
# CMPE 492 PROJECT REPORT

Authors:

Volkan BULCA

İbrahim Özgürcan ÖZTAŞ

Advised by

Assoc. Prof. Dr. Arzucan ÖZGÜR

Date: 7 May 2021

# Table of Contents

# 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 the 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've researched over 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, transformers-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 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 benevolent 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]

## 3. Methods

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 is released, we will start to construct a solid strategy for ourselves to compete in this competition.

## 4. Future Work

Since our future work planning depends on both the released schedule of the competition and the ideas that we provide as a solution, there is no certainty in our future plan right now.

Until the release of the training data set, we are only able to provide the schedule of the competition and other related information.

**BioCreative Competition Schedule [3]**

14.05.2021: DrugProt evaluation script and Training set release

05.07.2021: Test set abstracts and entity annotations release

12.07.2021: Test set prediction submission instructions and DrugProt

24.08.2021: Test set prediction submission due

01.09.2021: Test set evaluation returned to participants

07.09.2021: Short technical systems description paper due

13.09.2021: Paper acceptance and review returned

27.09.2021: Test set abstracts, entity annotations and relations

## 5. 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/