COVID-19 press conference search engine using BERT

Anonymous ACL submission

Abstract

There have been multiple press conferences concerning covid-19, where governments express their work in fighting the virus. These briefings provide the reporters from different agencies with the platform for their questions to be answered. This work studies multiple press conferences from different governments and agencies, ranging from WHO to the Whitehouse to different state governors to even different governments. This work collects the transcripts of these press conferences, then using a custom mechanism, selects short exchanges between different speakers, hence selecting questions asked by reporters and their answers. A search engine is then built on these questions by fine-tuning the state-of-the-art BERT language model to the collected covid-19 press conference transcript dataset. Then a recently proposed architecture called SBERT (Reimers and Gurevych, 2019) is used to select the best embedding. Then with cosine similarity, a search engine is built capable of getting the most similar questions asked by press from a custom query. Our goal by this work is to help organize the press questions concerning covid-19 to help build an insight on the different efforts being taken to combat the virus.

1 Introduction

Press conferences are the channel of communication that governments/agencies use to communicate their efforts in fighting covid-19 with the world. Studying and analyzing the transcripts of these conferences would provide great insights on the different approaches and efforts these governments/agencies use in their fight against the virus.

This work aims to collect the transcript of multiple press conferences made since late January, made by different governments/agencies. Then using a custom mechanism, the short exchanges made throughout these conferences are selected, which mostly contains the questions raised by reporters and how they were answered by officials. The main goal of this work is to study these questions, analyze how they were answered across the different governments, to build an analysis on the different efforts made by countries across the world.

This analysis is made by fine-tuning BERT (Devlin et al., 2018), a state-of-the-art language modeling, to build a customized language model, capable of understanding the context of covid-19 fighting efforts. Then using the recent proposed architecture SBERT (Reimers and Gurevych, 2019) (which builds a mechanism for selecting the most optimized embedding for sentences from BERT), a search engine is built. This search engine gets the most similar questions and their answers (from the built dataset) to a user query.

The paper is structured in the following way: (Section 2) presents how the dataset was collected and the proposed method for selecting the questions and answers in a press conference. In (section 3), we view how BERT (Devlin et al., 2018) was fine-tuned to the collected dataset. (Section 4) views the used architecture for extracting the embeddings from the fine-tuned BERT using the newly proposed SBERT (Reimers and Gurevych, 2019). (Section 5) views some results on running the search engine. We have used google colab for scrapping, fine-tuning and building our search-engine, the code 1 is provided as jupyter notebooks to run seamlessly on google colab. The data 2 is hosted on google drive to connect seamlessly with google colab.

This work opens the opportunity to analyze how a certain question is answered across the different governments, hence building insights on the different fighting efforts being made across the world.

2 Building COVID-19 Press Dataset

This work uses the transcripts provided by REV


REV provides the transcripts of the press conferences made by multiple governments and agencies, these are:

- World Health organization press briefings
- United Kingdom Coronavirus briefing
- White house press conferences
- Justin Trudeau Canada COVID-19 Press Conference
- Press Conferences made by multiple US state governors (NewYork, Iowa, Florida, ... and many others)

2.1 Scrapping Transcripts

We have built a customized scraper using python to scrap the exchanges made by different speakers in a given press conference. We have scrapped 654 press conferences made since 23th January, till 12th May. We were able to obtain more than 66k exchanges throughout the collected transcripts. We tend to scrap the transcript text of each speaker, with the name of that speaker, with the timing of when this exchange was spoken within the press conference. We also record the name and the date of the press conference in addition to its url (from REV).

Since covid-19 is a continuously evolving situation, we would periodically run our scrapers to obtain the most up-to-date transcripts.

2.2 Custom mechanism for selecting short exchanges

This work aims to build a search-engine on the questions raised by reporters and how they were answered by officials. However selecting these exchanges from the scrapped dataset appeared quite challenging, as REV doesn’t provide a guide on the identity of each speaker, so work must be done in order to try and identify the identity of each speaker.

This work uses a custom mechanism to try and select the identity of the speakers. This mechanism is built over rules of when these speakers begin to speak and the amount their exchanges. The proposed rules are:

- The longest exchanges in the press conference are flagged to be spoken by the official giving the press conference (president, prime minister, governor or a health official).
- The first exchange, is flagged as been spoken by the presenter (the conductor of the conference). This can either be a reporter or the official himself.
- If the main official conducting the conference, mentioned other speakers, those speakers are flagged to be helpers to that official. In most cases these have been found to be either health officials (like in case of Dr Fauci in the white house conferences), or another official (either military or a financial official).
- We are most concerned with flagging the reporter questions. These have been found to be few exchanges in a single transcript made by each reporter (each speaker speaks either once or twice max). When this pattern is found (few exchanges made by a single speaker), these exchanges are flagged to be made by reporters, and are considered to be questions. Then the exchange right after it is flagged as its answer.

Using these rules, the previously collected transcript dataset was flagged with the proposed speakers (either conference-conductor, official, helper, or reporter). Another dataset is then created to only contain the questions and answers that were flagged, this built dataset contained about 8k questions and answers.

3 Fine-Tuning BERT to covid-19-press

BERT (Devlin et al., 2018) has proved itself as the state-of-art architecture for language modeling. It is built as an enhancement to the vanilla transformer (Vaswani et al., 2017). It is built to only contain an encoder structure, and to depend solely on self-attention.

BERT is unique in the approach used in its pre-training, where they used the “masked language model” (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). This approach randomly chooses words from the input text (15% of words), and the training objective is to predict these masked words. This training objective enables BERT to be pre-trained in an unsupervised
manner, where raw text is supplied to BERT, without having labels.

This training objective is also used in its fine-tuning, in our case, the collected dataset (covid-19 press of 66k exchanges) is used as the raw training text to fine-tune the pretrained BERT. Hugging Face (https://huggingface.co/) library was used to fine-tune BERT to the collected dataset, the BERT model provided by google (https://huggingface.co/google/bert_uncased_L-8_H-256_A-4) was used as our pre-trained BERT. Google colab was used as the platform for fine-tuning.

4 BERT to build a search-engine

Using BERT for sentence-pair regression (measuring how similar sentences are to each other, the technique used to build a search engine), proves to be inefficient for multiple reasons.

To begin with, for sentence-pair regression in BERT, the 2 sentences are provided to BERT with a special separator token in between them [SEP]. To build a search engine using this approach, one would need to supply each sentence to BERT (in addition to the query sentence). This would require running BERT each time in deployment for about 8k times (size of the dataset) to get the most similar question and its answer from all the dataset. This is simply unsuitable for building a search engine.

Another approach other than sentence-pair regression is often proposed, which is extracting the sentence embedding from BERT. First running BERT just once on the 8k questions, getting their embedding, and in deployment, just run BERT once on the query and use cosine similarity to get the most similar question and its answer. However this also exposes another disadvantage in BERT, as in BERT no independent sentence-embedding are computed, this makes it challenging to extract a good embedding from BERT (Reimers and Gurevych, 2019).

Multiple approaches were proposed to help extract good embeddings from BERT. ((May et al., 2019),(Zhang et al., 2019),(Qiao et al., 2019)) proposed using the [CLS] token from BERT as the fixed size vector embedding for a sentence. Another approach used by (Reimers and Gurevych, 2019), computes the mean of all output vectors.

In (Reimers and Gurevych, 2019), they trained a Siamese BERT network on SNLI data (Bowman et al., 2015) and on Multi-Genre NLI. They then evaluated different polling approaches to build embedding representation for sentences. Either using [CLS] or by averaging vectors to get [MEAN], they fine-tuned their architecture for classification objective function on the STS benchmark with regression objective function. They concluded that using the [MEAN] polling strategy outperformed that of using [CLS] strategy. This is the reason it was the selected pooling strategy in our work. As our work is considered as unsupervised, we don’t have a method of measuring out the best polling mechanism to be used.

5 Experiments

We experiment our dataset, against some of the most intriguing questions about the spread of the virus, the economical impact. Where first we fine-tuned BERT on the collected 8k questions, saved their embedding using the [MEAN] polling strategy. Then for each query, we run the fine-tuned BERT with the same polling strategy, and using cosine similarity we get the most similar questions asked in the collected press-conferences. We view the most similar question, its answer, and the name of the press conference it was found in.

The following are some examples from the search engine, the top 2 most similar questions and their answers are selected. With the name of the press briefing, its date, and the time within the briefing when this exchange was spoken.

<table>
<thead>
<tr>
<th>Input Sentence</th>
<th>Results</th>
</tr>
</thead>
</table>
| how many tests would we able to perform | Score: 0.8211

**question:** How many tests do we have right now and how many do we need?  
**answers:** I do not have that specific number. I can say that there are several hospitals that are now doing tests...  
**header:** Michigan Governor Gretchen Whitmer Coronavirus News Briefing Transcript  
**date:** Mar 18, 2020 (11:50)
How quickly can we have nurses and doctors being tested for Coronavirus? At all?

Healthcare personnel are being tested for Coronavirus based on CDC guidelines.

San Francisco Mayor London Breed Coronavirus Briefing Transcript

Table 2: Number of tests (cont.)

how the virus was started

The investigation into whether the virus escaped from this lab in Wuhan, how active is that, and when do you expect...

Well, we’re looking at it. A lot of people are looking at it. It seems to make sense. They talk about a certain kind of bat, but that bat.....

Donald Trump Coronavirus Press Conference Transcript April 17

score: 0.7392

Governor, where do things stand with the Unemployment Trust Fund and how long can we expect...

So I actually have been checking in with Director Donna Frio. We feel that it is still financially good shape but obviously...

Michigan Governor Gretchen Whitmer Press Conference Transcript May 7

score: 0.7663

How would you support the unemployed

Table 3: Virus origin

when would life return to normal

Table 4: Returning to normal

Table 5: Unemployed support
(cont).

Score: 0.7434

question: When will the state get the money from the CARES Act, the $600 a week?

answers: So I would need to talk to DEO about when exactly. It’s not supposed to be far out. We may have some, if not, we’re going to get some....

header: Florida Gov. Ron DeSantis Coronavirus Briefing Transcript April 13

date: Apr 13, 2020 (39:16)

Table 6: Unemployed support (cont.)

Input Sentence

do we have enough ventilators

Results

Score: 0.8227

question: Do you find the ventilators in the stockpile? How does that-

answers: That would be from their stockpile, yes. The number of sick cops in the NYPD continues to rise. Is there going to be a percentage...

header: Gov Andrew Cuomo New York COVID-19 Briefing April 3: NY Passes 100,000 Cases

date: Apr 3, 2020 (47:30)

Score: 0.8140

question: Are there going to be enough ventilators for what is coming?

answers: It depends on what you mean by what’s coming. Right now, we have 12,700 ventilators in the stockpile. We will use the stockpile as needed...

header: Dr. Anthony Fauci Coronavirus Interview Transcripts: NIH Immunologist Does Series of Sunday Interviews

date: Mar 15, 2020 (09:34)

Table 7: Ventilators number

As seen in the previous examples, the exchanges that were flagged as questions were indeed questions. This helps indicate that the used mechanisms for selecting questions from the different exchanges were successful.

6 Conclusions

In this work we present a new covid-19 data source, which is the press conference briefings, as a rich source for analyzing different governments' response for fighting the virus. We also present some mechanisms of selecting questions from these press briefings. We have used the state-of-art language models for building a semantic search engine to get the most similar questions from the press briefings.

We believe that this new data source can even be used in other areas of research, to understand and build insights on the different approaches the governments are taking in combating this virus.

References


Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the behaviors of bert in ranking.
