Identify Risk Factors of COVID-19 from Medical Literature

Anonymous EMNLP submission

Abstract

Identifying risk factors of COVID-19 is crucial for protecting vulnerable people, determining the proper way of managing infected cases, making public policies for slowing down the spreading of SARS-CoV-2, to name a few. There is a vast amount of literature that studies the risk factors of COVID-19. However, the information of risk factors is embedded in unstructured literature texts, which is difficult to extract. In this paper, we develop information extraction approaches to extract risk factors of COVID-19 from medical preprints. Experiments on the CORD dataset demonstrate the effectiveness of our methods.

1 Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease that has infected 6,226,409 individuals all over the world and caused 373,883 deaths, as of June 1 in 2020. COVID-19 can cause symptoms such as fever, cough, short of breath, etc., ranging from mild to very severe. The level of severity varies in different infected individuals, depending on characteristics or medical conditions that increase their risk, such as being 65 years of age or older or having serious underlying medical conditions. These are commonly called risk factors. Identifying risk factors of COVID-19 is crucial for protecting vulnerable people, determining the proper way of managing infected cases (e.g., hospitalization, intensive care, self quarantine at home), making public policies for slowing down the spreading of SARS-CoV-2, to name a few.

There have been a number of studies (Rahman and Sathi, 2020; Zhou et al., 2020) recently on the risk factors of COVID-19. But such information is contained in long and unstructured articles, which are difficult to digest. Manually reading these articles to identify such information is very time-consuming. Besides, the number of articles studying risk factors of COVID-19 is rapidly growing. It is difficult for human to track such information continuously. This motivates us to build automated natural language processing (NLP) pipelines to extract the risk factors of COVID-19 and represent the extracted information in structured form that is easy to search and analyze.

The major contributions of this paper are:

- We develop an NLP pipeline that is able to automatically extract the risk factors of COVID-19 from large-scale unstructured literature.
- Experiments demonstrate that our methods can accurately extract the risk factors of COVID-19.

The rest of the paper is organized as follows. In Section 2, we introduce the dataset. Section 3 and 4 present the methods and experimental results. Section 5 concludes the paper.

2 Dataset

We used the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020) for our study. In response to the COVID-19 pandemic, the White House and a coalition of research groups prepared the CORD-19 dataset. It contains over 45,000 scholarly articles, including over 33,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. These articles are contributed by hospitals and medical institutes all over the world. We use a total of 13202 papers in the dataset: 803 papers from biorixiv_medrxiv, 9000 papers from comm_use_subset, 1973 papers from nonComm use_subset, and 1426 papers from pmc_custom_license. We download the dataset on March 13th, 2020 and the dataset has been updating.

3 Method

Our goal is to develop natural language processing (NLP) methods to analyze a large collection of COVID-19 literature and discover the risk factors of COVID-19. To achieve this goal, we need to address two technical challenges. First, in the large collection of COVID-19 literature, only a small part of sentences are about COVID-19's risk factors. It is time-consuming to manually identify these sentences. Simple methods such as keyword-based retrieval will falsely retrieve sentences that are not about COVID-19's risk factors and miss sentences that are about COVID-19's risk factors. How can we develop NLP methods to precisely and comprehensively extract sentences containing COVID-19's risk factors? Second, given the extracted sentences, they are still highly unstructured, which are difficult for decision-makers to digest and index. How can we extract phrases containing COVID-19's risk factors from the unstructured sentences?

3.1 Extracting Sentences Containing Risk Factors of COVID-19

To address the first challenge, we develop a sentence classifier to judge whether a sentence contains COVID-19's risk factors.

Data Labeling We first label sentences as either containing the risk factors of COVID-19 or not. To facilitate such labeling, we collect 41 risk factors of COVID-19 (shown in Table 1) by searching the web. They can be divided into four categories: (1)sub-population, (2)diseases and symptoms, (3)biomarkers, (4)society and economics. Given these 41 risk factors, we retrieve sentences containing these risk factors using keyword matching. In the end, we collected 101,468 such sentences from 13,202 papers in the CORD-19 dataset. Among them, 2170 sentences contain "covid", "cov" or "coronavirus". These 2170 sentences can be categorized into three groups:

- Sentences about the risk factors of COVID-19 (P)
- Sentences about the risk factors of other diseases, such as SARS and MERS ($N1$)
- Sentences about the risk factors of coronavirus diseases, but not sure about which disease ($N2$)

We also select sentences that contain "COVID-19" or "2019-nCoV", but are not about risk factors ($N3$). The rest of sentences are neither about COVID-19 or about risk factors ($N4$). In addition, we manually read the sentences and select 200 that are about risk factors of COVID-19 but not contained the seed risk factors in the table 1. In the end, we obtain a labeled dataset containing 400 positive sentences from P, 500 negative sentences from $N1 \bigcup N2 \bigcup N3$, and 500 negative sentences from $N4$. We use this dataset (denoted by train&val) for model training and validation. In addition, we label a test set containing 300 sentences: 100 positive ones from P, 100 negatives from $N1 \bigcup N2 \bigcup N3$, and another 100 negatives from $N4$.

Sentence classification Given these positive and negative training sentences, we use them to train a sentence classifier which predicts whether a sentence is about the risk factors of COVID-19. We use the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) model for sentence classification. BERT is a neural language model that learns contextual representations of words and sentences. Given a sequence of tokens, BERT uses a Transformer encoder to encode each token. The encoder is a stack of building blocks, each consisting of a self-attention layer and a position-wise feedforward layer. Residual connection and layer normalization are applied around each layer. BERT pretrains the weights of the Transformer encoder by randomly masking some tokens in the input sequence and forcing the encoder to correctly reconstruct the masked tokens. To apply the pretrained BERT to a downstream task such as sentence classification, one can add an additional layer on top of the BERT architecture and train this newly-added layer using the labeled data in the target task.

The original BERT is pretrained on general-domain corpus. Its size is very large, enabling BERT to effectively learn latent semantics of general-domain words. However, the general-domain corpus has a large domain shift with medical texts. The original BERT may be biased to general-domain texts and be less effective in representing medical texts. To solve this problem, BioBERT (Lee et al., 2020) finetunes the original pretrained BERT on biomedical datasets (including PubMed abstracts and PubMed Central full-text articles) for adapting the representations to the biomedical domain. In our work, we take a

Sub-population	Older adults, pregnant women
Diseases and symptoms	lung disease, heart disease, cardiovascular disease, chronic kidney disease, coronary heart disease, liver disease, nervous system disease, chronic respiratory disease, coagulation dysfunction, diabetes, HIV, cancer, asthma, sepsis, hypertension, arrhythmia, myocardial infarction, pneumonia, chronic renal failure, dyspnea, high fever
Biomarkers	neutrophilia, lymphocytopenia, methylprednisolone, leukocytosis, higher lactate dehydrogenase, plasma urea, serum creatinine, IL-6, low CD4 cell count, CD3 T-cell, CD8 T-cell, increased high-sensitivity cardiac troponin, increased high-sensitivity C-reactive protein, D-dimer, aspartate aminotransferase, alanine aminotransferase
Society and economics	medical resources, socioeconomics

Table 1: 41 seed risk factors in four categories

step further and finetune pretrained BioBERT on the CORD dataset (we call it COVID-BERT) to adapt the representations to COVID19-related literature. RoBERTa (Liu et al., 2019) has shown that removing the Next Sentence Prediction (NSP) task can slightly improve model's performance in downstream tasks, therefore we pretrain COVID-BERT only using the Masked Language Model (MLM) task.

3.2 Risk factor extraction

Given the sentences containing risk factors, the next step is to extract the risk factor phrases from the sentences. A standard way is to formulate this problem as a named entity recognition (NER) problem: annotate risk factor entities in sentences and train an NER model. To achieve this goal, we need to label a lot of such entities, which is time-consuming. To reduce human-annotation effort, we take a different approach: formulate the problem as a reading comprehension (RC) problem and leverage external labeled RC datasets to train an RC model without additional annotation efforts.

We use the Stanford Question Answering Dataset (SQuAD v1.1) dataset, which contains 100K question/answer pairs (Rajpurkar et al., 2016). The task is to find the answer of the question in the given text. We fine-tune our COVID-BERT model on SQuAD v1.1 dataset and use *"What is the risk factor?"* to query all the positive sentences to get the risk factors in them. Given the extracted answers, we filter out those meeting the following rules.

- Length$\leq$3
- contain "[CLS]" or "[SEP]" or "risk factor'"
- Contain "covid" or "2019 - ncov" or "sars"
- Not contain [a-z]

4 Experiments

4.1 Experiment settings

For sentence classification, the train&val set was randomly split into a training set and a test set with a radio of 3:1. The architecture of COVID-BERT model is the same as that of BioBERT-base, where the number of hidden layers is 12, the hidden size is 768, the number of self-attention heads is 12, and the total number of weight parameters is 110M. WordPiece tokenization was used to avoid the out-of-vocabulary issue. The vocabulary ofCOVID-BERT is the same as BioBERT-base. The model is trained for 10K optimization steps with a minibatch size of 8 on 4 GPUs. Training is stopped when the validation loss starts to increase. We finetune COVID-BERT on SQuAD with a learning rate of 3e-5 for 2 epochs.

4.2 Results on sentence classification

Method	Validation	Test
TextCNN	0.87	0.76
BERT	0.93	0.90
BioBERT	0.93	0.93
COVID-BERT	0.93	0.95

Table 2: Sentence classification performance of different methods on the validation and test set.

Table 2 shows the performance of different methods on the test set and validation set. From this

table, we make the following observations. First, BERT based models outperform TextCNN. Second, BioBERT performs better than BERT. This is because BioBERT is pretrained on biomedical corpora while BERT is pretrained on general-domain corpora. The learned representations in BioBERT is more more suitable for biomedical NLP tasks, such as our COVID-19 sentence classification task. Hence, BioBERT achieves better performance. Third, COVID-BERT outperforms BioBERT. COVID-BERT is pretrained on COVID-19-related texts which include COVID-19-specific terminologies that may not contained in general biomedical corpus. Hence COVID-BERT achieves better performance. We apply COVID-BERT to 1,048,575 sentences in CORD-19 dataset. 39,150 of them are predicted as being positive.

The importance of selecting negative examples In the experiments, we found that which negative examples to use for model training has a large impact on the testing performance. We compare three settings for selecting negative examples in the train&val dataset. The positive examples in the train&val dataset remain the same, which are 400 sentences from P. The test set remains the same. Table 3 shows the results.

Setting	Negative examples	Val Acc	Test Acc
Setting-1	500 from $N1 \bigcup N2 \bigcup N3$, 500 from $N4$	0.94	0.96
Setting-2	1000 from $N1 \bigcup N2 \bigcup N3$	0.93	0.76
Setting-3	1000 from $N4$	0.97	0.79

Table 3: Performance under three settings of negative sentence selection.

From this table, we can see that Setting-1 works much better than Setting-2 and Setting-3. This indicates that it is important to use sentences from both $N1 \bigcup N2 \bigcup N3$ and $N4$.

4.3 Results on risk factor extraction

To evaluate the performance of the extraction model, we label a test set with 200 sentences. Considering different options to choose the risk factors from sentences, we give 3 answers of risk factors for each sentence and use the highest score among 3 while evaluating. Exact match (EM) and F1 are used as evaluation metrics. We compare with a named entity recognition (NER) baseline. We used a BiLSTM-CRF model to perform NER. It uses a bidirectional LSTM (Bi-LSTM) model to learn the contextual representations of words and uses a conditional random field (CRF) to capture the sequential correlation of labels. The word embedding size was 50. The hidden state size of LSTM is 32. The model was trained on 3501 samples containing the seed factors we choose for 5 epochs. Table 5 shows the results.

Table 4 shows some examples of extracted risk factors from positive sentences.

Positive Sentence	Risk Factor
Detection of specific antibodies in patients with fever can be a good distinction between COVID-19 and other diseases, so as to be a complement to nucleic acid diagnosis to early diagnosis of suspected cases.	Fever
Nevertheless, the renal function of patients with COVID-19 needs to be monitored regularly, especially in patients with elevated plasma creatinine.	Elevated plasma creatinine

Table 4: Examples of extracted risk factors from positive sentences

	F1	Exact match
Ours	0.79	0.29
NER	0.27	0.02

Table 5: Performance of risk factor extraction.

5 Conclusions

In this paper, we study how to extract risk factors of COVID-19 from medical literature. We develop a BERT-based model to select sentences containing risk factors of COVID-19 and an extraction model based on reading comprehension to extract risk factor phrases. Experiments on the CORD dataset demonstrate the effectiveness of our approaches.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Abdur Rahman and Nusrat Jahan Sathi. 2020. Risk factors of the severity of covid-19: a meta-analysis. medRxiv.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. arXiv preprint arXiv:2004.10706.

Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, et al. 2020. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. The lancet.