CovidBERT-Biomedical Relation Extraction for Covid-19

Shashank Hebbar, Dr. Ying Xie
Kennesaw State University
3391 Town Point Dr. NW
Kennesaw, GA 30144

Abstract

Given the ongoing pandemic of Covid-19 which has had a devastating impact on society and the economy, and the explosive growth of biomedical literature, there has been a growing need to find suitable medical treatments and therapeutics in a short period of time. Developing new treatments and therapeutics can be expensive and a time consuming process. It can be practical to re-purpose existing approved drugs and put them in clinical trial. Hence we propose CovidBERT, a biomedical relationship extraction model based on BERT that extracts new relationships between various biomedical entities, namely gene-disease and chemical-disease relationships. We use the transformer architecture to train on Covid-19 related literature and fine-tune it using standard annotated datasets to show improvement in performance from baseline models. This research uses the transformer BERT model as its foundation and extracts relations from newly published biomedical papers.

Introduction

Over 10 million articles are available on PubMed, with at least a million articles published every year. As a result, it poses a tremendous challenge for researchers Davis et al. to keep upto date with the latest knowledge. This problem is severely acute during a Covid-19 pandemic or any other public health emergency. The goal of biomedical relation extraction is to gain information between different entities, such as disease-gene association, protein-protein interaction and chemical-disease interaction. This helps in re-purposing existing drugs and develop therapeutics to combat symptoms related to Covid-19. The first step in finding relations is biomedical name entity recognition, currently there are several state of the art named entity annotation tools that recognize biomedical entities on corpora with high accuracy. The second step is to identify if there exists a relationship between entity pairs at a sentence level or document level. We focus on two specific popular relationship types namely chemical-disease and gene-disease associations. chemical-disease is a binary classification problem which identifies whether the entity pair has a semantic relationship or not, whereas gene-disease is a multiclass classification problem which detects the semantic relationship between the entity pair and classifies them into a specific type.

Literature Review

Statistical methods have been traditionally used for natural language processing but since it suffers from the curse of dimensionality in learning the joint functions of language models. Hence representation learning has gained popularity, which involves representing words or phrases in a low dimensional space. Word embeddings operate on the principle that words with similar meaning tend to occur together. It captures the semantic meaning of the neighbours of a word. One of the main benefits of word embedding is that they capture the similarity between words.

Bengio et al. proposed a neural model that learnt the joint distributions of words. He contended by combining word representations using sequence probability, sentence representations can be learned which can be used to detect semantically similar sentences. One of the most popular type of word distributions was proposed by Mikolov et al. known as the CBOW (Continuous bag of words model) and the skip-gram model. The CBOW method models the conditional probability of predicting a target word given its surrounding context of a specific window size, whereas the skip-gram does the opposite. It models the conditional probability of predicting the context with a specified window given the target. The target word embedding is determined by the accuracy of the prediction. As the size of the target word embedding increases, the accuracy increase up to a certain convergence point. Some of the limitations of this method include its inability to represent phrase as embeddings. Combining embeddings of words do not necessarily represent the embedding of the phrase.

BERT Devlin et al. is a contextualized word representation model that outperforms the global word representation models that is based on a masked language model and pre-trained using bidirectional transformers Vaswani et al.. It achieves state of the art results on many NLP tasks, it uses masked language modeling and next sentence prediction as pre-training auxiliary objective functions.

Proposed Model

To further improve the classification power of BERT for covid, we further pre-train the BioBERT Lee et al. model.

BioBERT improves over the base BERT Devlin et al. model by pre-training BERT on a wide variety of biomedical corpus. This included PubMed abstracts and PMC full-text articles. We take the pre-trained Biobert model and train it on additional Covid-19 related corpus mentioned below.

- The CORD-19 open research dataset from the Allen Institute of AI. The dataset contains all COVID-19 and coronavirus-related research (e.g. SARS, MERS, etc.) from PubMed's PMC open access corpus using the query (COVID-19 and coronavirus research).Wang et al.
- iSearch COVID-19 Portfolio, Comprehensive, expert-curated portfolio of COVID-19 publications and preprints that includes peer-reviewed articles from PubMed and preprints from medRxiv, bioRxiv, ChemRxiv, and arXiv.
- LitCovid, it is a curated literature hub for tracking up-to-date scientific information about the Coronavirus Disease 2019 (COVID-19). Chen, Allot, and Lu.
- PubChem is part of the National Center for Biotechnology Information. It contains papers on small molecule compounds, bioactivity data, biological targets, bioassays, chemical substances, patents, and pathways based on coronavirus.

The version of BioBert used for training was BioBERT-Base v1.0 (+ PubMed 200K + PMC 270K) which is based on BERT-base-Cased, therefore it has the same vocabulary. The resulting pre-trained model is named as Covid-BERT. It has the same architecture style as the BERT base model. An encoder with 12 transformer blocks , 12 self-attention heads and the hidden size of 768. It takes an input sequence of no longer than 512 tokens and outputs the representation of a sequence. Each training sequence in BERT has two types of special tokens known as [CLS] which indicates the start of a token and contains the special classification embedding and the [SEP] token is used for separating segments within a training instance. In text classification, BERT takes the final hidden state [CLS] of the first token as a representation of the entire sequence. A dense layer with a softmax classifier is added on top of Covid-BERT to predict the probability of label k.

$$p(k|h) = softmax(Wh) \tag{1}$$

where W is the task specific parameter matrix. During fine tuning, only the parameters of the dense layer are jointly trained by maximizing the log probability of predicting the correct label.

Parameters	Values
Training steps	500000
Warmup steps	50000
Max sequence length	128
Max predictions per seq	20
Masked lm prob	0.15
Batch size	32
Learning rate	$2e^{-}5$

Table 1: Parameters used to train Covid-BERT

The Covid-BERT pretraining hyper-parameters are given in Table 1. and the corresponding losses given in Table 2. The pre-training was done using the T-4 GPU for 40 hours. After the pretraining process, named entity recognition is applied on the experimental datasets using the state of the art hunflair tagger Weber et al.. After the entity tagging process, the input text is fed into Covid-BERT to generate contextualized word embeddings. The base BERT architecture encoder block has 12 layers. Each layer captures different features of the input text. To improve performance during fine-tuning We experimented accuracy rate by extracting embeddings with different layers.

Parameters	Values
Loss	0.868987
Masked lm accuracy	0.7952974
Masked lm loss	0.8541765
Next sentence accuracy	0.99625
Next sentence loss	0.015645374

Table 2: Training loss for Covid-BERT

Experiments

Experiments were run for two different relationship types by fine-tuning on standard annotated datasets. During fine-tuning, the pretrained layers were frozen and after fine-tuning on those standard datasets, the trained model was applied on unlabeled corpus to extract new relationships. Name entity recognition on the unlabeled corpus was done using the hunflair tagger Weber et al.. The model was evaluated using precison, recall and F-1 score.The baseline models for comparison were BioBERT Lee et al.and Kernel-SVMAlam et al. by selecting the best kernel function.

Chemical-Disease Model

Relations between chemicals and diseases (Chemical-Disease Relations or CDRs) play critical roles in drug discovery, biocuration, drug safety, etc. Although some well-known manual curation efforts like the Comparative Toxicogenomics Database (CTD) project Davis et al. have already curated thousands of documents for CDRs, the manual curation from literature into structured knowledge databases is time-consuming and insufficient to keep up to date. Hence, we use the BioCreative V (BC5) annotated dataset Li et al. to fine-tune the model. Unlike traditional biomedical relation extraction datasets where the annotations are at sentence level for example in disease-gene association, here the annotations are at document level where relationships could be described across multiple sentences.

To indicate if a sentence expresses a relationship, the chemical/disease entity mentions in the sentence are compared with the annotation provided by the experts. If a sentence contains the entity pair that matches with the given annotation, the sentence is given a label of 1 indicating a relationship , similarly a label of 0 is given if it doesn't contain any matches. After processing all the 500 document instances, 1424 positive relations were generated and 2039 negative relations were generated.

Hyperparameters	Values
batch size	128
learning rate	0.001
validation split	0.3
Number of layers	7

Table 3: Hyperparamters for CDR Dense Layers

The dataset was split into training and test set using a 70/30 split. The deep learning model was implemented using Keras using the Adam Optimizer. The hyperparameters of the dense layer were fine tuned and the best values are shown in Table 3. To further investigate the model, word embeddings were extracted from different layers of the transformer model as shown in Table 4.

Layer	F-1
Layer 8	0.86
Layer 9	0.88
Layer 10	0.89
Layer 11	0.89
Layer 12	0.91

Table 4: Fine-tuning using different layers

As seen from Table 4, embeddings from the last layer gave the best performance. In table 5, the performance of Covid-BERT was compared with Bio-BERT keeping the hyperparamters constant and using the last layer for extracting the word embeddings. On average Covid-BERT performs better than Bio-BERT both in terms of recall and precision.

Model	Precision	Recall	F-1
Bio-BERT	0.87	0.88	0.874
Covid-BERT	**0.91**	**0.91**	**0.91**
K-SVM	0.83	0.81	0.82

Table 5: Performance Comparison

New Relations: To extract new relationships from unlabeled corpus, we applied the model on newly published abstracts in CORD-19 Wang et al. from 12/01/2020 to 12/31/2020. Before applying the trained model, new text was passed through the hunflair tagger to do name-entity recognition. Below are some examples of newly extracted relations with the smallest frequency or relations that are rare, along with the sample text.
Example 1:This article presents a case of calciphylaxis induced by warfarin in a COVID-19 patient. **Disease**: calciphylaxis **Chemical**: warfarin

Gene-Disease Classification

Extracting Gene Disease relationships is crucial for various biomedical applications such as drug re-purposing. It helps understanding disease etiology in order to prevent manifestation and further spread of the disease. The dataset used to train gene disease relationships is provided by DisGeNet database. The DisGeNET database integrates information of human gene-disease associations (GDAs) and variant-disease associations (VDAs) from various repositories including Mendelian, complex and environmental diseases. The integration is performed by means of gene and disease vocabulary mapping and by using the DisGeNET association type ontology.Piñero et al..

The dataset contains the variable ”score” which ranges from 0 to 1, and takes into account the number of sources, and the number of publications supporting the association. 1 indicating strong confidence and 0 indicating weak confidence. The distribution of score by quantiles is shown in Table 6. The problem is converted to a multilabel classification problem, by assigning labels according to the distribution of score as shown in Table 7. Hence, a total of 3 labels are created based on the continuous variable ’score’. Since this is a multilabel classification problem, we take the weighed average of precision , recall and F-1 score of the individual classes. Its weighted by the number of samples.

Quantile	Score
0.15	0.01
0.25	0.02
0.50	0.10
0.75	0.20
0.85	0.40
0.90	0.60
0.95	0.75
0.98	1.00

Table 6: Quantiles of the variable Score

Score	Label
0 - 0.10	1
0.10 - 0.40	2
0.40 - 1.00	3

Table 7: Multiclass Labels

The dataset with 2580 instances was split into training and test set using a 70/30 split. Keras with an Adam optimizer was used for implementation. The hyperparameters of the dense layer were fine tuned and the best values are shown in Table 8. To further investigate the model, word embeddings were extracted from different layers of the transformer model as shown in Table 9.

As seen from Table 9, embeddings from the last layer gave the best performance. In table 10, the performance of Covid-BERT was compared with Bio-BERT keeping the hyperparamters constant and using the last layer for extracting the word embeddings. On average Covid-BERT performs better than Bio-BERT both in terms of recall and precision.

Hyperparameters	Values
batch size	64
learning rate	0.0001
validation split	0.2
Number of layers	5

Table 8: Hyperparamters for GDA Dense Layers

Layer	weighted F-1
Layer 8	0.57
Layer 9	0.59
Layer 10	0.59
Layer 11	0.60
Layer 12	0.61

Table 9: Fine-tuning using different layers

Model	wt.Precision	wt.Recall	wt.F-1
Bio-BERT	0.59	0.60	0.59
Covid-BERT	0.59	**0.63**	**0.61**
K-SVM	0.52	0.55	0.53

Table 10: Performance Comparison

New Relations: Similar to disease-chemical example, we extract relationships from newly published papers from 12/01/2020 to 12/31/2020 available in in CORD-19 Wang et al.. Since this is a multi-class problem, we classify relations into three categories namely strong,mild and weak relations

Strong Relation: As more data accumulate about the immune responses and the kinetics of neutralizing antibody (nAb) production in SARS- CoV-2 infected individuals , new applications are forecasted for serological assays such as nAb activity prediction in convalescent plasma from recovered patients. **Disease**: SARS **Gene**: nAb

Mild Relation: The expression of mce operons depends on many factors , such as the growth phase , the culture medium , and the localization of M. tuberculosis tuberculosis infection. **Disease**: tuberculosis infection **Gene**: mce operons

Weak Relation: Our findings from juxtaposing IgG and PCR tests thus reveal that some SARS - CoV-2-positive patients are non - hospitalized and seropositive , yet actively shed viral RNA (14 of 90 patients). **Disease**: SARS **Gene**: IgG

Conclusion

In our paper, we propose an improved way of detecting biomedical relationships in the context of Covid-19 pandemic. We develop CovidBERT which is a transformer based model trained on large amounts of covid-19 related text and corpus. After fine-tuning we demonstrate the improvement in performance over baseline BioBERT, as well as newly extracted gene-disease and chemical-disease relationships from newly published papers. All these experiments are performed on manually curated datasets annotated by experts. After fine-tuning the model, we illustrate examples of relationships extracted from newly published biomedical papers.

References

Alam, S.; Kang, M.; Pyun, J.-Y.; and Kwon, G.-R. 2016. Performance of classification based on pca, linear svm, and multi-kernel svm. In *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*, 987–989. IEEE.

Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, C. 2003. A neural probabilistic language model. *Journal of machine learning research* 3(Feb):1137–1155.

Chen, Q.; Allot, A.; and Lu, Z. 2020. Litcovid: an open database of covid-19 literature. *Nucleic Acids Research*.

Davis, A. P.; Murphy, C. G.; Saraceni-Richards, C. A.; Rosenstein, M. C.; Wiegers, T. C.; and Mattingly, C. J. 2009. Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic acids research* 37(suppl_1):D786–D792.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4):1234–1240.

Li, J.; Sun, Y.; Johnson, R. J.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A. P.; Mattingly, C.; Wiegers, T. C.; and Lu, Z. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation* 2016.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26:3111–3119.

Piñero, J.; Ramírez-Anguita, J. M.; Saüch-Pitarch, J.; Ronzano, F.; Centeno, E.; Sanz, F.; and Furlong, L. I. 2019. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* 48(D1):D845–D855.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, L. L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Eide, D.; Funk, K.; Kinney, R.; Liu, Z.; Merrill, W.; et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Weber, L.; Sänger, M.; Münchmeyer, J.; Habibi, M.; Leser, U.; and Akbik, A. 2020. Hunflair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. *arXiv preprint arXiv:2008.07347*.