EXTRACTING KNOWLEDGE GRAPH OF COVID-19 THROUGH MINING OF UNSTRUCTURED BIOMEDICAL CORPORA

Guided By
Dr. G. Sudhakaran

T. Athiban - 2018103013
N. Prathesh - 2018103576
M.Syed Mohamed Asif - 2018103612

INTRODUCTION

COVID-19 is a global epidemic with a considerable fatality rate and a high transmission rate, affecting millions of people world-wide since its outbreak. The search for treatments and possible cures for the novel Coronavirus has led to an exponential increase in scientific publications, but the challenge lies in effectively processing, integrating and leveraging related sources of information.

Scientific publications regarding COVID-19 contains various data about related diseases, genes, drugs and so on. The data in such publications are vastly unstructured.

Most of the articles published under the title COVID-19 are gathered under the name of CORD-19. We introduce a fully automated generic pipeline consisting of an Information Extraction (IE) system followed by Knowledge Graph construction.

OVERALL OBJECTIVES

- To extract information from CORD-19 in a fully autonomous way using NLP techniques.
- To gather named entities such as diseases, genes, drugs from the CORD-19 dataset.
- To extract relations between entities (i.e., drug-induced-disease relations, drug-gene interactions, disease-gene interactions) from the CORD-19 dataset.
- To organize the found entities and relations in the form of Knowledge Graph.

LITERATURE SURVEY

The Covid-19 Open Research Dataset (CORD-19) is a growing resource of scientific papers on Covid-19 and related historical coronavirus research. CORD-19 is designed to facilitate the development of text mining and information retrieval systems over its rich collection of metadata and structured full text papers.

JNLPBA, NCBI, CHEMDNER, BC5CDR, CHEMPROT are gold standard datasets for the tasks of Named Entity Recognition and Relation Extraction in the biomedical Domain.

Named Entity Recognition (NER) is the task of finding entities across the document. NER can be done in both Supervised and Semi-supervised manner. BI-LSTM CRF (Huang et al., 2015), BI-LSTM-CNN (Chiu and Nichols, 2016), BI-LSTM-CRF (Lample et

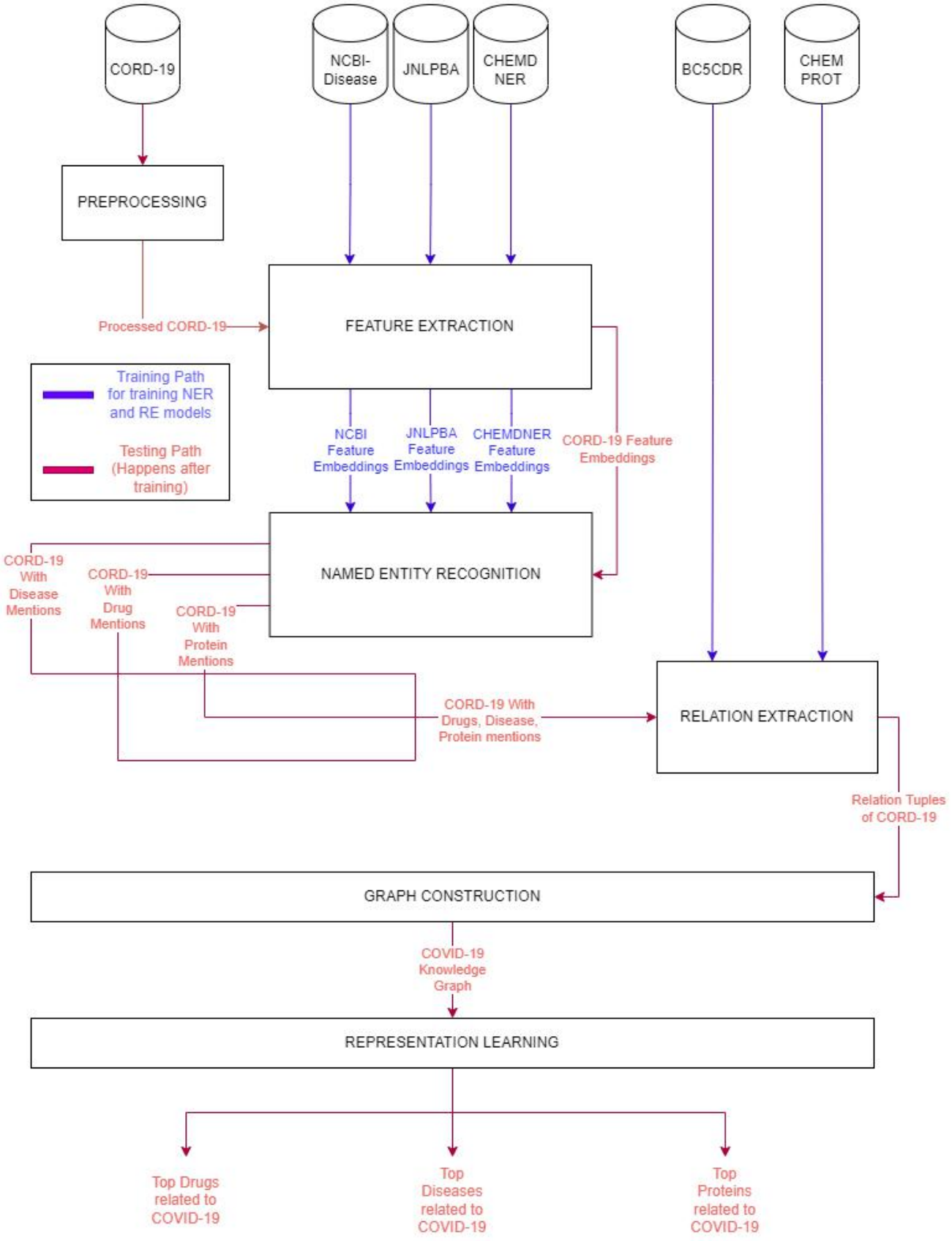
al.,2016) and LSTM-CNN-CRF (Ma and Hovy, 2016) are few such architectures.

Relation Extraction (RE) is the task of finding relations between two entities. Most of the recent RE systems use pretrained language models on unannotated text like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and XLNet (Yang et al., 2019).

KGs were immensely used in different fields like Life Science (Chen et al., 2009), Decision Support System (Russell and Norvig, 2010) etc. KGs in financial fields(Repke T., Krestel R, 2021) are used for investigative tasks such as legal monitoring and so on.

Smaller KGs have been constructed for COVID-19 like (Domingo-Fernandez ´et al., 2020) which covers 145 articles consisting of 3945 nodes and 9484 relations covering 10 entity types. Previously built KGs have also been employed for COVID-19 drug discovery (Richardson et al., 2020). KG on other COVID-19 datasets are also carried out using Deep Learning Methods (Kim, T.; Yun, Y.; Kim, N. 2021).

ARCHITECTURE DIAGRAM

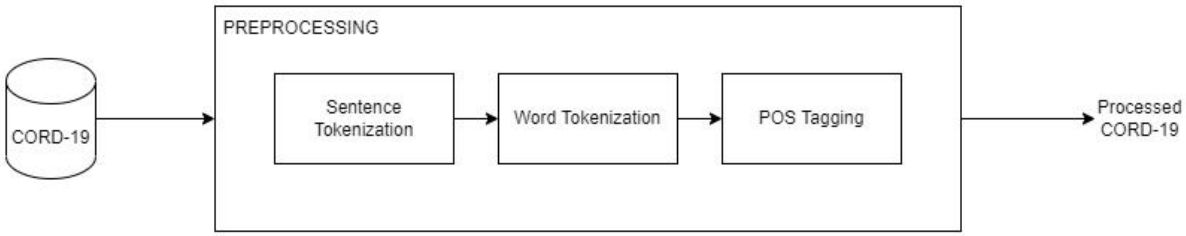


MODULES DESCRIPTION

There are 6 modules for this project. They are as follows

1. Preprocessing Module
2. Feature Extraction Module
3. Named Entity Recognition Module
4. Relation Extraction Module
5. Graph Construction Module
6. Representation Learning Module

1. PREPROCESSING MODULE

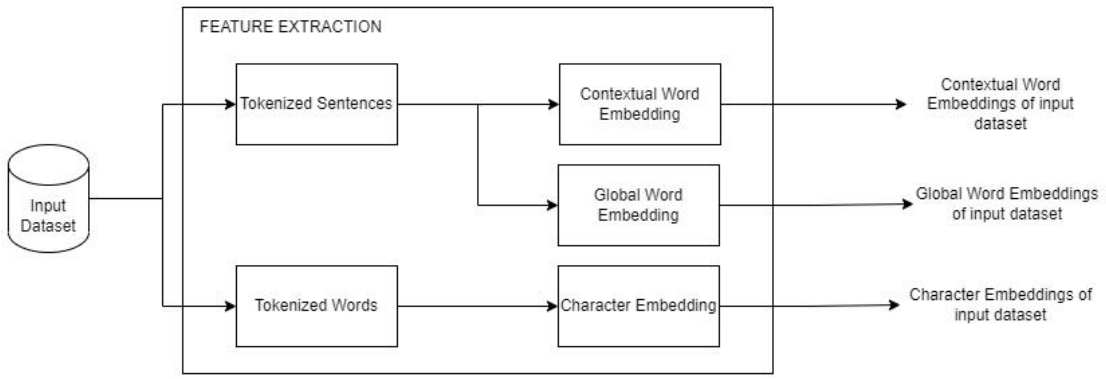


INPUT - CORD-19 Dataset
OUTPUT - Tokenized and Processed CORD-19

In this module, the CORD-19 Dataset is splitted into individual sentences using NLTK Sentence Tokenizer and these individual sentences are further tokenized using Word Tokenizer and the results are stored.

After the tokens are POS tagged for future use. Tokenized CORD-19 is necessary for Named Entity Recognition and Relation Extraction Modules.

2. FEATURE EXTRACTION MODULE



INPUT - NCBI-Disease, CHEMDNER, JNLPBA, Processed CORD-19

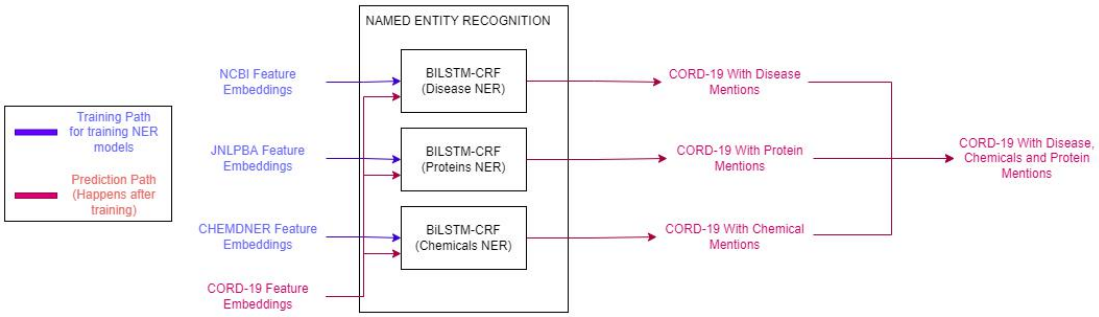
OUTPUT - Contextual Word Embedding, Global Word Embedding, Character Embedding of input

In this module, the datasets needed for Named Entity Recognition Module is fed as input.

These 3 datasets are initially preprocessed where all the sentences are padded so that they are of same length. Then we use contextual word embedding techiques such as ElMo or Transformer based word embeddings. The global word embedding is used for averaging multiple uses of same word in different contexts. Character embedding maps each character into corresponding vectors.

These 3 vectors are combined are the output of this module

3. NAMED ENTITY RECOGNITION MODULE



INPUT - Word and Character embeddings of NCBI-Disease, CHEMDNER, JNLPBA, Processed CORD-19

OUTPUT - CORD-19 with Disease, Drug, Protein Mentions

In this module, the NCBI-Disease dataset is used for recognition of diseases. CHEMDNER dataset is used for recognition of Drugs. JNLPBA dataset is used for recognition of Proteins.

The embeddings of each dataset are taken. Then the tokens and their corresponding named entity tags are associated.

Then each individual datasets are fed into a BiLSTM-CRF model, and the results are tested. Now there are 3 models, which are capable of finding diseases, drugs and proteins respectively.

Now the Processed CORD-19 is fed into each model and the entity tags of CORD-19 dataset are found.

The tags are for each word token that are encoded in BIO Scheme. Here B-Entity refers to the beginning of the entity, I-Entity

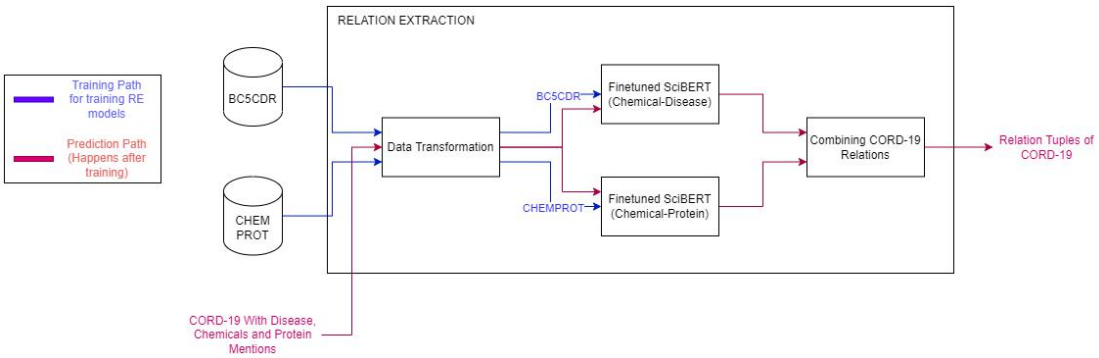
refers to the inside of the entity and O refers to the outside of the entity.

Example encoding,

```
The O
severe O
anemia O
( O
hemoglobin B-GENE
1 I-GENE
. I-GENE
2 I-GENE
g I-GENE
/ I-GENE
Dl I-GENE
) O
appeared O
to O
be O
the O
primary O
etiologic O
factor O
. O
```

These Tags are combined so that the final output contains CORD-19 dataset with all the diseases, drugs and proteins mentions.

4. RELATION EXTRACTION MODULE



INPUT - BC5CDR, CHEMPROT, CORD-19 with entity mentions
OUTPUT - Relation tuples of CORD-19

In this module, BC5CDR dataset is used for extraction of chemical induced disease relations. CHEMPROT dataset is used for extraction of chemical-protein relations.

The 2 datasets are preprocessed where the tokens are associated with its entities, and the sentences are processed as the first sentence contains the relation entities and the second sentence contains the text containing the relations.

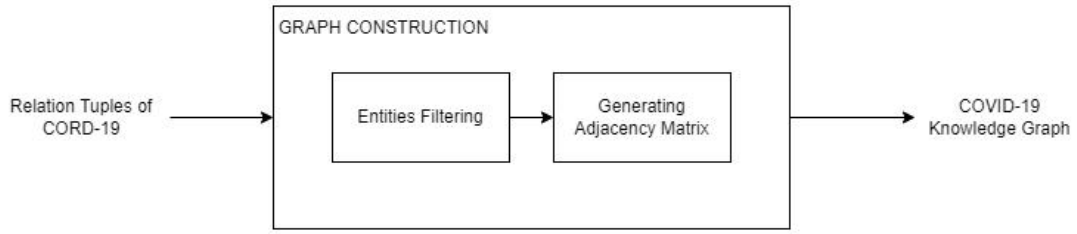
The Drug-disease relations dataset is fed into the SciBERT model with 1d-CNN output layer which produces the relations.

The CHEMPROT dataset is fed into individual SciBERT model which will be finetuned for finding relations in that particular dataset. Finally the performance of the models are measured.

Now the CORD-19 dataset with entity mentions is preprocessed where only the sentences with two or more entities are forwarded into the model. And based on the type of entities, the sentence is fed to one of two models and the model predicts whether a relation exists between two entities exists or not.

Finally the two models' outputs are combined and tuples are generated of the form (Entity1 ,Entity2) or (Entity1, Entity2, Relation).

5. GRAPH CONSTRUCTION MODULE



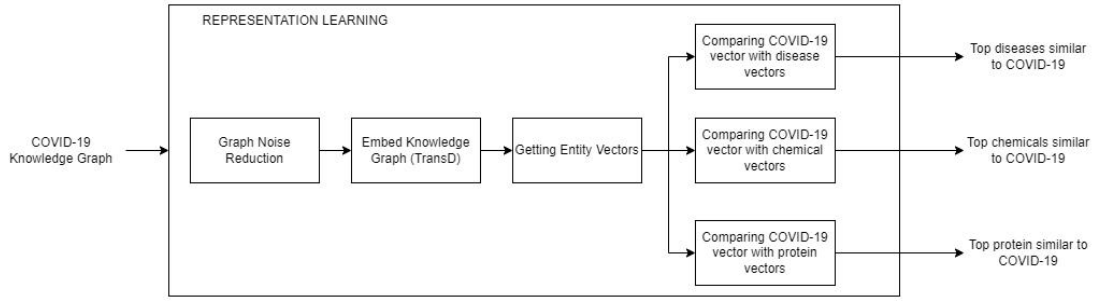
INPUT - CORD-19 Relations along with their entities
OUTPUT - COVID-19 Knowledge Graph

We construct a KG which is defined as KG = (E, R, G), where,

- E: a set of nodes representing disease/ protein/ drug entities
- R: a set of labels representing chemical-protein relation or chemical-disease
- $G \subseteq E \times R \times E$: a set of edges that represent facts connecting entity pairs.

Here entities with no relations or in-degree less than 5 (for example) can be removed which helps with the density of the resultant knowledge graph.

6. REPRESENTATION LEARNING MODULE



INPUT - COVID-19 Knowledge Graph
OUTPUT - Top Diseases, Chemicals, Proteins related to COVID-19

The noise of the input knowledge graph is reduced by removing entities with in-degree less than N. Then the knowledge graph is embedded using geometric method TransD.

After embedding the Knowledge Graph, the vectors of all entities in the Knowledge graph is obtained. From that, the COVID-19 vector is taken.

This vector is compared with all the remaining vectors using cosine similarity and top COVID-19 related diseases, chemicals and proteins are found and is returned.

DATASET DESCRIPTION

1. CORD-19

CORD-19 is a resource of over 500,000 scholarly articles, including over 200,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease.

2. NCBI-Disease

The NCBI disease corpus is fully annotated at the mention and concept level to serve as a research resource for the biomedical natural language processing community.

Corpus characteristics:

- 793 PubMed abstracts
- 6,892 disease mentions
- 790 unique disease concepts
- divided into training, developing and testing sets

3. CHEMDNER

The CHEMDNER corpus is a collection of 10,000 PubMed abstracts that contain a total of 84,355 chemical entity mentions labeled manually by expert chemistry literature curators, following annotation guidelines specifically defined for this task. The abstracts of the CHEMDNER corpus were selected to be representative for all major chemical disciplines.

4. JNLPBA

JNLPBA is a biomedical dataset that comes from the GENIA version 3.02 corpus (Kim et al., 2003). It was created with a controlled search on MEDLINE. From this search 2,000 abstracts were selected and hand annotated according to a small taxonomy of 48 classes based on a chemical classification. 36 terminal classes were used to annotate the GENIA corpus**.**

5. BC5CDR

Created by Li et al. at 2015. BC5CDR corpus consists of 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases and 3116 chemical-disease interactions.

6. CHEMPROT

ChemProt consists of 1,820 PubMed abstracts with chemical-protein interactions annotated by domain experts and was used in the BioCreative VI text mining chemical-protein interactions shared task.

PERFORMANCE MEASURES

Since Named Entity Recognition is a classification problem, where a token is classified as a particular named entity,

classification performance metrics can be used here. Relation Extraction is also an classification problem so the same measures can be used here as well. They are,

1. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

2. Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

3. F1-Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

4. Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

$$\textbf{\textit{Accuracy}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Since, the CORD-19 dataset doesn't possess any ground truth information. The relations extracted can only be verified via

fact-checking known websites like The Comparative Toxicogenomics Database | CTD (ctdbase.org).

Also the system can be evaluated by fact-checking the final found COVID-19 related entities by manually checking the relation between them.

REFERENCES

1. Ayoub Harnoune, Maryem Rhanoui, Mounia Mikram, Siham Yousfi, Zineb Elkaimbillah, Bouchra El Asri, BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis, Computer Methods and Programs in Biomedicine Update, Volume 1, 2021, 100042,ISSN2666-9900, https://doi.org/10.1016/j.cmpbup.2021.100042.

2. Minsoo Cho, Jihwan Ha, Chihyun Park, Sanghyun Park, Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition, Journal of Biomedical Informatics, Volume 103, 2020, 103381, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103381.

3. Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Frontiers in cell and developmental biology, 8, 673. https://doi.org/10.3389/fcell.2020.00673.

4. Zheng, S., Rao, J., Song, Y., Zhang, J., Xiao, X., Fang, E., Yang, Y. and Niu, Z., 2020. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in Bioinformatics*,.

5. Lu Wang L, Lo K, Chandrasekhar Y, et al. CORD-19: The Covid-19 Open Research Dataset. Preprint. ArXiv. 2020;arXiv:2004.10706v2. Published 2020 Apr 22.

6. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform. 2014 Feb;47:1-10. doi: 10.1016/j.jbi.2013.12.006. Epub 2014 Jan 3. PMID: 24393765; PMCID: PMC3951655.

7. Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, Leaman R, Lu Y, Ji D, Lowe DM, Sayle RA, Batista-Navarro RT, Rak R, Huber T, Rocktäschel T, Matos S, Campos D, Tang B, Xu H, Munkhdalai T, Ryu KH, Ramanan SV, Nathan S, Žitnik S, Bajec M,

Weber L, Irmer M, Akhondi SA, Kors JA, Xu S, An X, Sikdar UK, Ekbal A, Yoshioka M, Dieb TM, Choi M, Verspoor K, Khabsa M, Giles CL, Liu H, Ravikumar KE, Lamurias A, Couto FM, Dai HJ, Tsai RT, Ata C, Can T, Usié A, Alves R, Segura-Bedmar I, Martínez P, Oyarzabal J, Valencia A. The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform. 2015 Jan 19;7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S2. doi: 10.1186/1758-2946-7-S1-S2. PMID: 25810773; PMCID: PMC4331692.

8. Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at jnlpba. In NLPBA/BioNLP.

9. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. ArXiv, abs/1508.01991.

10. Ma, X., & Hovy, E.H. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. ArXiv, abs/1603.01354.

11. Su, P., Peng, Y., & Vijay-Shanker, K. (2021). Improving BERT Model Using Contrastive Learning for Biomedical Relation Extraction. BIONLP.

12. Daniel Domingo-Fernandez, Shounak Baksi, Bruce´ Schultz, Yojana Gadiya, Reagon Karki, Tamara Raschka, Christian Ebeling, Martin HofmannApitius, and Alpha Tom Kodamullil. 2020. Covid19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of covid-19 pathophysiology. bioRxiv

13. Peter Richardson, Ivan Griffin, C. Tucker, D. Smith, Olly Oechsle, Anne Phelan, Michael Rawling, Edward Savory, and J. Stebbing. 2020. Baricitinib as potential treatment for 2019-ncov acute respiratory disease. Lancet (London, England), 395:e30 – e31.

14. Kim, T.; Yun, Y.; Kim, N. Deep Learning-Based Knowledge Graph Generation for COVID-19. Sustainability 2021, 13, 2276. https://doi.org/10.3390/su13042276

15. Repke T., Krestel R. (2021) Extraction and Representation of Financial Entities from Text. In: Consoli S., Reforgiato Recupero D., Saisana M. (eds) Data Science for Economics and Finance. Springer, Cham. https://doi.org/10.1007/978-3-030-66891-4_11