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

David Oniani
Mayo Clinic
Rochester, MN, USA
oniani.david@mayo.edu

Dr. Yanshan Wang
Mayo Clinic
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. We have utilized 774M GPT-2 [2] model and applied transfer learning to retrain the model on this corpus. Ultimately, we have created a COVID-19 conversational chatbot. In order to improve the performance of the chatbot, we have applied 4 different text semantic similarity techniques using pretrained models including BERT [3], BioBERT [4], and Universal Sentence Encoder (USE) [5] as additional layers on top of GPT-2-based chatbot responses. We present the results annotated by experienced medical personnel at Mayo Clinic.

KEYWORDS

gpt-2, covid-19, cord-19, bert, biobert, universal sentence encoder, 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, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Chatbots have been much studied in recent years and with the advancements in the fields of artificial intelligence and natural language processing, 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. The paper takes an approach which is a marriage of these two and brings a different perspective on the chatbot creation. We first let a human ask a question and make GPT-2 come up with an answer. Then we further process the answer with additional filters and ultimately, apply a different model for finding the sentences that are most relevant to the question.

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

2 CORPUS

We harvested the data from the initial *commercial use subset* of COVID-19 Open Research Dataset (CORD-19) [6] 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.

3 TRANSFER LEARNING

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

4 SEMANTIC SIMILARITY

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

5 RESULTS

We have generated over 300 answers to the questions. We have asked experienced medical personnel at Mayo Clinic to rate our question.

Approach	Score
TfidfVectorizer + Cosine Similarity	TBD
BERT + Cosine Similarity	TBD
BioBERT + Cosine Similarity	TBD
Universal Sentence Encoder (USE) + Inner Product	TBD

Figure 1: Comparison of 4 Different Approaches.

6 CONCLUSION

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean porta purus et sem gravida rutrum. Maecenas blandit nulla ac luctus tempus. Nam finibus posuere ante, et lacinia massa vestibulum sit amet. Nulla velit arcu, efficitur quis turpis nec, sollicitudin lobortis nisi. Vivamus ut diam ut eros faucibus fringilla. Suspendisse pellentesque magna nec velit tristique sollicitudin. Morbi ultrices nec augue et molestie. Nam sapien ante, ullamcorper elementum convallis id, faucibus in lectus. Fusce pellentesque mollis velit efficitur porta. Sed finibus ligula quam, et lacinia velit posuere auctor. Donec ligula lorem, dictum nec lectus in, vehicula tincidunt massa. In hac habitasse platea dictumst.

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