BIA667: QUESTION ANSWERING SYSTEM FOR COVID 19

– Question Answering and Text Summarization

Yiyi,[1] Yili Wang,[1] Shengjie Liu,[1]
[1]Stevens Institute of Technology
yyi2@stevens.edu, ywang387@stevens.edu, sliu88@stevens.edu

ABSTRACT

In this study, in order to address the need for specific information in this COVID-19 pandemic, we adopt a deep learning based system which uses state-of-art NLP(natural language processing) technique in Question Answering (QA) and text summarization for mining the available medical literature. In our system, Information Retrieval (IR) module and QA technique are applied to extract the relevant paragraph given a query from user. Then, elegant summaries are provided by using text summarization to help people quickly understand answer.

1 Introduction

During this COVID-19 pandemic, many scholarly articles concerning this virus are published throughout the world. Meanwhile, there is a huge surge of the query request demand for this virus. Therefore, a COVID-19-specific QA system which can provide people efficient answers is crucial, especially for people from the medical field who seek the solution to treat patient and find the cure.

To build this kind of system, we propose a deep learning based system which uses state-of-art NLP technique in QA and text summarization for mining the available medical literature. This is a network based system which can answer questions related to COVID-19 such as the problems from $this\ Kaggle\ Project$. Through our system, the user can get these two answers

- top n relevant paragraphs to the query from user
- elegant summaries based on the returned relevant paragraphs. They are provided by using text summarization technique

Our system is composed of three parts (1) Information Retrieval (2) Question Answering (3) Text Summarization. The first part is to find the relevant publications given the query from user; the second part is to find the top n relevant paragraphs given the selected publication and then the third part will return the fluent summaries based the selected paragraphs to the user.

2 Related Work

Since the release of $COVID-19\ Research\ Dataset$, there are many systems built to make researchers and public to know the valuable information related to COVID-19. For example, CORD-19 Search is a search engine that utilized the CORD-19 Dataset processed by Amazon Comprehend Medical; Covidex makes use of multi-stage search architectures which can extract different features from data.

However, the above systems such as CORD-19 are pure search engine. In our system, we not only make the search engine to find the most relevant publications, but also find the top n relevant paragraphs and do the text summarization to produce elegant and efficient solutions.

3 Data Source

There are two public sources/data used in this project

- Reference text data in QA system: COVID-19 Open Research Dataset published in Kaggle
- Question Answering Dataset for pretrained Bert: SQuAD 1.0 (the Stanford Question Answering Dataset)

For reference text data, there are total 123513 json files which contain the publication concerning Covid 19, a sample json file can be found via $this\ link$.

4 Data Preprocessing step

There are currently 3 steps for the data preprocessing stage (may add more steps in the final report)

- For each json sample file, only the title and body text will be kept. Title will be served as an information source index and the body text are used to find the relevant paragraph concerning the user's query.
- Keep paragraphs separate rather than concatenating them, since the BasicBertQA and BioBert QA only accepts the input tokens with length less than 521.
- In the baseline BertQA model, the bert-tokenizer is used to coded the input tokens which include "Sep", "CLS" special tokens.

5 Methodology and Technology

The methodology and technology used are consistent with data source, to summarize [1]:

- Document (publication) Retrieval: Elasticsearch is used to created the search engine to retrieve a preliminary candidate set of publications. Elasticsearch uses Lucene indexing to create an easy-to-understand information retrieval module.
- Question Answering System: Baseline basic Bert and advanced BioBERT [2] (a QA model which is finetuned on the SQuAD and medical dataset) will be applied in our question answering part. Instead of finetuning on COVID-19 articles, we focus on maintaing the general ability of this network and perform zero-shot QA.
- Text Summarization: This part is based on the pre-trained language model Bart [3]. For each query, we generate a summary for the top n paragraphs from QA module and then pass to the user.

6 Preliminary Results

We have successfully import the json files into elasticsearch and it will return relevant publications based on specific user query (details included in the document retrivial jupyter notebook). For the Basic Bert QA model, given the questions "What do we know about COVID-19 Covid risk factors?" and the reference text relevant with pregnancy and covid 19 which is returned by elasticsearch, the model gives us back the answer as follows

```
Answer:   we observed that ob ##ste ##tric intervention may influence the clinical course of the disease .
```

where the # here is caused by bertokenizer when it come across some unknown tokens. If we can further do text summarization, the result will be better.

7 Evaluation Metric

The evaluation metric is two-fold:

- Comparision with official answers provided by Kaggle will be used in the Question Answering part
- Rogue Score will be used in the text summarization part

References

[1] SU Dan, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham Barezi, and Pascale Fung. Caire-covid: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management. 2020.

[2] 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, 2020.

[3] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.