Comparison of CORD-19-trained GPT-2-based Chatbot Responses Across Different Text Semantic Similarity Approaches

David Oniani
Mayo Clinic
Kern Center for the Science of Health Care Delivery
Rochester, MN, USA
oniani.david@mayo.edu

Dr. Yanshan Wang
Mayo Clinic
Division of Digital Health Sciences
Rochester, MN, USA
wang.yanshan@mayo.edu

ABSTRACT

On March 16 (2020), per request of the White House Office of Science and Technology Policy, new COVID-19 machine readable dataset (CORD-19) [1] has been released. Yet, there still was a need for making this dataset more interpretable and understandable for the general audience. Besides, there was an urgent need of fast and efficient information exchange. To accomplish this goal, we built an interactive chatbot that answers questions related to COVID-19. This not only made the information from the CORD-19 dataset more understandable to the general audience, but also put it into real-world use and helped efficiently obtain information concerning COVID-19. We have utilized 774M GPT-2 [2] model and applied transfer learning to retrain the model on the CORD-19 corpus. Ultimately, we have created an interactive COVID-19 conversational chatbot. In order to improve the performance of the chatbot, we have applied 4 different text semantic similarity techniques using pre-trained models including BERT [3], BioBERT [4], and Universal Sentence Encoder (USE) [5] as layers on top of GPT-2-based chatbot. For performance evaluation purposes, we have asked experienced personnel at Mayo Clinic to rate our answers and also present these annotations in the paper. Additionally, we have created a user-friendly interactive web application to be hosted online and made available for anyone to get more information regarding COVID-19. Our work has produced good results in both designing a chatbot that produces high-quality answers to COVID-19-related questions (hence, improving COVID-19 information exchange) and comparing the performance of several embedding generation and vectorization techniques in the biomedical domain.

KEYWORDS

gpt-2, covid-19, cord-19, bert, biobert, universal sentence encoder, tf-idf dataset, nlp, ai, semantic similarity

ACM Reference Format:
David Oniani and Dr. Yanshan Wang. 2020. Comparison of CORD-19-trained GPT-2-based Chatbot Responses Across Different Text Semantic Similarity Approaches. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

COVID-19/Novel coronavirus is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It was first identified in December of 2019 in Wuhan, China, and has since spread globally, resulting in an unprecedented pandemic. As of 17 May 2020, more than 4.66 million cases have been reported across 188 countries and territories, resulting in more than 312,000 deaths. The White House Office of Science and Technology Policy has requested a COVID-19 machine readable dataset, now known as CORD-19. Since its rapid emergence started, many countries have declared the state of emergency and the need for fast and efficient information sharing as well as tracking progress related to COVID-19 became crucial. For addressing these issues, we decided to use CORD-19 for designing a chatbot that would answer questions related to COVID-19. The chatbot would not only help improve information sharing, but also as a knowledge base for COVID-19.

A chatbot is a software which is able to conduct a conversation via text and/or other means. It is designed to emulate the way people communicate and give a sense that the conversation only involves humans. Therefore, cutting edge chatbots can hold a human-level conversation and usually, the performance of the chatbot is measured via the quality of its responses to humans. Chatbots have been much studied in recent years and with the advancements in the fields of artificial intelligence (AI) and natural language processing (NLP), both the the functionality and the performance have seen drastic improvements. Semantic similarity of texts, on the other hand, has been studied for a long time and recent breakthroughs allowed for development of new models such as BERT, BioBERT, and Universal Sentence Encoder. Today, one of the state-of-the art conversational artificial intelligence models is GPT-2. GPT-2 is a pre-trained model, so we have applied transfer learning utilizing CORD-19 for retraining purposes. The resulted chatbot gave irregularly long responses that would not be typical of a human. We have therefore decided to further filter the responses via applying embedding generation algorithms and models such as tf-idf, BERT, BioBERT, and Universal Sentence Encoder (USE) and then using semantic similarity approaches such as cosine similarity and inner product. In other words, we first let a human ask a question and make GPT-2 come up with an answer. We the further processed the response‘ with additional filters and ultimately, applied an embedding generation model for finding the sentences that are most relevant to the question.

Our study has produced a chatbot that is both performant and extensible. Additional layer of filters have shown success in classifying sentences. The chatbot is also able to be retrained and readjusted to new data, in case there are new discoveries and scientific achievements related to COVID-19. Furthermore, chatbot responses have been annotated by experienced staff at Mayo Clinic and the resuts were consistent across the annotators.

We will proceed by discussing the related work and the efforts that have been made previously. A few sections will be dedicated to specific approaches and materials used in this work. We will discuss the chatbot evaluation strategy as well as present the annotation results. Finally, we will also discuss the web-based tool designed to be used with this chatbot the source code of which (alongside with everything discussed) has been made available online.

2 RELATED WORK

In the GPT-2 domain, some efforts were made by Wang et al. [6] where GPT-2 was retrained in order to automatically generate answers to AD-related consumer questions posted by caregivers and evaluate how good AI is at answering those questions. Lee and Hsiang [7] have fine-tuned GPT-2 for generating patent claims. Klein and Nabi [8] have applied GPT-2 in conjunction with BERT for automatic questin generation purposes. We are unaware of the work which applied 775M GPT-2 model for transfer learning purposes on the CORD-19 dataset.

In regard to the work related to comparing pre-trained AI models, Jin et al. have made some efforts conducting probing experiments and comparing BERT, ELMo [9], and BioBERT. Sharma and Daniel [10] compared the performance of BERT networks to that of FLAIR [11].

In the general artificial intelligence based chatbot domain, Serbal et al. [12] have applied deep reinforcement learning for building a conversational AI chatbot. Adiwardana et al. [13] have developed a multi-turn open-domain chatbot trained end-to-end on data mined social media conversations. Yin et al. [14] have developed a deep learning based chatbot for psychological therapy purposes.

3 MATERIALS

We harvested the data from the initial *commercial use subset* of COVID-19 Open Research Dataset (CORD-19) [15] containing 9000 scholarly articles in the form of JSON files. We extracted the abstract and the main body of the article from every JSON file, combined them together, and used as a corpus for retraining the GPT-2 model.

4 METHODS

We applied a hybrid approach for generating responses. The answer for the question first gets generated by GPT-2. Then some filtering is done by handling irregular patterns, extra spaces, and redundant punctuation. Additional filtering is done by leaving only the sentences that are most semantically similar to the question, usually eliminating a large portion of the response that would otherwise still be present and potentially make the response confusing. Such hybrid approach to the response generation produced high quality responses to COVID-19-related questions. Figure 1 illustrates this approach.

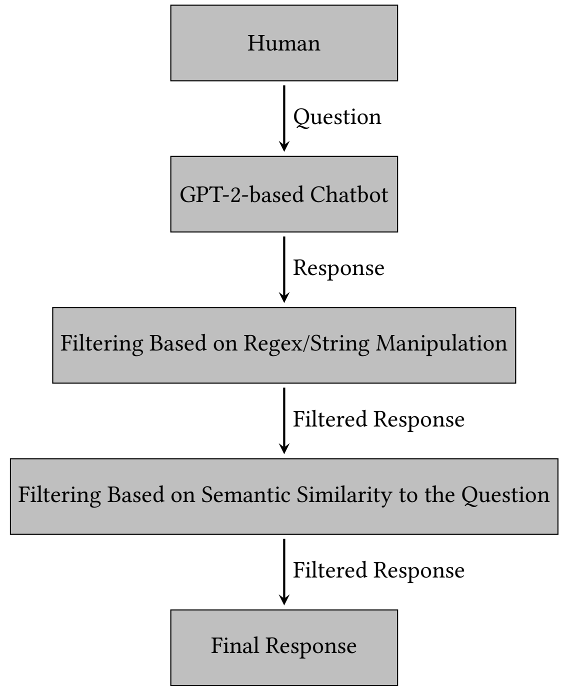

Figure 1: Response Generation.

4.1 Transfer Learning

We utilized GPT-2 774M model and ran transfer learning for 2500 iterations with the batch size of 8. We used Adam [16] as the optimizer and set the learning rate of 0.0001. The link for downloading the model is available on our GitHub page.

GPT-2 has a Transformer [17] based architecture which, in many ways, is similar to Open AI GPT model[2][18]. There are 4 different versions of GPT-2. It features models with 124 million, 355 million, 774 million, and 1.5 billion parameters. We have utilized the version with 774 million parameters. The original GPT-2 was written in tensorflow [19] and this is the version we have used. That said, for retraining purposes, we have utilized the TPU-trainable version of the GPT-2 [20].

Adam an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments [16]. It is highly memory-efficient and has shown good results in retraining our chatbot. We have also tried SGD [21], yet Adam has shown a better performance and therefore, we have released the Adam-based GPT-2 retrained model.

For building a web interface for our chatbot, we have used Python's package flask [22] and integrated it with tensorflow. This resulted in a simple and user-friendly user interface.

4.2 Semantic Similarity

Semantic similarity is a metric that quantifies the degree to which two texts or text documents are similar to each other. The two approaches we have used include cosine similarity and inner product.

We used these approaches as an additional layer to the language model.

Cosine similarity is one of the most commonly used approaches in calculating semantic similarity of texts. Therefore, it is naturally employed in natural language processing tasks. Many NLP applications need to compute the semantic similarity between two short texts. Its flexibility allows one to apply it under virtually any settings, as long as documents can be represented as vectors. Besides, finding cosine similarity is usually not a time-consuming task and can be done really quickly. Therefore, it is also commonly used for benchmarking purposes [23].

The GPT-2 responses are usually very lengthy and for the most part, the answer is not relevant to the question. To solve this problem, we chunked the answer into separate sentences and found the ones that are most *semantically similar* to the question asked. For this task, we have tested and applied 4 different approaches:

- BioBERT large v1.1 (+PubMed 1M) model based on BERT-large Cased (custom 30k vocabulary)
- Universal Sentence Encoder (USE), version 3, large
- Bert-Large, uncased (24 layers and 340M parameters)
- Scikit learn's Tfidfvectorizer

Once the embeddings were generated, we have applied cosine similarity and in some cases, inner product and took 5 most similar sentences. Some additional regex and text filters were added for fixing punctuation and other grammar-related errors.

5 QUESTIONS AND EVALUATION

In order to evaluate the performance of the approaches as well as the overall performance of the chatbot, it was important to have the questions that both are frequently asked and linked to COVID-19. For this purpose, we decided to use 12 questions from the Kaggle's COVID-19 Open Research Dataset Challenge (CORD-19) [24]. Most of the questions included a term "COVID-19," but some did not, in which case we appended the term to the end of the question. Figure 2 presents all 12 questions.

Number	Question
#1	Are there geographic variations in the mortality rate of COVID-19?
#2	What is known about transmission, incubation, and environmental stability of COVID-19?
#3	Is there any evidence to suggest geographic based virus mutations of COVID-19?
#4	Are there geographic variations in the rate of COVID-19 spread?
#5	What do we know about virus genetics, origin, and evolution of COVID-19?
#6	What has been published about ethical and social science considerations of COVID-19?
#7	What has been published about medical care of COVID-19?
#8	What do we know about diagnostics and surveillance of COVID-19?
#9	What do we know about COVID-19 risk factors?
#10	What has been published about information sharing and inter-sectoral collaboration of COVID-19?
#11	What do we know about vaccines and therapeutics of COVID-19
#12	What do we know about non-pharmaceutical interventions of COVID-19?

Figure 2: 12 Questions from Kaggle.

For each approach, we generated 5 different answers for the same question, resulting in the total of 240 answers. Additionally, we made all of the answers publicly available on GitHub [25]. We then asked experienced medical personnel at Mayo Clinic to rate these answers. The rating system was composed of 5 different categories:

Category	Description	Point(s)
Relevant	The answer partially or fully answers the question and/or makes clear attempts to do so and is related to the question	5
Well-formed	the answer makes a logical sense and is somewhat related to both the question and COVID-19, yet it does not (partially or fully) answer the question	4
Informative	The answer is not related to the question, but provides some information about COVID-19 and makes a logical sense	3
Acceptable	The answer makes some logical sense and is weakly related to the question or COVID-19, but is mostly difficult to understand	2
Poor	the answer is totally unrelated to the question or COVID-19 and/or does not make a logical sense	2

Figure 3: 5 Rating Categories.

Having 5 categories allowed for a flexibility of opinions and a broad range of scores which ultimately gave us a better way to evaluate our chatbot.

6 RESULTS

6.1 General Performance.

Annotator	Approach	Score
#1	TfidfVectorizer + Cosine Similarity	3.909
#2	TfidfVectorizer + Cosine Similarity	3.727
Average	TfidfVectorizer + Cosine Similarity	3.818
#1	BERT + Cosine Similarity	4.182
#2	BERT + Cosine Similarity	4.309
Average	BERT + Cosine Similarity	4.246
#1	BioBERT + Cosine Similarity	4.167
#2	BioBERT + Cosine Similarity	4.093
Average	BioBERT + Cosine Similarity	4.130
#1	Universal Sentence Encoder (USE) + Inner Product	3.696
#2	Universal Sentence Encoder (USE) + Inner Product	4.125
Average	Universal Sentence Encoder (USE) + Inner Product	3.911

Figure 4: Comparison of 4 Different Approaches Across the Annotators.

The first annotator has rated BERT as the best approach with the average score of 3.909. BioBERT has shown roughly the same performance (4.167), with BERT rated slightly above. tf-idf procedure performed well, yet could not outperform neither BERT nor BioBERT. USE had the worst performance out of all embedding generation techniques with the score of 3.696 out of 5.

The second annotator, similarly, has given the highest average score to BERT (4.309). USE comes the second with the score of 4.125 followed by BioBERT with approximately the same score (4.093). tf-idf vectorization has yielded the worst results.

In general, the results show the consistency across the annotators. BERT and BioBERT have shown the best performance, with BERT being the leader. Their average scores were 4.246 and 4.130 respectively. tf-idf (3.818) and USE (3.911), on the other hand, have shown roughly similar, yet inferior to BERT and BioBERT, performance. All four approaches, on average, can be considered to be in the "well-formed" category with BERT and BioBERT being close to the "Relevant" category.

7 WEB APPLICATION

In order to make the chatbot more accessible to the general audience as well as for automating the response generation and facilitating information exchange, we have built a web application for the chatbot [26]. The application is powered by Python's flask [22] package and gives a simple and user-friendly means for an interactive communication with the chatbot. In other words, any person could use this application to learn more about COVID-19 and get answers to their questions on this topic. The image below shows the web interface for the interactive chatbot.

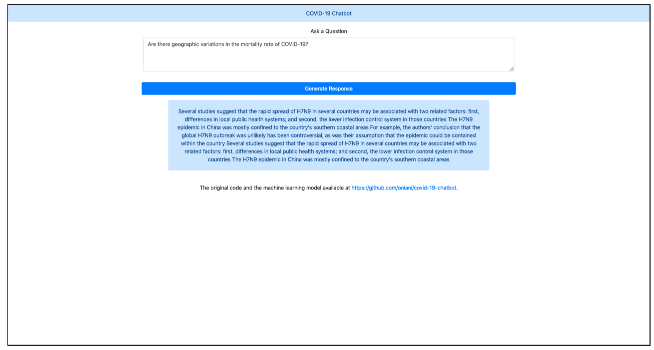

Figure 5: An Image Depicting the Interactive Web Chatbot.

8 CONCLUSION

We have explored the feasibility of utilizing GPT-2 for answering questions related to COVID-19. GPT-2 was retrained on the CORD-19 corpus featuring thousands of scholarly articles. Some of the example answers to the COVID-19-related were made available online. Four different embedding generation techniques (tf-idf, BERT, BioBERT, and Universal Sentence Encoder (USE)) were compared. The results were annotated by experienced medical staff at Mayo Clinic. The annotation results were consistent and concluded that BERT and BioBERT have the best average performance. USE came next with tf-idf showing the worst average performance.

REFERENCES

[1] The White House Office of Science and Technology Policy. 2020. Call to action to the tech community on new machine readable covid-19 dataset. https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset.

[2] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

[3] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: on the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, (September 2019). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. https://doi.org/10.1093/bioinformatics/btz682.

[5] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. eprint: arXiv:1803.11175.

[6] Yanshan Wang, Krishna B. Soundararajan, Sunyang Fu, Luke A. Carlson, Rebecca A. Smith, David S. Knopman, and Hongfang Liu. 2019. How good is artificial intelligence at automatically answering consumer questions related to alzheimer's disease? eprint: arXiv:1908.10678.

[7] Jieh-Sheng Lee and Jieh Hsiang. 2019. Patent claim generation by fine-tuning openai gpt-2. eprint: arXiv:1907.02052.

[8] Tassilo Klein and Moin Nabi. 2019. Learning to answer by learning to ask: getting the best of gpt-2 and bert worlds. eprint: arXiv:1911.02365.

[9] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. eprint: arXiv:1802.05365.

[10] Shreyas Sharma and Ron Daniel Jr. 2019. Bioflair: pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. eprint: arXiv:1908.05760.

[11] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, 1638–1649.

[12] Iulian V. Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Rajeswar, Alexandre de Brebisson, Jose M. R. Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2018. A deep reinforcement learning chatbot (short version). eprint: arXiv:1801.06700.

[13] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. eprint: arXiv:2001.09977.

[14] Junjie Yin, Zixun Chen, Kelai Zhou, and Chongyuan Yu. 2019. A deep learning based chatbot for campus psychological therapy. eprint: arXiv:1910.06707.

[15] 2020. Covid-19 open research dataset (cord-19). https://pages.semanticscholar.org/coronavirus-research.

[16] Diederik P. Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. eprint: arXiv:1412.6980.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need, 5998–6008. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[18] Alec Radford and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

[19] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. (2015). http://tensorflow.org/.

[20] Shawn Presser. [n. d.] https://github.com/shawwn/gpt-2/.

[21] Kiefer J. Kingma and Wolfowitz J. 1952. Stochastic estimation of the maximum of a regression function. https://projecteuclid.org/euclid.aoms/1177729392.

[22] [n. d.] Welcome to flask – flask documentation. https://flask.palletsprojects.com.

[23] Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Y. Hammerla. 2019. Correlation coefficients and semantic textual similarity. eprint: arXiv:1905.07790.

[24] Kaggle. 2020. Covid-19 open research dataset challenge (cord-19). https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks.

[25] 2020. Annotation results and generated answers on github. https://github.com/oniani/covid-19-chatbot/tree/master/annotations.

[26] Interactive Chatbot Web Application Based on Flask. 2020. Covid-19 open research dataset challenge (cord-19). https://github.com/oniani/covid-19-chatbot/tree/master/web-app.