CMPE 492 PROJECT REPORT

Authors:

Volkan BULCA

İbrahim Özgürcan ÖZTAŞ

Advised by

Assoc. Prof. Arzucan ÖZGÜR

Date: 1 July 2021

Table of Contents

1. Introduction and Motivation **3**

2. State of the Art **4**

3. Methods **6**

4. Results **8**

5. Conclusion and Discussions **9**

6. Future Work **10**

7. References **11**

Foreword

We would like to thank Assoc. Prof. Dr. Arzucan Özgür for infinite guidance and continuous support in our journey of project development.

Also, we would like to thank Hilal Dönmez for brainstorming and contributions to our project.

1. Introduction and Motivation

Information is the dominant power in the concurrent era since every action, every decision has to be an outcome of a process based on a set of information. Even in a daily action, we use our experience which is the lifelong collection of the outcomes of the events in which we have been a part of. As the importance of information increases due to rapid development in the global network and the internet infrastructure, the era of information has brought many advancements, especially the improvements in machine learning and its contributions to other areas. Natural Language Processing, Deep Learning, Artificial Intelligence and Neural Networks are the most popular outcomes of information retrieval and processing. Hence, with the new tools of bulk computation, deliberate topics such as computation of biological information and biological analyses gained accessibility with scarce resources.

Biological Natural Language Processing (BioNLP) has become a key area since it is a multidisciplinary area, a combination of biology and computer science. By applying the text analysis methods in a biological text or a sequence of characters consisting of genes or proteins, a huge set of information can be retrieved. It may be used to predict the outcome if there's a need for classification or identification of a new instance in mere seconds.

In this graduation project, we've decided to compose a BioNLP project. With the guidance of our advisor Assoc. Prof. Dr. Arzucan Özgür, we've become a contestant of the BioCreative VII competition. Out of five different topics to study, We've selected Track 1: Text mining drug and chemical-protein interactions (DrugProt) as the topic of our project and we will pursue this track.

Our motivation for selecting this competition and this track is because both of us want to do a research based project rather than a purely practical outcome. We thought that researching biological texts and creating a BioNLP project would be beneficial for us. At first, we researched many articles to seek inspiration about our method and our aim. We've selected a paper[1] which helped us find our purpose and clues of our possible methods that may be useful for our study. Then, we discussed our research with our advisor and she offered this competition based on BioNLP with different topics at each track. After a decision process, we've decided to continue with the track in which the relations between drug-gene pairs are inspected.

2. State of the Art

We have analyzed the academic paper: “Vapur: A Search Engine to Find Related Protein - Compound Pairs in COVID-19 Literature”[1] to understand the knowledge curated in this domain of research. As we aim to find drug-protein interactions in the literature of biomedicine, we have extracted and adopted several state-of-the-art conclusions from the paper mentioned before.

Below, there are several conclusions listed as paragraphs one by one, which we adopted for our project. At each paragraph, there’s the conclusion with its explanation of why we adopted it.

In recent years, biomedical literature experienced a huge growth. Therefore, indexing all papers related to the biomedical area becomes a challenging task to deal with. According to the paper, “Recently, transformer-based NER models attracted more attention, including BERN, which is a state-of-the-art biomedical named-entity recognition and normalization tool that uses BioBERT to identify and normalize the entities in a sentence.“[1] is stated. In the light of such a conclusion, we have been convinced to utilize BERN in our project as a named entity recognition extraction and normalization tool. In recent studies mentioned in the paper, BERN is adopted as the text-mining tool for COVID-19 related articles. Since our task is to find drug-protein pairs in papers of biomedical literature, we need a relation extractor and normalizer for biomedical purposes and BERN is the state-of-the-art tool that can be used in the way we need to use.

One such application of BERN is “Vapur”, a search engine focusing on finding protein-chemical relations for COVID-19 literature. As it is mentioned in the paper that we study, “Vapur is able to retrieve relevant documents to a query as categorized by the biochemically related entities thanks to its relation-oriented inverted index.”[1]. As the paper suggests, the indexing requires identification and normalization of the named entities in papers by using BERN as a pre-trained model. To find the same outcome for different representations of the same entity, Vapur creates equivalence classes for entities which store different representations of each entity and this is learned from the recognition and normalization steps of Vapur. As a result, Vapur provides a flexible search environment given the search query because these equivalence classes have been created from a wide range of mention types from free-text to chemical IDs [1].

As the paper provides, “Vapur represents each entity mention as a string and adopts a 3-gram based matching algorithm to search the queried entity in its index. Given a query, Vapur first creates a multi-set of all 3-grams of the query and computes the similarity of this set to all 3-gram multi-sets of the mentions in the index, which are pre-computed.“[1]. We believe it is still beneficial to have a form of n-grams, created by the query since many articles in the biomedical area may have used the terms in the search query, but in a different order. To retrieve similar articles in our task, applying n-grams to the search query and collecting results is the way we prefer to implement. Also, having typos in a query would decrease the precision of our system, however Vapur overcomes this issue by utilizing generalized Jaccard similarity in the search query which leads to a more powerful approach for article retrieval.

In addition to the previous article, we also consulted another academic paper: “Deep learning of mutation-gene-drug relations from the literature” and we observed that using named entity recognition systems in automated information extraction processes would be beneficial for many aspects. We’ve seen that the authors have manually curated a database containing mutation-gene-drug-disease relations based on the expert’s knowledge. Also, the authors have preferred a deep learning model with sophisticated configuration for the extraction of pairs designated as outputs of the analysis [2].

For further research on the BioBERT model, we read the article: “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. According to [4], BERT is a very robust model for NLP and text mining. However, it contains a very generic corpus, hence its approach of entity recognition is not that powerful in a specific domain. BioBERT, on the other hand, utilizes a very similar approach to BERT but it was trained on not only generic datasets but also biomedical datasets containing PubMed articles and other field-specific data. As a result, the BioBERT model can recognize and perform better in the biomedical field. The improved results are discussed in [4]: BioBERT outperforms currently popular models in six of the nine widely-used biomedical related datasets in terms of its micro f1 score in named-entity recognition(NER) tasks, achieved higher f1-scores in relation extraction(RE) datasets like CHEMPROT, and yields highest accuracy scores on all question answering(QA) tasks.

3. Methods

In our midterm report we stated that: “As we have become a contestant of the BioCreative VII competition, the planned schedule implies that the training data set and other utility items will be released on May 14th, 2021. Therefore, we could not offer a concrete structure of methods that we plan to use as a solution. When the training data set and other utility items are released, we will start to construct a solid strategy for ourselves to compete in this competition.”

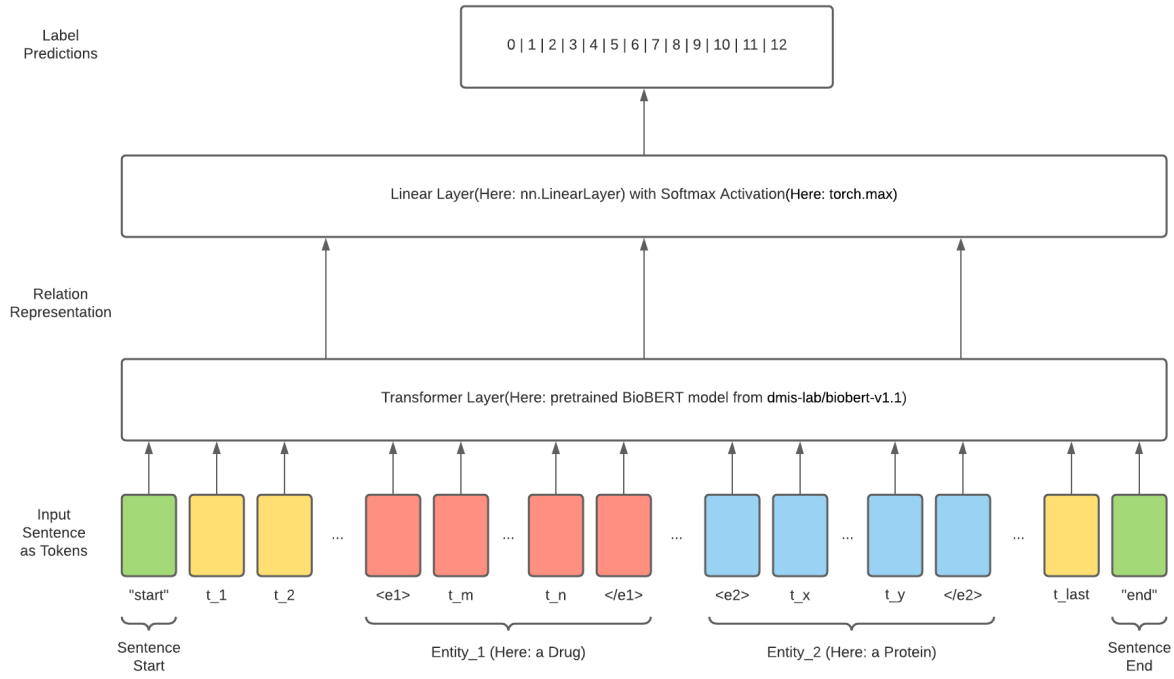
However, since the release of the DRUGPROT training set[3] which is the main dataset for the Track 1 of the BioCreative VII competition, was postponed twice, we decided to work on and study the relations and characteristics of the dataset given in the one of the previous BioCreative competitions: Chemprot. We started constructing our baseline model using the latter. With the help of our advisor, we directly started implementing a model using BioBERT, which is known to be a very powerful pre-trained model on chemical relation extraction. During our studies, since the DRUGPROT training set was finally released, we immediately shifted our efforts on adapting the baseline model to the new dataset. Luckily, both datasets are very similar to each other in terms of structure, although the labels for the relations between chemical-protein entities are different.

Before training the model, we first processed our raw data. Our first step is to split the abstracts of the PubMed articles given in the training dataset into its syntactically and semantically correct sentences. To achieve this, we used the GENIA Sentence Splitter [5]. It takes the complete abstract text as input and produces a list of sentences, including the PubMedID, title and abstract content.

After the first step, we continued with marking and expanding the entity pairs of each sentence. Initially, we extracted and saved the offset positions of both the start and end point of each entity in the sentences and registered the sentence number that includes the entity to be used in further steps. Then, we extracted the relations between entities from the given dataset and by using the previously registered offsets and the sentence counts of the entities, we found out which sentences contain multiple entity pairs and we expand them in a set of sentences which has exactly one unique pair. Hence, the number of sentences we expanded will be the same as the total available number of entities.

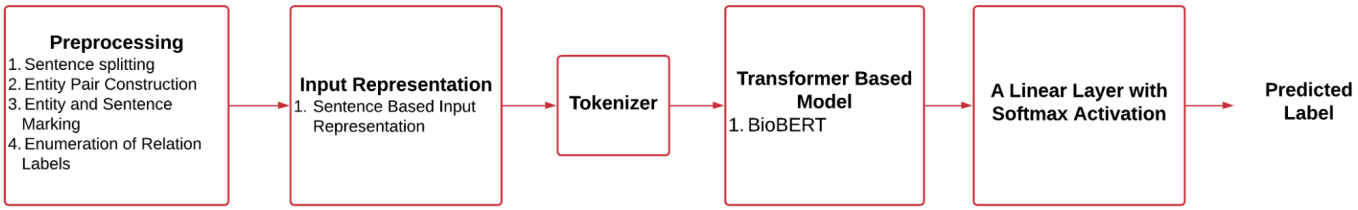
As our third step, we tokenized each sentence using the tokenizer of a pretrained BioBERT model [6]. However, the default tokenizer would only work for the words which exist in its vocabulary, and for the unrecognized words, the tokenizer splits the words even further until all sub-tokens are in its corpus. For our model to function, we have to insert marks before and after the entities as “<e1></e1>” and “<e2></e2>” so that the model can understand the words in between correspond to special entities in our dataset. Since the tokenizer does not recognize the entity markers, we’ve manually registered those markers as if they are defined in the corpus of the tokenizer. Furthermore, the structure of the sentences must be

preserved for the model; therefore, we also defined the start and end markers of the sentences to the tokenizer. We've found that some encodings are left unassigned to a specific word in the corpus, so we took and utilized them as the tokens of our special markers. Six indexes are available and sufficient to reassign the markers: {1, 2, 3, 4, 101, 102}, where {1, 2} are for <e1> and </e1> markers which specify the drug entities; {3, 4} are for <e2> and </e2> markers which are used for protein entities and remaining indexes are for start and end markers of the sentences. After the tokenization and entity marking, we've also created an attention matrix, a mask, to introduce our input list properly. Since the transformer-based models require which tokens are presented in their structure, we need to represent them explicitly. For any input token sequence, we've padded them until they have 512 tokens and their respective masks as well. Finally, we discarded 14 sentences having more than 512 tokens because BERT models are configured to have a token number limited to 512 at most. There are 17274 remaining sentences in our preprocessed version of our dataset, so discarding 14 sentences has a negligible effect on the training phase of our model.



The BioCreative VII competition Track 1 task has provided neither a development nor a test set yet, thus we had to split our dataset into training and test sets using a 80-20 ratio which seems like a fair split for us. Before splitting, we shuffled our data so that we eliminated the bias on the creation of train and test sets. After the split, we had to transform our data, labels and masks into tensors because the pretrained model accepts only the tensor type as input. Also, we transfer these tensors from CPU to GPU devices to accelerate the training process.

Now, the only task remains is to build and train our BioBERT model. At first, we initialized our BioBERT model as an AutoModel from the transformers library in Python3 and we've selected the version of "dmis-lab/biobert-v1.1" [6]. In our task, we've multiple relation labels to classify the entity pairs, thus we've added a linear layer after the model itself to properly classify. We've used "nn.Linear(768, 13)" as our linear classifier with parameters input_dimension=768 and output_dimension=13 due to the number of given unique relation types. In addition to that, we've selected an optimizer to optimize the loss in our model and we've preferred Adam from the "torch" library with learning rate 1e-5 and weight decay coefficient 3e-2. Along with those, we've calculated our loss with a multiclass cross entropy function from the "torch" library as "nn.CrossEntropyLoss()". Therefore, by defining our batch size as 8 and epoch count as 5, we've fine-tuned our pretrained BioBERT model.



4. Results

After training our model with the given configurations and the training set we created, we used the test set in the prediction phase of the model. Here, we used the ".eval()" function of the AutoModel. While doing predictions, for each input, mask, label tuple of the test data, we save the true label of it. Then, we run our trained model and get an output. This output contains several similarity metric values, which show the closeness of the prediction to each label. Among those output data, we need to take the label which corresponds to the largest metric; therefore, we take the maximum of it and assign it as our actual predicted label. Once every label is predicted for all the data, we simply look at the classification report of the prediction of our model.

From those predictions, we got a 0.77 accuracy value for our test set. The more detailed metrics including precision, recall and f1-score values for each label can be found in the table below.

Table 1: Model evaluation with lr=1e-5, wd=3e-2, only 1 iteration

	precision	recall	f1-score	# of true labels	# of predicted labels
0	0.86	0.84	0.85	1121	1094
1	0.58	0.80	0.67	148	206
2	0.63	0.76	0.69	416	502
3	0.68	0.70	0.69	244	254
4	0.67	0.73	0.70	266	289
5	0.89	0.88	0.89	204	202
6	0.77	0.66	0.71	244	210
7	0.83	0.84	0.83	144	146
8	0.84	0.72	0.77	504	433
9	0.66	0.54	0.60	146	119
10	0.00	0.00	0.00	7	0
11	0.00	0.00	0.00	1	0
12	0.00	0.00	0.00	10	0
accuracy			**0.77**	**3455**	**3455**

5. Conclusion and Discussions

As we further discuss our model, we know that transformer-based models are highly dependent on random selections, thus, one iteration of our model would yield a prediction result with a non-negligible variation. To overcome this issue, the model could have been run multiple times with the exact same dataset and hyper-parameters so that the effect of the randomness of the BERT token selection is further minimized when the results are averaged.

Secondly, we are curious about the effects of the hyper-parameters i.e. learning rate and the weight decay coefficient on the prediction results. Our single selection of hyper-parameters might not yield the most optimized outputs during loss minimization in training, hence could affect the predicted labels in the evaluation phase.

Finally, if we look at the prediction results in detail, we can see that when the number of relations corresponding to a label is small, the precision of predicting that label for a relation gets usually lower as well. This is probably because the model learns less about those labels while training and hence cannot properly predict that label when the model is fed with a different dataset. Also, as it can be seen, some labels did not even receive any attention from the model because the training set does not have sufficient number of relations corresponding to those labels.

In conclusion, extracting relations between entities in a given text requires delicate preprocessing and micro-scale adjustment. The set of generic models usually fails the task since the generalization leads to the diversity of the topic of the texts that leads to a decrease of focus on the target topic, whereas models like BioBERT are pre-trained with data sets containing vast amounts of either biomedical articles or its abstracts to increase the focus of training on the target topic.

As stated above, biomedical text mining depends on the models that have been constructed and fine-tuned with the set of biomedical articles, and BioBERT has offered such precise targeting on any biomedical context. Our project shows one instance of such an occasion that BioBERT can be utilized and provides fairly strong results. With further experimentations and future works such as explained below, biomedical text mining can be improved to a more precise and reliable state by the power of BioBERT models.

6. Future Work

As future work, we'd like to try out several different preprocessing approaches to reshape the training and test dataset. Since we've realized that the sentences in both the training and test datasets may include words and idioms that do not have any biomedical meaning, thus they are non-contributory for our model to function greatly. Our first future work is to rather than feeding our model with the unprocessed sentences, we would like to try to reconstruct the sentences with the most biologically contributory tokens and compare the result with our primary model.

In addition to that, we would like to inspect the given entity classes by creating graph networks that represent the clustered relations, which we think of as a supplementary decider for our BioBERT model. The graph should be constructed by the features of the drug groups and protein families. We'd like to research further on how to retrieve the protein family of any protein from various resources and include them in our structure. Also, finding the group of any drug is similar with the former statement, thus having both of these extra information may help us develop a more precise model with higher accuracy and F1-scores.

Even though these future works are based on theoretical reasons, we'd like to limit-test our model with different hyper parameters. By providing different learning rates and weight decay coefficients for our model and its optimizer, we'd like to observe which learning rate and weight decay coefficient suit our model. Also, we've learned that transformer based models are composed of random token guessing, therefore running our model 10 times and averaging the results of the 10 runs would be the logical approach for reporting the performance of our model.

We will continue working on this project for the BioCreative VII competition and seek to improve our results by trying the aforementioned suggestions of improvements. Below, the updated schedule for this competition can be found.

BioCreative Competition Schedule-Updated Version [3]

15.06.2021: DrugProt evaluation script and Training set release

05.07.2021: Test set abstracts and entity annotations release

30.08.2021: Test set prediction submission due

03.09.2021: Test set evaluation returned to participants

10.09.2021: Short technical systems description paper due

16.09.2021: Paper acceptance and review returned

27.09.2021: Test set Gold Standard annotations to participants

7. References

[1] A. Köksal, H. Dönmez, R. Özçelik, E. Özkırımlı, A. Özgür. “Vapur: A Search Engine to Find Related Protein - Compound Pairs in COVID-19 Literature” [Online], December 2020. Available: https://www.aclweb.org/anthology/2020.nlpcovid19-2.21.pdf

[2] K.Lee, B. Kim, Y. Choi, Sunkyu Kim, W. Shin, S. Lee, S. Park, Seongsoon Kim, A. C. Tan, J. Kang. “Deep Learning of Mutation-Gene-Drug Relations From the Literature” [Online], 25 January 2018. Available: https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/s12859-018-2029-1.pdf

[3] [Online]. Available: https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-1/

[4] J. Lee, W. Yoon, Sungdong Kim, D. Kim, Sunkyu Kim, C. Ho So, J. Kang, “BioBERT: a Pre-trained Biomedical Language Representation Model for Biomedical Text Mining”[Online], 10 September 2019. Available: https://arxiv.org/ftp/arxiv/papers/1901/1901.08746.pdf

[5] [Online]. Available: http://www.nactem.ac.uk/y-matsu/geniass/

[6] [Online]. Available: https://huggingface.co/dmis-lab/biobert-v1.1