Relationship extraction using BioBERT

Rasmus Lindqvist
mas15rli@student.lu.se

Viktor Bard
mas15vba@student.lu.se

Abstract

Research projects around the world are trying to understand the COVID-19 virus, but with the large amounts of information available this can be a difficult, time-consuming task. Relation Extraction (RE) can be a useful tool to more efficiently extract relevant information about corona-virus from large amounts of bio-medical text. This paper is part of a larger project under Aits Lab, that aims to develop a complete tool for information extraction from corona-virus texts. Using BioBERT, a state-of-the-art bio-medical text mining tool, this paper develops, and evaluates, a framework for relation extraction on corona-virus articles.The framework presents results comparable to those of the authors of BioBERT, F1 score of 80.32 and 79.89 on two test corpora, GAD and EUADR. Furthermore, the paper results in a framework able to find relations in CORD-19, a corona virus data set. Even though manual inspections show promising results on RE for CORD-19, no metrics are available here as no annotated data exists. The paper shows that BioBERT can be a useful tool in bio-medical RE, but that it requires extensive hardware.

1 Introduction

The recent outbreak of COVID-19 has created a global rush towards a better understanding of corona-viruses, to eventually find treatments and vaccines. However, the large amounts of articles related to corona-virus makes it virtually impossible for a human being to extract all the useful information in these texts. The goal of this paper is to develop, and evaluate, a method to mine bio-medical texts for information related to corona-viruses.

This paper is the one of the resulting papers of a research project under Aits Lab, where a group of students has been working on different aspects of retrieving information from biomedical research papers using natural language processing (Aits, accessed May 27, 2020).

1.1 Natural Language Processing

Natural language processing (NLP) is the process of making a computer analyze large of amounts of linguistic data. The field of NLP contains several subareas, with many methods and algorithms available. This article will focus on the task of relation extraction mainly, but named entity recognition is also mentioned as it is closely related to our task.

Named entity recognition (NER) is the task of identifying and classifying unstructured text into categories. The categories in this paper is mainly gene and disease, but with a final goal to expand into more categories; for example virus and protein, to mention a few.

Relation extraction (RE) is concerned with the task of extracting semantic relations in a text. Using RE on a large data set can be a useful tool to filter out important information in a text. An example of RE in bio-medical text could be: Gene A promotes disease B. Where A and B are named entities. Relation extraction is the task of understanding that A interacts with B.

1.2 Language model

As mentioned earlier, there are many different models and algorithms in the area of NLP. With the development of more powerful GPU's in the last decade and with that, the possibility of large matrix calculations, different neural networks algorithms has risen to perform especially well in the domain of NLP.

The model chosen and evaluated in this paper is a open source model called a BioBERT (Lee et al., 2019). It is an extension of BERT, which is neural network language model developed by Google (Devlin et al., 2018). BioBERT has been trained on bio-medical text and achieves state-of-the-art

performance on bio-medical text mining tasks such as NER and RE. BioBERT is trained in two steps. The first step is a pre-training step where it uses unsupervised learning, on large amount of unlabeled text, to get a general language understanding. In the second step the model is fine-tuned for specific tasks, such as NER or RE. Pre-trained version can be found on GitHub (DMIS-Lab, accessed May 27, 2020). The second step is supervised which means there is a need for annotated text. Though since the model is already pre-trained, the model can often achieve good results with even a small fine-tuning data set. (Lee et al., 2019)

1.3 Data sets

1.3.1 Training sets

To fine tune BioBERT we used two different corpora, GAD and EUADR, available from the BioBERT github page. GAD is a semi-automatically annotated corpora containing relations between genes and diseases (GAD, 2004). EUADR is a much smaller, manually annotated, corpora containing relations between drugs, disorders and genes (van Mulligen et al., 2012).

1.3.2 Test sets

As mentioned earlier, the goal is to build a framework for relation extraction on articles related to corona-virus. The data set that will be used is given by the kaggle challenge "COVID-19 Open Research Dataset Challenge (CORD-19)" (Allen Intitute for AI, accessed May 27, 2020) and consists of a large number of biomedical research papers on the past and current corona viruses.

GAD and EUADR is also used for evaluating the performance of BioBERT.

1.4 Project scope

The task governing this paper is relation extraction between different named entities found by other research groups in the larger project by Aitslab. The NER is performed on the CORD-19 data set by another group in the research project and is shared with us. The goal of this project was to establish a framework for RE using BioBERT. The vision of the finished product is to be able to search for any given entities in the data set and find the relations between them but in this report were using genes and diseases for evaluation on the performance of the BioBERT model and extracting relations in the CORD-19 data set for a manual inspection of the tagging.

2 Method

Our workflow basically consisted of two parallel processes, one more focused towards evaluating the performance of BioBERT in general and one more focused toward applying BioBERT relation extraction to the CORD-19 data set.

2.1 BioBERT environment

The nature of BioBERT, is that it is very computationally heavy, so in order to run it in reasonable time a large GPU RAM is needed. Our solution to this was to setup a Google Colab environment. Google Colab is online notebook that execute code on Google's cloud servers. Normally around 11 to 15 GB GPU RAM is available for free. As everything is reset between sessions on Colab, an efficient way of setting up the BioBERT environment was desired. The final solution was to clone BioBERT directly from GitHub and have the rest of the files saved on a google drive that was setup for this project. In this way, everything could be saved to Google Drive between sessions.

2.1.1 Fine-tuning

We fine-tuned the model on the specific task of recognizing relations between genes and diseases. To fine tune the model we used GAD corpora, containing 5330 relations between genes and diseases, and the EUADR corpora, containing 355 relations between drugs, disease and genes. In the original article of BioBERT these data sets are split into 10 smaller ones each. To achieve comparable results, this is also what we did. This meant that base model was fine-tuned on 20 different train sets, so we ended up with 20 different models.

The fine tuning was run in three different number of epochs to be able to investigate the performance versus computation time depending on the epochs. The three tests were run with 3, 5 and 10 epochs and we decided to run all models in 5 epochs since it gave the best result in approximately 6 minutes when 15GB RAM were allocated on Google's servers.

2.1.2 Model evaluation

All models were evaluated based on their precision and recall on both training sets to see how they performed withing the corpora it is trained and on a separate corpus. They were then compared to the results stated in the BioBERT article (Lee et al., 2019).

2.2 CORD-19 prediction

2.2.1 Pre-processing

To be able to make the models predict relations on the CORD-19 data set, some pre-processing was required. Firstly, two folders with NER PubAnnotation files, provided by another project group, were combined into one (PubAnnotation, accessed May 27, 2020). Combining PubAnnotation files is also an important part for the larger project, and our code is available on GitHub (NLP-relationextraction, 2020).

In order to run predictions on the combined PubAnnotation files, they were converted to the input format of BioBERT. BioBERT requires a .tsv format with a sentence with two tokenized entities for each row, like the following example.

"As to @DISEASE$, it is well established that @GENE$ have an important role in viral replication and de novo virus production"

2.2.2 Prediction

The pre-processed .tsv file was run on the 10 fine-tuned GAD models to generate predictions on relations in the sentences. The 10 GAD models were chosen as they showed the best out of sample performance, compared to the EUADR models. The output of BioBERT is a probability of a relation, scaling between [0,1]. Using a threshold of 0.5, these probabilities were converted to a label of 1, meaning a relation exists, or 0, meaning no relation exists.

2.2.3 Post-processing

Since predictions were made with 10 models, we had 10 prediction results. These were compared by taking the union and intersection and then converted back to PubAnnotation format, with the predicted relations added.

2.2.4 Evaluation

No labeled data on relations exists in the CORD-19 data set so a metric evaluation was not possible. Instead, a manual inspection of the results was performed with the help of Sonja Aits.

3 Results

3.1 BioBERT evaluation

The results of running the fine tuned models on the test sets, the GAD- and EUADR-models yielded the results as can be seen in table 1. The presented metrics are a mean value of the 10 generated models of each training set.

Model	Test set	F-score	Recall	Precision
GAD	GAD	80.32%	83.40%	77.55%
GAD	euadr	75.49%	87.64%	74.12%
euadr	GAD	24.02%	33.06%	28.39%
euadr	euadr	79.89%	83.21%	78.41%

Table 1: Evaluation metrics of test sets

3.2 CORD-19 results

When running the model pre-trained on GAD on the CORD-19 subset of 100 articles, the following results were achieved. The numbers in table 2 are the total number of sentences with a relation based on the union, intersection and average of the 10 models' prediction.

Total sentences	$\cup$	$\cap$	Average
1764	854	71	402

Table 2: Number of relations found in CORD-19 test set

3.2.1 Manual inspection

The following sentences are a small sample of the relations that were unanimously tagged by the 10 models.

- *Therefore , further studies , with a larger number of samples , are required in order to better establish the role of @GENE$ as a potential biomarker for @DISEASE$ progression .*
- *In the model describe here ; the animals were immune competent for HCV ; therefore , our findings provided further important evidence that @GENE$ was effective in the treatment of @DISEASE$.*
- *With this aggressive form of cancer on the rise , it is highly plausible the future treatments for @DISEASE$ could involve targeting @GENE$.*

A small manual inspection of 25 relations were done with the assistance of Sonja Aits. As can be seen in table 3, 12 out of 25 relations were correct. As seen in table 4, out of the 12 faulty relations, 4 were true relations but wrong entities, 2 were both wrong entities and relations and 7 were true relations but wrong relations.

True relations	12
Wrong relations	13

Table 3: Manual inspection results of the CORD-19 RE

True RE wrong NER	4
True NER wrong RE	7
Wrong NER and wrong RE	2

Table 4: Statistics of wrong relations tagged

4 Conclusion

4.1 BioBERT performance

Looking at the evaluations metrics of the fine tuned models we achieved comparable results as the authors of BioBERT (Lee et al., 2019). The small variations is most likely due to different number of epochs run while fine tuning. In the article, there are no metrics of when the models are run on other test sets than their respective training data came from. When running the GAD model on the EU-ADR test set the metrics are still comparable to the results of when it's run on its original data set. However, the same can not be said about the EU-ADR model whose metrics are far worse when ran on the larger, GAD, test set.

The numbers presented on BioBERT's relation extraction on CORD-19 are showing that although all the models are trained on the same corpora (different parts of it) it does not mean that they will have seen and are trained on all possible relations. All 10 models are unanimously agreeing on only 8.3% out of all relations.

4.2 Named Entity Recognition with Relation Extraction

An important note on the results of RE for it to be useful in real text mining operations is that the NER is just as important as the actual RE. If the tagged entities are entirely wrong, the relation extraction would try to find relations between irrelevant entities yielding an unusable result. This phenomena can be seen in table 3, as 6 faulty entities resulted in faulty relations.

4.3 Issues

This project was mainly troubled by lack of time and computer power. As earlier mentioned in the article, BioBERT demands quite extensive hardware to be able to run efficiently. Running the BioBERT environment on Google Colab was manageable but definitely not perfect. Depending on the availability on the servers we achieved different amount of GPU RAM allocated which drastically impact the time training a model. At full capacity, we experienced it ran as much as 4 times faster than other times. Preferably the fine tuning would be run locally if we had the right hardware. Further more, handling all the generated files and models via Google Colab turned out to be more work than initially expected since the only way to save them were through a Google Drive.

Regarding the time issue, since there currently are no annotations available of the CORD-19 data set it is difficult to easily evaluate the performance of the RE. Right now the evaluation has to be done manually by checking the relation annotations in the original texts which is a time consuming matter.

4.4 Future work

The next step to further evaluate the performance on the CORD-19 data set we consider is to produce an accurately annotated test set of relations in the data set. With proper annotations it would be possible to create valuable metrics regarding the performance of the models on the specific data set. Furthermore, it would be possible to evaluate both a single models as well as the union and intersection of all models created in this project.

We have not investigated properly why the model trained on the EUADR training set performs so poorly on the GAD test set. One reason for this could be that the data set is too small to be able to train the model properly. It could be a good experiment to re-train models on this training set with different epochs to see if this has an impact on the performance.

As the project developed we continuously discovered different areas of format handling, e.g. json-file to tsv-file or combining data. We created separate scripts for each occasion it was needed. In hindsight many of these scripts can be combined and integrated into the BioBERT environment to reduce the number of steps running the program.

Acknowledgments

We would like to express our gratitude towards our supervisors Sonja Aits and Pierre Nugues whom have been a great support and aid during this project. Furthermore, we would like to thank the rest of the larger project group, who it has been a great pleasure to work with.

References

Sonja Aits. accessed May 27, 2020. *Aits Lab*.

Allen Intitute for AI. accessed May 27, 2020. *COVID-19 Open Research Dataset Challenge (CORD-19)*.

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

DMIS-Lab. accessed May 27, 2020. *BioBERT*.

Genetic Association Database GAD. 2004. *The Genetic Association Database*, volume 36. Nature Publishing Group, Englewood Cliffs, NJ.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining.

Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. 2012. *The EU-ADR corpus: Annotated drugs, diseases, targets and their relationsships*, volume 45. Elsevier.

NLP-relationextraction. 2020. *Our project github*.

DBCLS PubAnnotation, by Database Center for Life Science. accessed May 27, 2020. *PubAnnotation*.