Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/12836
DOI URL: http://dx.doi.org/10.21474/IJAR01/12836

ISSN NO. 2320-5407

RESEARCH ARTICLE

A REVIEW ON QUESTION AND ANSWER SYSTEM FOR COVID-19 LITERATURE ON PRE-TRAINED MODELS

Bavishi Hilloni and Debalina Nandy
Atmiya University, Rajkot.

Manuscript Info

Manuscript History
Received: 10 March 2021
Final Accepted: 14 April 2021
Published: May 2021

Key words:-
BERT, CORD-19, COVID 19, NLP (Natural Language Modelling), Question Answering System, Specter

Abstract

The COVID-19 literature has accelerated at a rapid pace and the Artificial Intelligence community as well as researchers all over the globe has the responsibility to help the medical community. The CORD-19 dataset contains various articles about COVID-19, SARS-CoV-2, and related corona viruses. Due to massive size of literature and documents it is difficult to find relevant and accurate pieces of information. There are question answering system using pre-trained models and fine-tuning them using BERT Transformers. BERT is a language model that powerfully learns from tokens and sentence-level training. The variants of BERT like ALBERT, DistilBERT, RoBERTa, SciBERT alongwith BioSentVec can be effective in training the model as they help in improving accuracy and increase the training speed. This will also provide the information on using SPECTER-document level relatedness like CORD 19 embeddings for pre-training a Transformer language model. This article will help in building the question answering model to facilitate the research and save the lives of people in the fight against COVID 19.

Introduction:-

The COVID-19 pandemic has made all the researchers of medical community as well as the experts of artificial intelligence to work upon accelerating literature. There is a need of Question Answering Systems which is helpful to stay up-to-date with all the pieces of information regarding the corona virus. However, the CORD-19 dataset is unlabelled and so developing Question Answering Systems is a challenging task. Many teams had participated in COVID-19 Open Research Dataset Challenge where there were two rounds and tasks assigned in each for developing Question Answering Systems.

Apart from CORD-19 dataset, there are various other sources like LitCovid dataset, COVIDQA and COVIDGQA datasets which are useful in developing Question Answering System for COVID-19 related research work and improve its performance. Here, we have reviewed various Question Answering Systems presented the transformers: BERT based language models: ALBERT: distilBERT, RoBERTa, SciBERT, BioBERT. Moreover, the CORD-19 embeddings are challenging to train and so a new method SPECTER based on document-relatedness has been developed which can be applied for fine-tuning the model.

Corresponding Author:- Hilloni Bavishi
Address:- Atmiya University, Rajkot.

Datasets
In this section, the CORD-19 dataset has been described in detail and a brief idea about other COVID-19 related datasets has been given.

CORD-19
The COVID-19 Open Research Dataset (CORD-19) is source of around 60,000 papers which was released to facilitate the research among the computing and medical community. The dataset has full text articles as well as preprints from various sources like PubMed, the World Health Organization's COVID-19 database and preprint servers of bioRxiv, medRxiv and arXiv. Each paper has associated with it the bibliographic metadata, like authors publication venue and when integrated in COVID-19 dataset, it has a unique cord_uid.

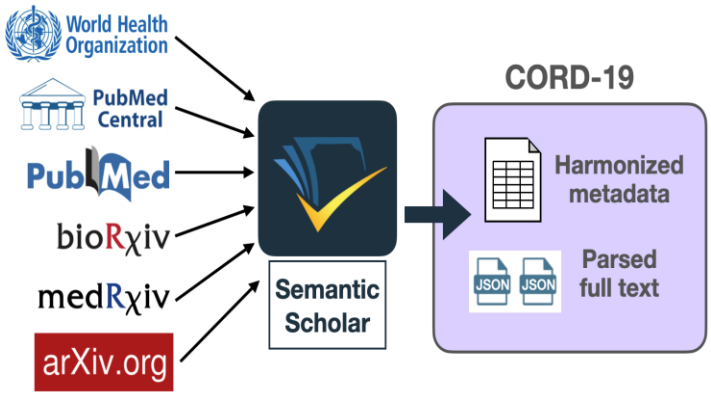

Fig.1:- CORD 19 dataset collected from different sources from Semantic Scholar[11].

The papers are in the form of PDF or PMC in the dataset. We have a full text JSON file which gives the complete details of the paper. The resource is been updated regularly with new articles. The Question Answering System was developed for the dataset based on the Tasks mentioned in Kaggle CORD-19 Challenge.

LitCovid
Unlike CORD-19 dataset which covers papers of COVID-19 along with SARS, MERS and related corona viruses LitCovid dataset has articles related to 2019 Novel Corona Virus. The articles are around 127044 (as of 8 May 2021) and every day the dataset is being updated. The LitCovid dataset is first among many others which provide COVID-19 related literature. It is the source to track the scientific information in PubMed and keeps all relevant articles which can further be used with advanced algorithms to derive improved results.

CORD-19 Embeddings
The CORD-19 embeddings are document-level embeddings unlike word or sentence embeddings. For this a method SPECTER is introduced which includes public endpoint of creating paper embeddings from papers' titles and abstracts. Such method of working with document-relatedness embeddings will be useful in pre-training a Transformer language model. The SPECTER mainly works with citation graph and can be applied for fine-tuning without task-specific pre-training. Here, the Natural Language Processing uses the Transformers language model to work with classification, prediction and recommendation.

BERT based Language Models for CORD-19
The BERT uses bidirectional transformer used for pre-training a machine learning model. The transformer has two machines: Encoder and Decoder. While BERT is Bidirectional Encoder Representation from Transformers, it works in both left-to-right and right-to-left direction on each layer for language modeling. Fine-tuning a specific task using BERT can be done for various purposes including question answering. Here, we include the BERT based language modeling done on COVID-19 related literature.

SciBERT
A pre-trained language model which helps in performing scientific tasks and based on BERT model is SciBERT. It addresses the large-scale labeled scientific data and deals with improving performance of unsupervised corpus of scientific publications.

The SPECTER method also involves the initialization of weights based on SciBERT pre-trained model. The training and testing data involves subset of Semantic Scholar Corpus. Here, the emphasize is laid on citation-based pre-training.

BioBERT
BioBERT is a pre-trained BERT-based language model which is trained on general as well as biomedical domain corpora. It shows language representation on large-scale biomedical corpora that is used for transfer learning. Again it allows fine-tuning and document classification. The BioBERT6 is applied on multi-labeled document classification models on LitCovid dataset and it surpasses the accuracy, F1 score compared to others. We can say that BioBERT performs better than original BERT model. In the document classification, BioBERT pre-trained on PubMed articles performs best.

ALBERT
The various issues like limitations in memory, requirement for more training time and degradation of performance while pre-training a language model can be addressed using A Lite BERT (ALBERT).

In developing the Question Answering Systems for COVID-19 related literature, ALBERT was used to find answers for questions mentioned in the Tasks of CORD-19. For this the Stanford Question Answering Dataset (SQuAD) was used to pre-train our Question Answering Systems. In[4], a light BERT model has been used gor developing question answering on CORD 19 tasks. As ALBERT has surpassed all other BERT models it is reasonably good to apply it at the time of urgency. It is successful in applying as it finds interesting relationships so is manageable.

DistilBERT
The Submissions to the COVID-19 Challenge involved use of a variation of BERT model DistilBERT which is cheaper, lighter and smaller.. This model proposes language representation and mainly used for extractive question answering tasks. DistilBERT has the advantages of reduced size compared to BERT, along with retaining its language understanding capabilities. So, DistilBERT is modeled from Transformers that are pre-trained on SQuAD 1.1.DistilBERT is a light Transformer model trained by distilling BERT.

For solving CORD 19 Tasks, there are submissions here DistilBERT is used. The HuggingFace DistilBertForQuestionAnswering and DistilBERTTokenizer on PyTorch is used[5].

RoBERTa
Language modeling with different pre-training has variations in performance. Some are better than the BERT model itself. This type of challenging situation is addressed by yet another model RoBERTa which is robustly optimized BERT pre-training approach and has found its application in drug discovery from CORD-19. This model solves the issues relating to training data size, hyperparameters and design choices.

As in[6], RoBERTa transformer-based model has shown its application in treating the challenge of discovering drug based on COVID-19 literature. Here, the transformer discovery method has proved better results compared to word2vec method for COVID-19 drugs, their combinations and side effects with respect to on-going clinical trials.

BioSentVec
BioSentVec are the biomedical embeddings done with sent2vec. It uses about 30 million documents from clinical notes from MIMIC-III Clinical Database. The aim for developing BioSentVec was to generate pre-trained encoders mainly for biomedical texts and thus it has helped a lot in carrying out research on COVID-19 literature. The texts are tokenized using NLTK and this pre-trained model is used to develop Question-Answer model for COVID-19 literature.

In this type of mode[2], the answers are retrieved using embeddings. The solution is compared with question string and sentence corpora, and finally using KNN algorithm, the best match answer sentence are retrieved.

Conclusion:-
We presented the various pre-trained models used to develop Question Answering Systems to compete in the Kaggle CORD-19 Challenge. We gave brief description of BERT-based model SciBERT, BioBERT, DistilBERT, ALBERT and BioSentVec on the dataset CORD-19 or LitCovid.

We consider that due to use of such powerful pretty transformers, the Question Answering Systems developed have been efficient and address the issues and needs of the hour. The Question Answering System developed by the Natural Language Processing experts would be of great help to medical research community.

References:-

1. Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, Daniel S. Weld SPECTER: Document-level Representation Learning using Citation-informed Transformers; arXiv:2004.07180v4 [cs.CL]
2. Blog on BioSentVec: https://rameshdatascientist.blogspot.com/2020/04/covid19-question-answering-model-based.html
3. CORD 19 Challenge: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge: Kaggle
4. [COVID-19] ALBERT-SQuAD for Q&A on CORD-19: https://www.kaggle.com/joljol/covid-19-albert-squad-for-q-a-on-cord-19
5. DistilBert for CORD19 QA: https://www.kaggle.com/madhuhegde/distilbert-for-cord19-qa/notebook:
6. Leo K. Tam San, Xiaosong Wang, Daguang Xu: Transformer Query-Target Knowledge Discovery (TEND): Drug Discovery from CORD-19; arXiv:2012.04682v2 [cs.CL]
7. LitCovid Dataset: https://www.ncbi.nlm.nih.gov/research/coronavirus/
8. Qingyu Chen, Alexis Allot, Zhiyong Lu: LitCovid: an open database of COVID-19 literature: Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D1534–D1540, 09 November 2020
9. Qingyu Chen, Yifan Peng, Zhiyong Lu BioSentVec: creating sentence embeddings for biomedical texts: arXiv:1810.09302v6 [cs.CL]
10. Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter; arXiv:1910.01108v4 [cs.CL]
11. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide Burdick D, Eide D, Funk K, Katsis Y, Kinney R, Liu Z, Merrill W, Mooney P, Murdick D, Rishi D, Sheehan J, Shen Z, Stilson B, Wade A, Wang K, Wang NX, Wilhelm C, Xie B, Raymond DM, Weld DS, sEtzioni O, Kohlmeier S (2020): CORD-19: The Covid-19 Open Research Dataset; arXiv:2004.10706v4 [cs.DL]
12. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov: RoBERTa: A Robustly Optimized BERT Pretraining Approach; arXiv:1907.11692v1 [cs.CL]
13. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations; arXiv:1909.11942v6 [cs.CL].