CORD-19 Answer generation: BERT Based Question-Answering System on Reduced CORD-19 Dataset

Nicolas Abule
nabule@torontomu.ca
Department of Computer Science
Toronto Metropolitan University

Jason Chen
jason1.chen@torontomu.ca
Department of Computer Science
Toronto Metropolitan University

Abstract

In this study, we explore the effectiveness of fine-tuning a BERT-based model for Question Answering (QA) tasks using a reduced CORD-19 dataset (specifically looking at documents pertaining to comorbidity). Our approach involves two key components: 1) A Document Retrieval System based on Cosine Similarity, which efficiently retrieves relevant text partitions from the scientific papers, and 2) A BERT model to generate the resulting answer component from the retrieved information. Our results demonstrate the potential of BERT models in extracting answers from biomedical literature, particularly in the context of COVID-19 research. Additionally, we compare our fine-tuned BERT model with the "bert-large-uncased-whole-word-masking-finetuned-squad" model and COBERT model, showing how the latter models outperformed the fine-tuned model but nonetheless had middling results in creating satisfactory answers.

1 Introduction

BERT (Bidirectional Encoder Representation) [Devlin et al., 2019] is a Language representation model that massively transformed the field of natural language processing in 2018. Developed by GoogleAI, BERT is a model based around bidirectional pre-training, a feature that no other model attempted before, as other models read left-to-right or vice versa. BERT is pre-trained with 2 methods, Masked Language Model and Next Sentence Prediction. Coupled with this and its ability to be fine-tuned with only one output layer, BERT produces state-of-the-art results that massively outperform prior models [Wang et al., 2020]. On March 16, 2020, a massive COVID-19 dataset under the name of CORD-19 was released to the public. This release invited data analysts and natural language processing specialists to further develop the information gathering and language processing of medical metadata. This resulted in over 200,000 downloads in 2 months and 4 years later, an expedited development in question answering systems and specialized language models meant for medical data [Wang et al., 2020]. Some examples of these medical language models are bioBERT and sciBERT. We use this dataset in a reduced form such that it contains papers that relate to comorbidity reducing the size of the dataset by around 80%. In this paper, we will be examining our language model created specifically with biomedical papers, and comparing their effectiveness to the standard BERT QA system.

2 Formal Description

This approach uses a fine-tuned BERT model to find the most relevent information within the context retrieved from a document in the corpus with the highest cosine similarity with the question asked. The fine tuning was done on the pre-trained "bert-base-uncased" model and from that 3 epochs of training using the AdamW algorithm achieving a validation loss

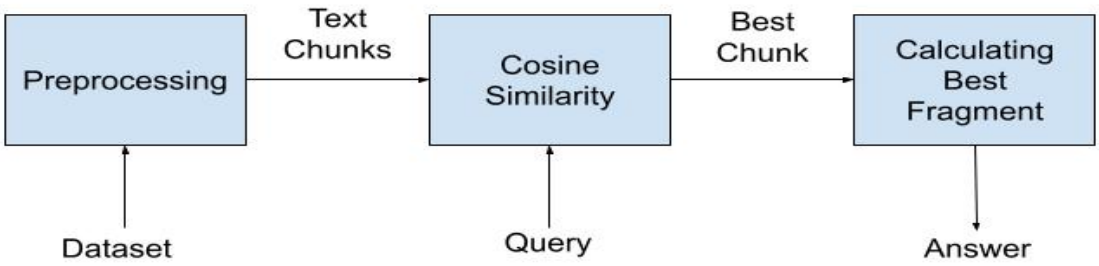

Figure 1: Computational flow of creating answers from given Query

of 2.38.e^{-06}. The retrieval of the document is gotten through iterating through every document in the corpus C and calculating its respective cosine similarity S_c with the question Q and returning its maximum: *max(for all c in C (S_c(Q,c)))*. The sentence fragment picked is found through finding the word with the highest probability being the start of an answer span. The best starting word is calculated through the dot product of S and T_i divided by the softmax of $S.T_j$ where S is a starting vector and T_i is a hidden vector for the i^{th} input token. *max(for word in Words ($e^{S.T_i}$ / $sum_j(e^{S.T_j})$))* (Devlin et al, 2019). The ending word is calculated through trying to maximize $S.T_i + E.T_j$ where E is a end vector and $j > i$. *max(for word in Words ($S.T_i + E.T_j$))* [Devlin et al., 2019]. A diagram of this can be seen in Figure 1.

3 Related Works

Creating QA systems for the medical industry was a long-sought topic, with papers reporting back all the way from 2003. Pierre Jacquemart and Pierre Zweigenbaum used traditional NLP concepts (NER) and information retrieval methods to grab required information [Jacquemart and Zweigenbaum, 2003]. Paired with an outline of every question structure, this managed to answer questions within a database.

	Semantic model
1	[which X]–(r)–[B]
	[A]–(r)–[which Y]
2	does [A]–(r)–[B]
3	why [A]–(r)–[B]

Figure 2: The Semantic Model of a 2003 question-answering system. [Jacquemart and Zweigenbaum, 2003]

In 2008, Andrea Andrenucci had attempted to make a medical QA system using 3 approaches, answering questions open ended questions requested by a user [Andrenucci, 2008]. Though somewhat successful, each approach had glaring flaws. Deep NLP suffered from a long calculation time due to complex structure. Information Retrieval and shallow NLP heavily relied on redundant sentence structures, recognizing answers only through semantic similarity. Lastly, template based questions depended on manually created QA pairs which are used by the system to answer similar questions. With the release of CORD-19, [Tang et al., 2020] tested several models on the dataset with supervised and unsupervised variations. Most of these models were based on BERT, but had different variations from our model, such as being trained on different datasets (compared to this model, which used SQuAD and tensorFlow) or being a different strand of BERT (e.g: bioBERT). This QA system was named CovidQA. Surprisingly, a different model developed by google, T5 Text-To-Transformer was the best model for Text-chat Questions as well as keyword querying. Another CORD-19 based question-answering system named CoBERt [Alzubi et al., 2021] has a model closely resembling the model in this paper, using cosine similarity to find articles with closest answers, and the same SQuAD library to train the BERT model . In addition, Alzubi et al., creates an additional BERT

model named Distil BERT, a simpler model that sacrifices parameters to have a 60 % faster runtime, achieving up to 97 % of original effectiveness. This paper as well as the one above will be the main source of comparison in our next section.

4 Results and Comparisons

Using our QA system, it was found that questions with accurate answers must have a cosine similarity at least larger than 13 %. Questions that were too broad or did not have an exact answer in a paper resulted in a cosine similarity of 0 and an unintelligible answer. Upon using our QA system, queries had to match the exact same keywords as the answers found within the paper's paragraph. For example, if a paragraph contained the phrase "the primary cause of iCU admissions", using the query "what caused the most ICU admissions?" would not work as the QA system could not match the context of "the most" with "primary".

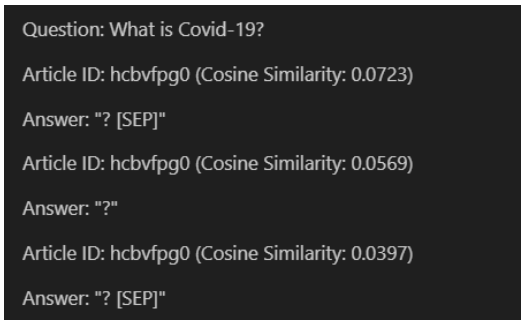

Figure 3: Broad question of Covid-19 results in random answers

4.1 COBERT

COBERT's system was capable of answering questions that had syntactically different sentence structures to the paragraph containing the answer.

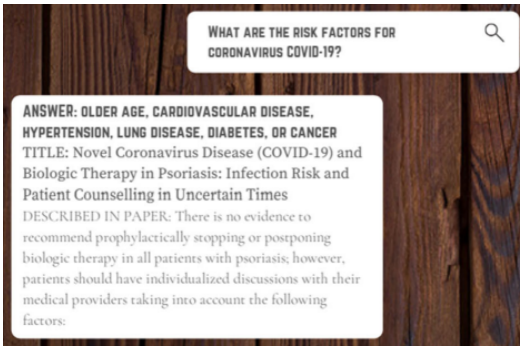

Figure 4: COBERT's QA system able to answer risk factors despite only containing "factors" within paragraph.

From Figure 4, we can see that the COBERT system is capable of matching implied sentences, where keywords of "risk", "covid-19" are implied in the context before the exact paragraph answered. Attempting to answer the question on our system resulted in unintelligible answers as seen in figure 5.

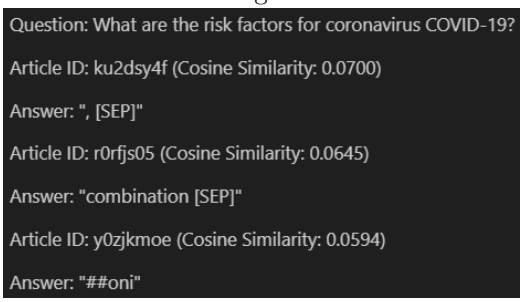

Figure 5: our QA system's attempt at answering Figure x (above figure)'s question

4.2 SQuAD

Looking at "bert-large-uncased-whole-word-masking-finetuned-squad" and comparing that to our fine-tuned model we see that the same papers are picked but often we get different answers. Looking at the question "What is Comorbidity?" we can see two very different answers the former giving "the increased probability of having a certain pair of conditions" and the latter giving "multiple pathologies". The squad model in this example gives a much longer answer and that was typical in testing and also preferable as longer answers were typically better, but both of these answers are reasonable for the task of answering the question.

One thing to note was that both of these models often returned what would be seen in figure 3 and 5 and were not useful for the task.

5 Limitations

While BERT combined with cosine similarity offers a powerful approach for document retrieval, it is essential to recognize its limitations. Firstly, BERT's training on sentence pairs may hinder its generalization to longer texts, particularly when dealing with extensive document corpora. Additionally, the BERT implementation used had a 512 token cap and this limitation could lead to information loss particularly between adjacent sentences. Furthermore, cosine similarity, the way that document retrieval was implemented was not very robust as it would only pick out documents that contains the exact words as in the query and nothing for their variations. Another problem that comes from this is that of semantic differences between different documents or even within the same document in different document fragments. To improve the document retrieval TF-IDF, topic models, knowledge graphs or other methods to capture the semantic information within the texts and compare that to the query would also work well.

Another limitation is in the way the responses were generated specifically in the fact that this was done through a statistical model based on the both the query and the document fragment given to the BERT model. This method often did not give answers unless questions were very specifically worded, and even then these answers would often contain parts of the document fragments that were completely unrelated such as punctuations or other irrelevant words. As BERT is not a generative model the information retrieved from the documents is all that can be returned as an answer which is hardly satisfactory for what would be considered readable answers, but feeding the resulting information into a generative model such as GPT3 or GPT4 would likely give adequate answers. Further training on the model could produce better results in keeping punctuations and other irrelevant words as not suitable as start or end words when doing the statistical analysis.

6 Conclusion

The results from this approach showed promise in usage of cosine similarity for document retrieval but much more must be added for a truly robust model to be built based on it. Creating readable answers through BERT also seems difficult without heavily altering our model to create stops and starts between different sentence fragments and combining them. When looking at the fine-tuned BERT model and comparing it to COBERT or SQuAD models we get disappointing results showing that these broad based models give the same or better responses even from similar contexts. But regardless of that if given good contexts all models perform about the same.

References

[Alzubi et al., 2021] Alzubi, J. A., Jain, R., Singh, A., Parwekar, P., and Gupta, M. (2021). Cobert: Covid-19 question answering system using bert. *Arabian Journal for Science and Engineering*, 48(8):11003–11013.

[Andrenucci, 2008] Andrenucci, A. (2008). Automated question-answering techniques and the medical domain. *Proceedings of the First International Conference on Health Informatics*.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Jacquemart and Zweigenbaum, 2003] Jacquemart, P. and Zweigenbaum, P. (2003). Towards a medical question-answering system: a feasibility study. *Studies in Health Technology and Informatics*, 95:463–468.

[Tang et al., 2020] Tang, R., Nogueira, R., Zhang, E., Gupta, N., Cam, P., Cho, K.,

and Lin, J. (2020). Rapidly bootstrapping a question answering dataset for covid-19.

[Wang et al., 2020] Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A., Wang, K., Wang, N. X. R., Wilhelm, C., Xie, B., Raymond, D., Weld, D. S., Etzioni, O., and Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset.