillation. With many variants of BERT developed every year, there can be a dilemma as to which version of BERT to choose when deploying a question answering system. Thus, in this study, we compared the performances of seven BERT architectures on SQuAD 1.1 and CovidQA. Covid-19 rules and regulations in Singapore are continually being adapted to the severity of the viral situation. These rules and regulations come in the form of unstructured text and are very extensive. It is crucial for this information to be easily accessible and adhered to by the public. Thus, in this study, we also built a question answering system that answers covid-19 questions relating to these rules and regulations. The architecture we utilized to build this Covid-19 question answering system is the "retriever-reader" architecture. We use the ElasticSearch with BM25 algorithm as the retriever and the RoBERTa architecture as the reader.

Keywords – 'Retriever-reader' architecture, RoBERTa, Covid-19 QA system

1 Introduction

1.1 Background: Question answering (QA) systems produce an accurate answer, given a natural language question. There are two main approaches: textual QA and Knowledge Base QA. Textual QA extracts answers from unstructured text documents while Knowledge Base QA from extracts answers from a manually constructed Knowledge Base (Zhu et. al., 2021).

1.2 Scope: This study focuses on Textual QA, more specifically, under the task setting Open-domain QA where the system retrieves relevant documents with respect to the question and generates an answer. While there are many QA architectures available, this study will build a covid-19 QA system using the 'retriever-reader' architecture (figure 1). We also seek the answer the following research questions to optimize the performance of the covid-19 QA system:

- Does replaced token detection yield better performance?
- Does cased BERT yield better performance than uncased BERT?
- Does RoBERTa outperform BERT?
- Does ALBERT outperform BERT?
- What effect does distillation have on BERT?
- How does performance change as the number of token input increase?

1.4 Outline: We will first describe and explain the components of the 'retriever-reader' architecture in section 2. In section 3, we will describe how we constructed the CovidQA dataset, which is composed of question-and-answer pairs regarding the covid-19 situation in Singapore. Following that, in section 4, we describe the experiments and findings to answer our research questions. Next, in section 5, we describe fine-tuning of the reader model on CovidQA. Lastly, in section 6, we discuss some limitations of our study and our covid-19 QA system.

2 Architecture

2.1 'Retriever-Reader' architecture:

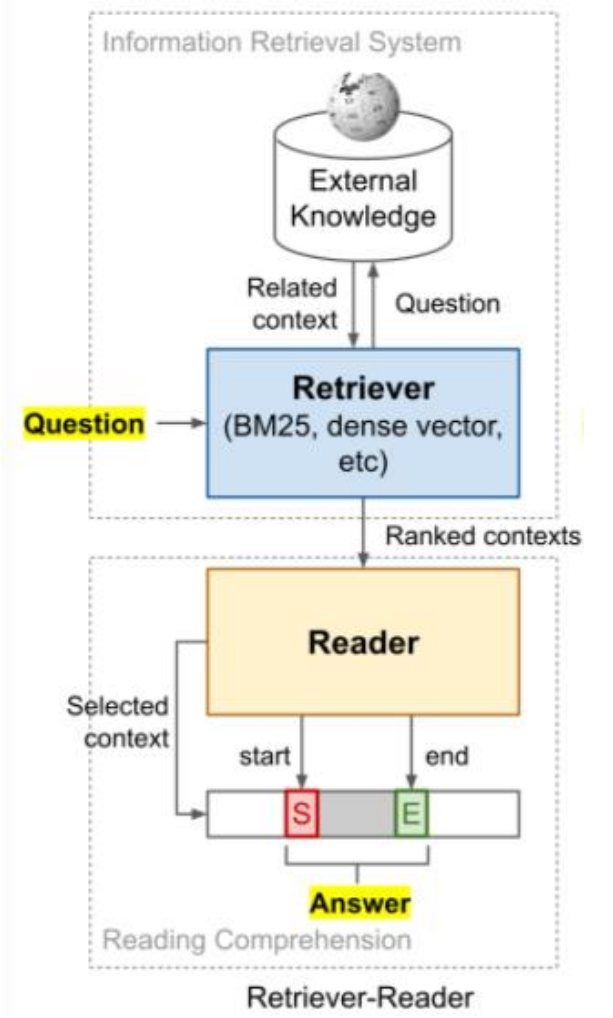

Figure 1. 'Retriver-reader' architecture.

The architecture we utilized to build a covid-19 QA system is the 'retriever-reader' architecture (figure 1). The retriever is responsible for retrieving relevant documents with respect to a given question, while the reader extracts an answer span from the received documents (figure 1). The architecture was first proposed in DrQA (Chen et al., 2017). Since then, many improved variants were developed such as the Multi-passage BERT (Wang et al., 2019) and the Retrieval-Augmented Generation (Piktus et. al., 2021). These QA systems employed the same 'retriever-reader' architecture but differed in the utilization retriever and reader models.

2.2 Retriever: For the retriever, we employed ElasticSearch with BM25 algorithm, a bag-of-words retrieval function.

2.3 Reader: For the reader, we employed RoBERTa (Liu et. al., 2019) because it was found to have the best performance on both SQuAD 1.1 and CovidQA dataset in our experiments. RoBERTa is an improved version of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et. al., 2019), that was pre-trained using dynamic masking, on more data and over a longer duration compared to the original BERT.

BERT and its variants are essentially pre-trained language models that can be fine-tuned to a question answering task. This method of using pre-trained language models and fine-tuning on a natural language task has been shown to be effective (Radford et al., 2018). The following lists the exact names of all pre-trained models that we compared in our experiments:

- albert-base-v1
- bert-base-cased
- bert-base-uncased
- distilbert-base-uncased
- distilroberta-base
- google/electra-base-discriminator.
- roberta-base

3 Covid-QA Dataset

3.1 Documents: To acquire a set of covid-19 related documents to be used as out QA system datastore, we scraped textual data from three reliable websites: straitstimes.com, gov.sg and channelnewsasia.com. These textual data can be covid-19 news articles or covid-19 official rule and regulations documents implemented by the Singapore government. These documents were indexed into ElasticSearch using Haystack, an open-source framework for building search systems.

3.2 QA pairs: To evaluate whether our QA system answers covid-19 questions reliably, we manually constructed and annotated 305 question-and-answer pairs on our covid-19 documents using the Haystack Annotation Tool. This Covid-QA dataset will be used to evaluate our QA system.

4 Experiments and findings

4.1 Motivation: Although the original BERT achieved 87.433 exact match on SQuAD 1.1, exceeding the human performance of 86.831, it still has many weaknesses such as being 'significantly undertrained' (Liu et. al., 2019) and requiring 'large amounts of compute to be effective' (Clark et. al., 2020). This led to the invention of many new pre-training tasks and parameter reduction techniques to improve BERT. It can be difficult to choose between many of these BERT variants, thus we conducted experiments to compare the performances of BERT variants listed in section 2.3.

4.1.1 Fine-tuning: We fine-tuned each pre-trained model on 20000 randomly selected training samples from SQuAD 1.1.

4.1.2 Evaluation: These pre-trained models were evaluated on two sets of validation data: 5000 randomly selected validation samples from SQuAD 1.1 and the entire Covid-QA. The metric we use to evaluate the model performances was F-1 score (figure 2).

$$F1 = \frac{2*Precision*Recall}{Precision + Recall}, where$$

$$Precision = TP/(TP+FP)$$
$$Recall = TP/(TP+FN)$$
$$TP: number\ of\ tokens\ that\ are\ shared\ between\ the\ correct\ answer\ and\ the\ prediction.$$
$$FP: number\ of\ tokens\ that\ are\ in\ the\ prediction\ but\ not\ in\ the\ correct\ answer.$$
$$FN: number\ of\ tokens\ that\ are\ in\ the\ correct\ answer\ but\ not\ in\ the\ prediction.$$

Figure 2. Definition of F-1 score

4.1.3 'Max_length' parameter: After relevant documents are retrieved by the retriever, they are split into shorter passages because BERT has a maximum input sequence length of 512 (Devlin et. al., 2019). Although past research suggests that splitting documents into passages with 100 words seem to work best (Wang et al., 2019), we suspect that this number can differ depending on the model and dataset utilized. Thus, we also evaluate our models over varied 'max_length', ranging from 100 to 400 with increments of 30 to determine the optimal 'max_length'.

4.2 Findings:

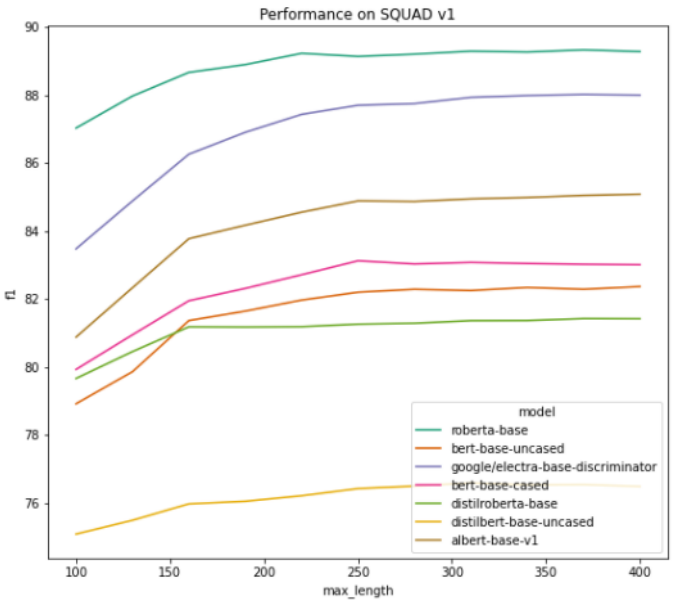

Figure 3. F-1 score against 'max_length' on SQuAD 1.1.

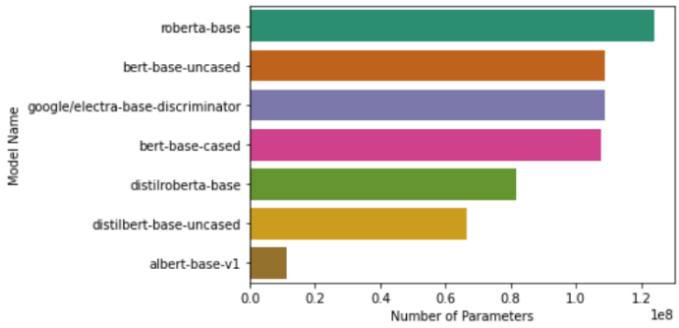

Figure 4 shows the number of parameters of each model

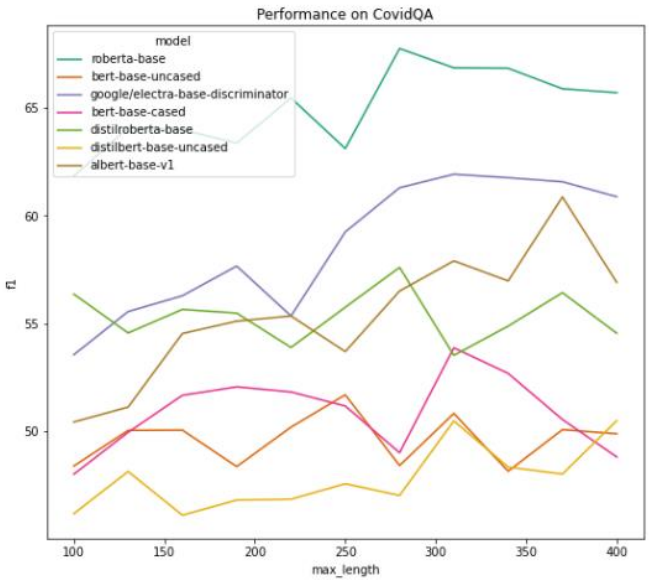

Figure 5. F-1 score against 'max_length' on CovidQA.

4.2.1 Does replaced token detection yield better performance?

The original BERT was pre-trained using masked language modelling (MLM) task – a subset of tokens in the corpus were masked out and the model must predict the masked word. ELECTRA (Clark et. al., 2020) made use of a different pre-training task, replaced token detection – a subset of tokens was randomly replaced, and the model must predict, for every token, whether it was replaced or not. Based on our experiments, electra-base outperforms bert-base-uncased in all 'max_length' parameters (figure 3 & 5) despite having the same number of parameters (figure 4).

4.2.2 Does cased BERT yield better performance than uncased BERT?

Cased BERT outperforms uncased BERT under SQuAD 1.1. However, this was not always true in CovidQA as uncased BERT yields better performance under some 'max_length' settings (figure 5).

4.2.3 Does RoBERTa outperform BERT?

RoBERTa is essentially a BERT variant that was trained model longer, with bigger batches over more data, without next sentence prediction objective and while dynamically changing the masking pattern applied to the training data (Liu et. al., 2019). Based on our experiments, RoBERTa outperforms all other BERT variants (figure 3 & 5). However, it is to be noted that this improved performance also could be attributed to greater number of parameters (figure 4).

4.2.4 Does ALBERT outperform BERT?

To address memory limitations and longer training times, two parameter reduction techniques were utilized to create ALBERT (Lan et. al., 2020). They were factorized embedding parameterization and cross-layer parameter sharing, which improved parameter efficiency and led to models that scale better. Based on our experiments, ALBERT outperforms BERT (figure 3 & 5) despite having a lot less parameters (figure 4).

4.2.5 What effect does distillation have on BERT?

Knowledge distillation is a "compression technique in which a compact model – the student – is trained to reproduce the behavior of a larger model – the teacher -or an ensemble of models" (Sanh et. al., 2020). The goal of distillation is to reduce model size and increase training speed while retaining as much performance as possible. Based on our experiments, both distilroberta-base and distilbert-base-uncased saw a large decrease in F-1 score when compared to their teachers, roberta-base and bert-base-uncased (figure 3 & 5).

4.2.6 How does performance change as the number of token input increase?

In SQuAD, performance increases steadily as 'max_length' parameter increases (figure 3) due to more context available for the reader to accurately predict the answer span. This increase plateaus at around 250 for all models. On the other hand, this effect is not always true on CovidQA. Increasing 'max_length' from 100 to 300, we see increase in F-1 score for only some models like roberta-base and electra-base-discriminator. On the contrary, increasing 'max_length' beyond 300 causes F-1 score to decrease for some models like

distilroberta-base and bert-base-uncased (figure 5). We suspect that this drop in performance could be attributed to other plausible answers appearing as more text is available.

5 Further fine-tuning

5.1 Fine-tuning and validation on CovidQA: Training of models is computationally intensive. Thus, for our final reader model, we leveraged on a fully fine-tuned reader, deepset/roberta-base-squad2, available at huggingface.co/models. Next, we further fine-tuned this model on 70% of CovidQA and validated on the remaining 30%, in batches of 5 and over 5 epochs. The log-loss and F-1 scores are shown below:

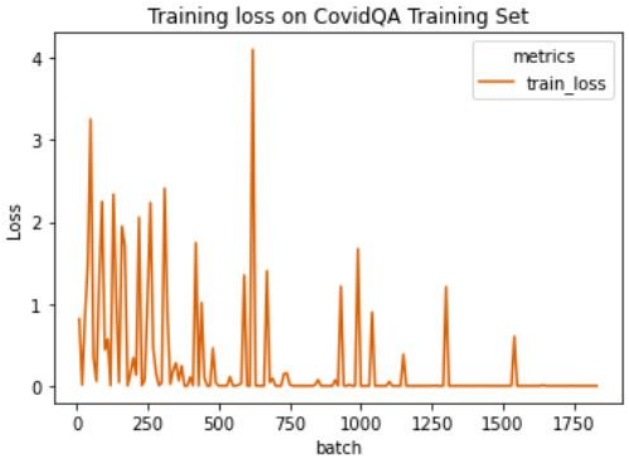

Figure 6. Training loss over 5 epochs.

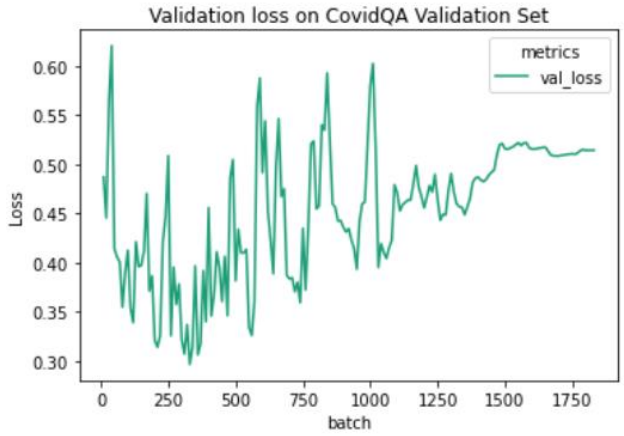

Figure 7. Validation loss over 5 epochs

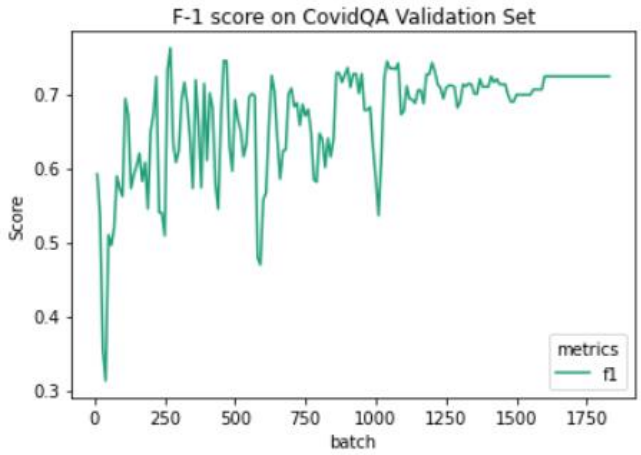

Figure 8. F-1 score over 5 epochs

F-1 score plateaued around the 300th batch which is around the 1st epoch (figure 7). The validation loss starts also starts to increase around the 300th epoch (figure 6). This suggests that, fine-tuning beyond 1 epoch only led to overfitting. Thus, our final reader model will utilize deepset/roberta-base-squad2, fine-tuned over only 1 epoch on CovidQA.

5.2 Predictions on unseen questions:

	Question	Answer
0	when was the official launch of covid 19 vaccination for women?	June 14
1	what is the likelihood of getting serious illness due to moderna vaccine?	0.004 per cent
2	how fast can ART kits give results?	less than 20 minutes
3	why do some seniors avoid taking the vaccine?	fear of complications or side effects
4	how much does nasal swab cost??	$10
5	what was MOH's response after the Victoria Junior College student contracted...	quarantined 95 students and eight staff
6	what is a serology test	detects the presence of antibodies and can show if the person might have bee...
7	what is the capacity of recreational facilities?	50
8	what are the benefits of COVID-19 Driver Relief Fund?	$500
9	what is the capacity of live performances without PET?	50 pax

6 Limitations

Due to limitations of time and resource, there are numerous other pre-trained models that were not investigated in our experiments.

In addition, distil-models like distilroberta-base and distilbert-base were disadvantaged in terms of the number of parameters, thus our experiments do not portray the full potential of distillation and parameter reduction techniques.

Lastly, the CovidQA dataset is small, so fine-tuning on covid-19 related questions is limited. Furthermore, the questions were manually written by our research team, which may not be representative of actual questions asked by the public. Future studies could collect genuine questions about covid-19 rules and regulations from the public so that the QA system can be fine-tuned on greater numbers of and more realistic QA pairs.

7 Conclusion

We compared the performances of seven BERT variants on SQuAD and CovidQA. Based on our experiments, roberta-base model is the best performing model in terms of F-1 score. We build a Covid-19 QA system using ElasticSearch with BM25 algorithm as the retriever and deepset/roberta-base-squad2 as the reader: chingfhen/URECA-Covid-19-Question-Answering-Research (github.com). For our covid-19 datastore, we scraped covid-19 documents from three reliable websites and indexed the documents in ElasticSearch.

Future studies could explore better retriever models such as the dense passage retriever (DPR) which outperforms Lucene-BM25 system by 9-19% (Karpukhin et. al, 2020). Furthermore, questions regarding covid-19 rules and regulations can be collected from the public so that the reader can be fine-tuned on more realistic training data.

Acknowledgement

I would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project.

This project was conducted at Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU), which is a collaboration between Singapore Telecommunications Limited (Singtel) and Nanyang Technological University (NTU) that is supported by A*STAR under its Industry Alignment Fund (LOA Award number: I1701E0013).

References

Danqi Chen, Adam Fisch, Jason Weston and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. arXiv:1704.00051v2.

Kevin Clark, Minh-Thang Luong, Quoc V. Le and Christopher D. Manning. 2020. Electra: Pre-training Text Encoders as Discriminators rather than Generators. arXiv:2003.10555v1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. arXiv:2004.04906.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma1 and Radu Soricut. 2020. Albert: A Lite Bert for Self-supervised Learning of Language Representations. arXiv:1909.11942v6.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401v4.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Victor Sanh, Lysandre Debut, Julien Chaumond and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108v4.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati and Bing Xiang. 2019. Multi - passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering. arXiv:1908.08167v2.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria and Tat-Seng Chua. 2021. Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering