COVID-19 related Natural Language Processing using BioBERT

Antton Lamarca Arrizabalaga

18th August 2020

Contents

Abstract

The COVID-19 pandemic has resulted in an unprecedented amount of scientific publications related to its biological mechanisms, spread of the virus and potential treatment of the disease being published. Almost every field of science has produced new research in an attempt to minimize the impact of the pandemic. The amount of publications itself is also a problem, however, as it makes identifying relevant findings and promising research lines more difficult. To avoid missing potentially interesting publications, a specialised text-mining tool could "process" the articles that are being continuously released to extract the most relevant information. This way, it would effectively remove a bottleneck in the research process. While text-mining approaches have been proposed before to deal with the high output of new research, tools specialised in finding information specifically related to COVID-19 have become a necessity with the coming of the pandemic. In this project, the BioBERT language-representation model was applied to perform Name Entity Recognition on a corpus of COVID-19-related scientific publications. Four new models were created by fine-tuning the base BioBERT model with a dataset built from that corpus. The performance of the models when detecting mentions of diseases and proteins related to the SARS-CoV-2 virus was then compared to that of the default BioBERT models.

Introduction

The global pandemic of COVID-19 has resulted in an unprecedented volume of scientific publications related to the disease and the SARS-CoV-2 virus that causes it. According to Nature Biotechnology, 7,136 COVID-19-related papers had already been uploaded to PubMed in 2020 by the 5th of May [1]. The COVID-19 literature database managed by the World Health Organization currently contains almost 20.000 scientific articles on the topic, and the amount of papers being published on the topic will keep increasing.

Keeping up with the constant output of relevant papers has been an issue for researchers even before the pandemic caused an explosion on the amount of publications. One of the solutions proposed to solve this problem has been the development of effective text mining tools that will be able to extract information from an otherwise overwhelming amount of research articles [2]. Such a tool would be built following principles of Natural Language Processing (NLP).

NLP and its uses

Natural Language Processing is a field within Computer Science and Artificial Intelligence focused on the computer-based processing and analysis of human languages. NLP-based technology has been applied to develop things such as machine translations, automatic audio-to-text processing and the construction of knowledge-graphs. It can also be a valuable tool in biomedical research and healthcare [3], assisting in diagnostics or information extraction (extracting information from literature, medical or not).

Historically, NLP and text mining have been "rule-based". A number of handcrafted rules were used to create NLP systems such as chatbots. However, most modern approaches to NLP rely heavily on Neural Network-based methods [4].

One of such modern NLP systems is the language representation model BERT [5]. Developed by Google, BERT has become popular in the NLP community for allowing the creation of state-of-the-art models with minimal fine-tuning of a pretrained BERT model. On top of that, several projects have attempted to increase the effectiveness of BERT for specific fields by creating domain-specific variants of the original BERT such as SciBERT, ClinicalBERT or BioBERT, which are suitable for the biomedical domain [6, 7, 8].

Common NLP tasks

Natural Language Processing is a broad field that deals with several challenges. Within NLP, there are some common tasks that models attempt to carry out. A distinction can be made between sentence-level tasks, which focus on whole sentences, and token-level tasks. In NLP, a token is a string of characters between spaces or punctuation marks, often a word. Common token-level tasks include Relationship Extraction (RE), Question-answering (QA) and Named Entity Recognition (NER) [9].

In Named Entity Recognition (NER), an NLP model is used to recognize and classify "Named Entities", e.g. a location, person, or protein, into different categories. NER is a crucial sub-task for information extraction, as it allows for automatic identification of key terms in the text.

BioBERT

BioBERT is the medical/biological variant of BERT, first introduced by Lee *et al.* [8]. While BERT is a general-purpose tool trained on English Wikipedia and BookCorpus, BioBERT is designed specifically to deal with biomedical text. It builds on top of BERT and was trained using PubMed abstracts and full-text articles from PMC. This makes it more appropriate to handle words that are specific to the biomedical field. BioBERT is able to perform three of the aforementioned text mining tasks: Named Entity Recognition (NER), Relationship Extraction (RE) and Question Answering (QA).

In this project, BioBERT was used exclusively for NER. The goal of the project was to generate new BioBERT models that would be better than those presented in the original BioBERT paper at detecting mentions of COVID-19, the SARS-CoV-2 virus, and drugs that could potentially be used in the treatment of patients suffering from the disease.

Methods

The Colab environment

Google Colaboratory (https://colab.research.google.com/), also known as just Colab, is a cloud-based environment for Jupyter Notebook developed by Google. It allows its users to run python code through the browser, without the need of any setup process. Users can also get access to limited GPU resources for free, which makes it ideal for machine learning-related tasks. This project was carried out almost exclusively in Colab, with data being loaded from and stored in Google Drive.

Fine-tuned Models

The GitHub repository for the BioBERT project (https://github.com/dmis-lab/biobert) provides a number of different versions of the base BioBERT model [8]. In this project, I used the **BioBERT-Base v1.1 (+ PubMed 1M)** model. The repository also includes 8 datasets that were used by the developers of BioBERT to fine-tune the base BioBERT model for NER. Particularly relevant among these are: the NCBI-disease dataset [10], which includes annotated disease and symptom names; the JNLPBA dataset [11] and the BC2GM dataset [12], both of which contain annotated gene/protein names; and the two variants of the BC5CDR dataset [13], one focused on drugs and chemicals, and a second one focused on diseases. The base BioBERT model and the datasets were loaded into a notebook in Colab, where the datasets were used to perform fine-tuning of the base model.

Following the steps of Lee *et al.* [8], fine-tuning was performed for three of the datasets: NCBI-disease, JNLPBA and BC5CDR-Chem. This resulted in three separate models specialised in detecting mentions of diseases, proteins/genes and chemicals/drugs respectively. In each case, training was carried out for 10 epochs, using the default BioBERT parameters.

The CORD19 research challenge

In March of 2020, the Allen Institute For AI issued a challenge through the Kaggle platform that included several machine-learning and text-mining tasks aimed at advancing research related to COVID-19 (https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge). The challenge also provided access to the CORD-19 dataset [14], a dataset of more than 140,000 articles in JSON format

that are all related to the COVID-19 disease or coronaviruses. CORD-19 was specifically designed to be used by the global community in machine-learning projects. In this project, a 100-paper subset of the dataset (https://github.com/je5720kas/EDAN70/blob/master/) was isolated and extensively used in several tests, with the idea of making every step scalable for the full database.

Prediction

While BioBERT claims to have a separate "prediction mode" for performing NER with fine-tuned models, this functionality appears to not have been fully implemented yet. Instead, prediction can be done by repeating the steps of the fine-tuning process for a model that has already being fine-tuned. Introducing data into the BioBERT model for prediction, as well as extracting the results from the output file proved challenging, and extensive pre- and post-formatting of the data was required. The prediction itself was performed with a learning rate of 5e-05 and a maximum sentence length of 512 tokens.

Text Processing

In order to introduce data into BioBERT, both for fine-tuning and prediction, the data needs to be stored in a .tsv file structured in a variation of the CoNLL-2003 named entity data format [15]. While the exact data-formatting steps used by the BioBERT team in creating this file have not been made available by the developers citing issues related to copyright and the PubMed API, an approximation was implemented through a collection of custom-made python scripts. This pipeline was later used to generate the necessary .tsv file form the data provided in the CORD19 challenge. The final version of the pre-processing pipeline (https://github.com/AnttonLA/BINP37/tree/master/formatting) was able to extract information from either files of the CORD19 database or JSON files in the standard PubAnnotation format [16].

(http://www.pubannotation.org/docs/annotation-format/).

Performing Named Entity Recognition with BioBERT produces a single output file in a format similar to the CoNLL-2003 challenge format [15]. This file was reconverted to more readable JSON files in the standard PubAnnotation format by another custom script.

(https://github.com/AnttonLA/BINP37/tree/master/formatting/eval).

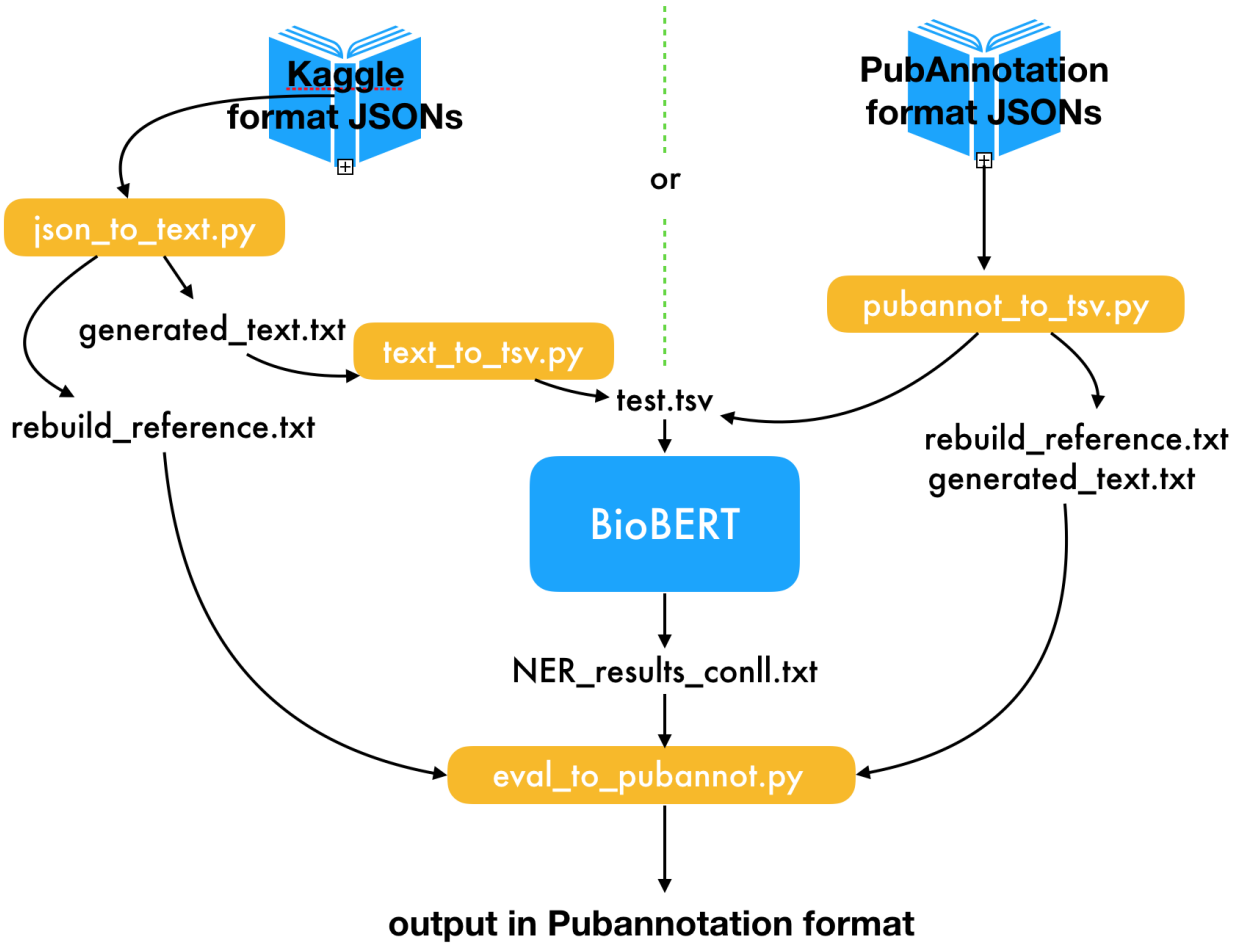

Figure 1: Diagram detailing the text processing pipeline.

Models built with expanded training data

In order to be able to fine-tune models with data specific to COVID-19-related articles, new data-sets were generated from the results of dictionary-based taggers applied to the CORD19 data. The tagged files were provided by other students in the Cell Death, Lysosomes and Artificial Intelligence group at Lund University, and were in PubAnnotation format. The files were converted into CoNLL-2003 format .tsv files. These data-sets contained each the same subset of 100 papers taken from the CORD19 database, but automatically annotated for the tags Disease and Protein/Gene. While far from perfect and prone to miss words not in the dictionary or make mistakes, these annotations were deemed good enough to train new BioBERT NER models with.

Apart from the three basic BioBERT models mentioned before, four more models were trained with this new data. Two of them using only the automatically annotated files, and another two using a combination of the databases provided used in the BioBERT paper and these new annotated files.

In total, 4 models were fine-tuned with data-sets containing CORD19 data: dictionary-only disease, dictionary-Disease and NCBI-Disease combined, dictionary-only protein and dictionary-Protein and JNLPBA combined.

Evaluation of results

Two different evaluation scripts were used when comparing the results of different models. When attempting to reproduce the steps of the BioBERT paper by Lee *et al.* [8], the models fine-tuned with the datasets of the original BioBERT publication were evaluated using the native BioBERT evaluator script in order to replicate the steps of the article as closely as possible. The rest of the model-evaluations, however, were carried out using a script provided by other students in the group (https://github.com/Aitslab/corona/tree/master/prototypes). This script was used to evaluate the precision and recall of both the original BioBERT models and the models built from custom databases when dealing with coronavirus-related data. Precision and Recall can be defined as:

$$Precision = \frac{\#ofTruePositives}{\#ofTruePositives + \#ofFalsePositives}$$

$$Recall = \frac{\#ofTruePositives}{\#ofTruePositives + \#ofFalseNegatives}$$

The evaluation was made with a manually annotated Gold Standard Corpus consisting on 10 publications with specific mentions of COVID-19 and the SARS-CoV-2 virus [17]. The results obtained by the models for these articles were then compared to the manual annotations in order to evaluate the precision and accuracy of each model.

Results

The first step of the project was to re-create the fine-tuned BioBERT models described in the original BioBERT paper. The results depicted in Table 1 represent the evaluations obtained for the BioBERT models when tested using the test files used in the original BioBERT publication. The results of the article were reproduced for all three of the tested models with small variations from the original scores. A second evaluation was performed with the test file of another dataset with the same tag type for the NCBI-Disease and JNLPBA models. The second evaluation showed a lower precision and recall, as is usually the case.

Table 1: Results for Biomedical Named Entity Recognition

Model	**Test Data**	**Precision (Reproduced)**	**Precision (BioBERT)**	**Recall (Reproduced)**	**Recall (BioBERT)**
NCBI-Disease	NCBI-Disease	93.64%	88.22%	93.76%	91.25%
	BC5CDR-Disease	68.94%		67.59%	
JNLPBA	JNLPBA	70.81%	72.24%	83.03%	83.56%
	BC2GM	51.01%		65.71%	
BC5CDR-Chem	BC5CDR-Chem	86.58%	93.68%	89.38%	93.26%

Note: the Model and Test Data columns specify which model and test-dataset was used. The columns marked as 'Reproduced' contain the values obtained in this project, while those marked as 'BioBERT' contain the values given in Lee et al. [8].

The same models were later used on the Gold Standard Dataset [17] which contained COVID-19-related abstracts. In addition, 4 models were fine-tuned using COVID-19-specific silver standard corpora generated with dictionary-based taggers that detect SARS-CoV2 and COVID-19 synonyms. These models were also used to perform predictions on the same Gold Standard Dataset. The results were then evaluated using a custom script that compared PubAnnotation output files in a tag-by-tag basis. As shown in Table 2, the models fine tuned with a combination of the original NCBI-Disease data-set and the Silver Standard Corpora seems to yield the best results.

Table 2: Results for Biomedical Named Entity Recognition

Tag-Type	Model	Test Data	Precision	Recall
Disease	NCBI-Disease	Gold Standard	31%	35%
	Dict-Disease		26%	30%
	NCBI-Disease + Dict-Disease		33%	28%
Protein	JNLPBA		17%	18%
	Dict-Protein		0%	0%
	JNLPBA + Dict-Protein		0%	0%

Note: results on the Gold Standard data-set [17] for each of the models.

Discussion

Reproducing the results of Lee *et al.*

BioBERT is one of the best performing BioNLP models for NER. Considering it is publicly available and it can also do other downstream tasks such as Relationship Extraction, it made sense to asses its capabilities as a tool for COVID-19 related NLP. It is worth mentioning, however, that there are other tools available for BioNLP. One example is the Flair framework [18].

The results obtained when reproducing the steps presented in the BioBERT article by Lee *et al.* [8] are close to the ones reported in the paper. The slight difference in the values could be explained by the parameters used when performing the testing, as values such as the number of epochs or maximum sentence length were not reported and therefore a mismatch in the parameters is likely.

It is interesting to note that the values for the model fine-tuned with the NCBI-Disease model are slightly higher than those stated in the original publication. This, again, was most likely caused by small discrepancies in the training parameters of the models.

When using the test sets of other databases of the same tag type, the results of the NCBI-Disease model and the JNLPBA model, while good, are significantly worse. This seems to indicate a mismatch in the annotation style of the databases. Unsurprisingly, the models perform best with publications coming from the same database as the papers they were trained with. It is typical for NLP models and Machine Learning models in general to not generalize very well.

New models

Neither the base model fine-tuned with the NCBI-Disease data-set nor the models fine-tuned from the results of dictionary-based tagger on the 100-paper subset are very good. In general, the models fine-tuned for this project are still not good enough for real-world applications. The BioBERT architecture has shown promise in other fields [8], so it is unlikely that this is the cause of the limitations. A possible cause is the wide scope of the most of the data-set annotations, resulting in a high amount of false positives. For example, the model trained with NCBI-Disease detects all kinds of diseases and symptoms in their predictions, but only COVID-19-related terms are annotated in the Gold Standard data-set.

The data-sets built from the dictionary-taggers are likely to be a limiting factor as well. The disease dictionaries lack symptom terms, and when evaluated as part

of another student project, the disease tagger has 0.8 precision and 0.31 recall. The protein tagger has not yet been evaluated properly. In the protein-related models, the base JNLPBA model does a really poor job in the Gold Standard data-set. The model fine-tuned with dictionary data, however, performs even worse, not getting a single true positive. Although this is probably affected by the lack of protein tags in the Gold Standard Database. It is possible that the true precision and recall values for all protein models would be slightly higher.

Nevertheless, the results show certain promise for the application of BioBERT NER in the field of BioNLP. A model trained on a corpus built from imperfect dictionary taggers, —which can be quickly generated in a health crisis unlike an expensive, manually curated Gold Standard— gives almost similar results to the BioBERT model trained on a Gold Standard from another domain. On top of that, combining both seems to allow to slightly boost precision. This means that using Silver Standards has potential when no Gold Standards are available and new tools need to be deployed quickly. It is also possible that training with a larger silver standard, such as the entire CORD-19 data-set could further improve results.

Limitations and future work

The models used in this project can be considerably improved upon. One of the unresolved issues encountered when working with BioBERT was its limitation when dealing with long sentences. If a sentence is longer than 512 tokens long, every token coming after that threshold gets ignored by the model and is not considered for prediction. Circumventing this issue would probably be possible with appropriate additional pre- and post-processing of the data.

As shown by this project, the precision and recall of a fine-tuned BioBERT model depend considerably on the data used in training. As such, having access to more, higher quality data would benefit the effectiveness of the model as well. Furthermore, the evaluation results were influenced by the limitations of the Gold Standard Corpus, which unfortunately included only a limited amount of tags of each of the types, and particularly few protein/gene tags.

That being said, the best possible results for COVID-19-related NER would most probably come from a combination of different approaches, not just the BioBERT model. Dictionary-based taggers, in particular, seem to be a powerful tool to combine with neural networks.

Acknowledgements

My thanks to supervisor Sonja Aits for her advice and guidance during the development of this project. Thanks to Prof. Pierre Nugues, Dr. Marcus Klang and to Salma Kazemi Rashed for their time and help. Finally, thanks to Emil, Peter, William, Annie, Sofi, Viktor and Rasmus, who worked in parallel on related projects, as well as to everyone else whose code or data I used during this project.

Bibliography

[1] 'All that's fit to preprint'. In: *Nature Biotechnology* 38.5 (May 2020), pp. 507–507. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0536-x. URL: https://doi.org/10.1038/s41587-020-0536-x.

[2] Vaishali M Kumbhakarnaa, Sonali Kulkarni and Apurva D Dhawaleb. 'Clinical Text Engineering Using Natural Language Processing Tools in Healthcare Domain: A Systematic Review'. In: *Dhawaleb, Apurva, Clinical Text Engineering Using Natural Language Processing Tools in Healthcare Domain: A Systematic Review (March 28, 2020)* (2020).

[3] Chung-Chi Huang and Zhiyong Lu. 'Community challenges in biomedical text mining over 10 years: success, failure and the future'. In: *Briefings in bioinformatics* 17.1 (2016), pp. 132–144.

[4] Basemah Alshemali and Jugal Kalita. 'Improving the reliability of deep neural networks in NLP: A review'. In: *Knowledge-Based Systems* 191 (2020), p. 105210.

[5] Jacob Devlin et al. 'Bert: Pre-training of deep bidirectional transformers for language understanding'. In: *arXiv preprint arXiv:1810.04805* (2018).

[6] Iz Beltagy, Kyle Lo and Arman Cohan. 'SciBERT: A pretrained language model for scientific text'. In: *arXiv preprint arXiv:1903.10676* (2019).

[7] Kexin Huang, Jaan Altosaar and Rajesh Ranganath. 'Clinicalbert: Modeling clinical notes and predicting hospital readmission'. In: *arXiv preprint arXiv:1904.05342* (2019).

[8] Jinhyuk Lee et al. 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining'. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[9] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[10] Rezarta Islamaj Doğan, Robert Leaman and Zhiyong Lu. 'NCBI disease corpus: a resource for disease name recognition and concept normalization'. In: *Journal of biomedical informatics* 47 (2014), pp. 1–10.

[11] Jin-Dong Kim et al. 'Introduction to the bio-entity recognition task at JNLPBA'. In: *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*. Citeseer. 2004, pp. 70–75.

[12] Xuan Wang et al. 'Cross-type biomedical named entity recognition with deep multi-task learning'. In: *Bioinformatics* 35.10 (2019), pp. 1745–1752.

[13] Gamal Crichton et al. 'A neural network multi-task learning approach to biomedical named entity recognition'. In: *BMC bioinformatics* 18.1 (2017), p. 368.

[14] Lucy Lu Wang et al. *CORD-19: The COVID-19 Open Research Dataset*. 2020. arXiv: 2004.10706 [cs.DL].

[15] Erik F Sang and Fien De Meulder. 'Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition'. In: *arXiv preprint cs/0306050* (2003).

[16] Jin-Dong Kim and Yue Wang. 'PubAnnotation-a persistent and sharable corpus and annotation repository'. In: *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. 2012, pp. 202–205.

[17] Salma Kazemi Rashed, Johan Frid and Sonja Aits. 'English dictionaries, gold and silver standard corpora for biomedical natural language processing related to SARS-CoV-2 and COVID-19'. In: *arXiv preprint arXiv:2003.09865* (2020).

[18] Alan Akbik, Duncan Blythe and Roland Vollgraf. 'Contextual String Embeddings for Sequence Labeling'. In: *COLING 2018, 27th International Conference on Computational Linguistics*. 2018, pp. 1638–1649.