DD-RDL: Drug-Disease Relation Discovery and Labeling

Jovana Dobreva, Milos Jovanovik$^{(\boxtimes)}$, and Dimitar Trajanov

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
in Skopje, Skopje, North Macedonia
{jovana.dobreva,milos.jovanovik,dimitar.trajanov}@finki.ukim.mk

Abstract. Drug repurposing, which is concerned with the study of the effectiveness of existing drugs on new diseases, has been growing in importance in the last few years. One of the core methodologies for drug repurposing is text-mining, where novel biological entity relationships are extracted from existing biomedical literature and publications, whose number skyrocketed in the last couple of years. This paper proposes an NLP approach for drug-disease relation discovery and labeling (DD-RDL), which employs a series of steps to analyze a corpus of abstracts of scientific biomedical research papers. The proposed ML pipeline restructures the free text from a set of words into drug-disease pairs using state-of-the-art text mining methodologies and natural language processing tools. The model's output is a set of extracted triplets in the form (drug, verb, disease), where each triple describes a relationship between a drug and a disease detected in the corpus. We evaluate the model based on a gold standard dataset for drug-disease relationships, and we demonstrate that it is possible to achieve similar results without requiring a large amount of annotated biological data or predefined semantic rules. Additionally, as an experimental case, we analyze the research papers published as part of the COVID-19 Open Research Dataset (CORD-19) to extract and identify relations between drugs and diseases related to the ongoing pandemic.

Keywords: Drug-disease relations · NLP · Knowledge extraction

1 Introduction

Drug discovery is a time-consuming, expensive, and high-risk process and it usually takes 10–15 years to develop a new drug, with the success rate of developing a new molecular entity being only around 2% [31]. Because of that, studying drugs that are already approved to treat one disease to see if they are effective for treating other diseases, known as drug repurposing, has been growing in importance in the last few years [27]. The core methodologies of drug repurposing approaches can be divided into three categories: network-based approaches, text-mining approaches, and semantic approaches [31]. The text-mining approach's main goal is the processing of medical and biological literature and extracting

L. Antovski and G. Armenski (Eds.): ICT Innovations 2021, CCIS 1521, pp. 98–112, 2022.
https://doi.org/10.1007/978-3-031-04206-5_8

fruitful novel biological entity relationships. The importance of the text-mining approach has been increasing in the last couple of years due to the fact that the amount of scientific publications related to the biomedical and life sciences domain is growing at an exponential rate, with more than 32 million medical publications available on PubMed. On the other hand, the latest advancements in natural language processing (NLP) have significantly improved the efficiency of language modeling [4]. Text mining in the medical and biological domain can be divided into four phases [31]: information retrieval (IR), biological name entity recognition (BNER), biological information extraction (BIE), and biological knowledge discovery (BKD). In the IR step, the relevant documents are selected from the identified source. In the BNER step, the biological entities are identified in each of the retrieved documents. This step is accomplished with the use of the NLP subtask for named entity recognition (NER) [16]. NER includes the automated implementation of natural language processing tasks like tokenization, handling punctuation, stop terms and lemmatization, as well as the ability to create custom (case-specific) text processing functions like joining consecutive tokens to tag a multi-token object or performing text similarity computations. In other words, we can tell that NER uses the grammatical sentence structure to highlight the given entities.

The task of dealing with the syntactic rules in the sentences is a part of co-reference resolution [24]. Co-reference resolution, i.e. the replacement of all expressions referring to the same entity in the text, is an important step for many higher-level NLP tasks which involve understanding the natural language, such as summarizing documents, answering questions, and extracting information [25]. On the other hand, applying semantic role labeling (SRL) on the processed texts leads to observing the semantic part of the sentences and extracts the represented roles from it [29]. The BIE and BKD steps include the tasks of knowledge extraction from the biological corpus, such as drug-disease pairs and drug-action-disease triples. Based on our previous research experience in the field of applying NLP techniques to the biomedical domain [3,10], we recognized a need for a new model that will be able to learn new information about potential therapies for given diseases, based on published medical research. In this paper, we present a literature-based discovery (LBD) model which employs a series of steps to analyze the corpus of abstracts of scientific biomedical research papers. For that purpose, the model restructures the free text from a set of words into drug-disease pairs using the text mining techniques and the NLP tools. With this, it is able to build a network of known diseases and drugs which are correlated. The graph representation used by the model represents an abbreviated form of the knowledge that covers all drugs and diseases which are covered throughout the entire corpus of abstracts. In it, each node represents a disease or a drug entity that is interconnected with other nodes via a weighted edge, which denotes the verb that occurs most frequently with the given pair. The inclusion of verbs in the process allows us to add an additional label to the relation, which renders the knowledge extraction more precise. We used datasets which are publicly available and represent publications from the medical domain. Therefore, the corpus was very large, but the data was not labeled or annotated. In order to use the raw data and get comparable results we used NLP state-of-the-art models, such as AllenNLP, BERT, RoBERTa, GTP, etc. The evaluation of our model is based on the gold standard dataset for drug-disease cases that contains 360 drug-disease relationships which are selected from the therapeutic target database [22]. As an experimental case, we also analyze the COVID-19 by analyzed the abstracts of the research papers published as part of the COVID-19 Open Research Dataset (CORD-19) [26].

2 Related Work

Applying text mining and NLP methods to discover drug-disease pairs has been an exciting research field in recent years. In [27] the authors propose a system that extracts the drug-disease pairs by using the vocabulary of drugs and diseases from the Unified Medical Language System (UMLS). Due to the large drug and disease dictionary size, an optimized variant of the string searching algorithm Aho-Corasick [2] is used. The algorithm is a dictionary-matching algorithm that locates elements of a finite set of strings (dictionary) within an input text and matches all strings simultaneously. The learned patterns are then used to extract additional treatment and inducement pairs. A similar approach is used in [30], where the extracted patterns are detected from the sentences that have the following format "DRUG Pattern DISEASE" or "DISEASE Pattern DRUG". For example, the top five patterns are: "in", "in the treatment of", "for", "in patients with", "on", and we can note that some of them do not express the real nature of the relationship. Based on literature mining, the creators of [12] built a knowledge base that is used to find biological entities linked to the COVID-19 disease. They collect disease-drug interactions from the CORD-19 literature and categorize them as positive or negative (labels). The positive label indicates that the medicine is moderately effective in curing the condition, whereas the negative label indicates that the drug is ineffective in curing the ailment. The proposed platform does not find more details regarding the relationship between drugs and diseases and how they are connected. The authors of [17] expand the pattern-based technique beyond single drugs and diseases to drugs combinations. They look for trends in the relationship between drug combinations and diseases. They use a pattern-based technique to extract illness and medication combination pairings from MEDLINE abstracts and build a word meaning disambiguation system based on POS tagging. Using many language characteristics, including lexical and dependency information, the authors of [8] proposed a supervised learning technique for automatically extracting chemical-induced disease connections. Furthermore, the suggested technique makes use of the MeSH restricted vocabulary to aid in the training of classifiers and to solve the issue of relation redundancy during the extraction process. Using gold-standard entity annotations, the system obtained F-scores of 58.3% on the test dataset.

The authors of [19] investigate the many sorts of relationships that exist in LBD systems. They utilize the basic A-B-C model [9], where six types of relations are covered in this overview: c-doc, c-sent, c-title, SemRep, ReVerb, and the

Stanford parser. C-doc looks for term co-occurrence over the whole document (in this case, a document is an abstract). C-sent takes a more stringent approach, considering phrases to be co-occurring if they appear in the same sentence inside a document (abstract) and c-title just examines document titles. SemRep [13] is a free accessible tool that uses underspecified syntactic processing and UMLS domain knowledge to extract subject-relation-object triples (such as "X treats Y") from biomedical literature. Based on enforced syntactic and lexical constraints, the freely accessible ReVerb information extraction technology extracts binary relations conveyed by verbs. The Stanford parser extracts phrase structure trees and creates typed grammatical connections, such as subject, between pairs of words. For the purpose of assessment, the authors create a gold standard dataset. With time slicing, this is possible: hidden knowledge is created from all data up to a given cut-off date and compared against fresh ideas provided in publications after the cut-off date. Identifying innovative ideas in publications after the cut-off date, on the other hand, is not straightforward: extracting all newly co-occurring pairs of CUIs, for example, will obviously provide a huge and noisy gold standard, favoring LBD number over quality.

SemRep [13] is an NLP system that uses linguistic principles and UMLS domain knowledge to extract semantic relations from PubMed abstracts. The assessment is based on two different datasets. They employ a manually annotated test collection and undertake thorough error analysis in one study[1]. SemRep's performance on the CDR dataset, a typical benchmark corpus annotated with causal chemical-disease correlations, is also evaluated. On a manually annotated dataset, a rigorous assessment of SemRep provides 0.55 precision, 0.34 recall, and 0.42 F_1 score. A more lenient evaluation provides 0.69 accuracy, 0.42 recall, and 0.52 F_1 score, which more properly represents SemRep performance.

SemaTyP [22] is a biomedical knowledge graph-based drug discovery approach that mines published biomedical literature to find potential medicines for illnesses. They first use SemRep to create a biomedical knowledge graph from the relationships extracted from biomedical abstracts, then train a logistic regression model by learning the semantic types of paths of known drug therapies that exist in the biomedical knowledge graph, and finally use the learned model to discover drug therapies for new diseases. They provide a gold standard[2] of drug-disease instances for the assessment from TTD[3]. They chose TTD's 360 drug-disease correlations as the gold standard for drug rediscovery testing. NEDD [33] is a computational technique based on meta-paths. Using Hin2Vec [5] to generate the embeddings, they first create a heterogeneous network as an undirected graph by combining drug-drug similarity, disease-disease similarity, and known drug-disease correlations. The low dimensional representation vectors of medications and diseases are generated by NEDD using meta pathways of various lengths to explicitly reflect the indirect links, or high order closeness, inside drugs and diseases. Experiments using a gold standard dataset [7], which contains 1,933

[1] https://semrep.nlm.nih.gov/GoldStandard.html.

[2] https://doi.org/10.6084/m9.figshare.6389870.v1.

[3] http://db.idrblab.net/ttd/.

verified drug-disease correlations, demonstrate that NEDD outperforms state-of-the-art methods in terms of prediction outcomes. The authors offer a technique in [21] that differentiates seven connections between two semantic concepts, "therapy" and "disease". Five graphical models and a neural network are represented. Only three relations were found, namely, "cure", "prevent", and "side effect", which were represented with accuracy levels of 92.6, 38.5, and 20, respectively. The papers [1,11,21,32] represent different classification models for relationships between drugs and diseases. In each of them the main idea is to first apply a named entity recognition for the medical terms and then to use approaches based on linguistic patterns and domain knowledge. For the classification part, the authors choose diverse types of ML models, such as neural networks, SVM, logistic regression, etc.

The proposed methods for drug-disease pairs discovery are mainly lexicon-based and do not enable discovering the actions (the verbs) connecting the disease and the drug. Starting from this drawbacks we propose a model that, besides discovering the drug-disease pairs, also allows for labeling the relationship with the verb that connects the detected entities. The models that are used in the proposed ML pipeline are based on publicly available state-of-the-art NLP models that are mainly trained on standard English text. With the obtained results, we have shown that we can get comparable results without using large amounts of annotated biomedical data or specially coded grammatical and semantic rules.

3 Drug-Disease Relation Discovery and Labeling (DD-RDL)

The main goal is to build a text mining tool for biomedical abstracts which will provide useful knowledge extraction with the help of general NLP techniques. Our model, unlike other models, does not use CUIs or UMLS, and is solely based on entity extraction and building a co-occurrence matrices. Therefore, our experiment is based on trying various NLP tools and techniques which will lead to a good LBD system, so the focus is on exploring and testing if the newest text mining tools will prove useful in the biomedical field. On the other hand, building and training a whole new domain-specific model especially when we are building a graph, takes a lot of time. Therefore, our model uses domain specific and globally used NLP tools for faster knowledge extraction, where a big data corpus is not necessary for achieving good accuracy. The state-of-the-art models for this problem are building NLP tools that are domain specific. Our idea is to try already existing NLP models for English language into the medical field, we know that the words and meanings are different into the spoken English and medically based one. In other word, the training of each NLP task from the DD-RDL model is on English language, but we are transferring that knowledge into domain specific one in order to see is the relation extraction just as successful as in the language tasks. The Drug-Disease Relation Discovery and Labeling (DD-RDL) model is composed of six stages grouped in two subsystems (Fig. 1). The first subsystem is composed of a pipeline that has three steps, starting with named entities recognition (NER), continuing with co-reference resolution (CR), and ending with semantic role labeling (SRL). The second subsystem is composed of knowledge extraction tasks, and its pipeline is comprised of three stages: verb grouping, adjacency matrix construction, and knowledge extraction. The code is publicly available on GitLab[4].

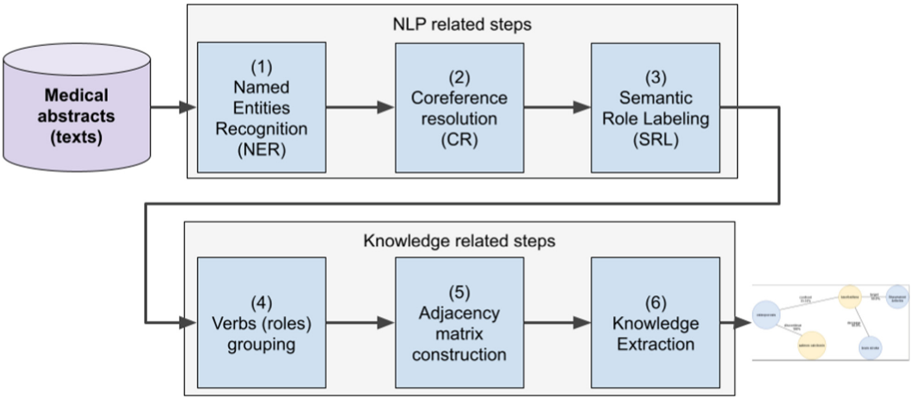
Medical abstracts (texts)
NLP related steps
(1) Named Entities Recognition (NER)
(2) Coreference resolution (CR)
(3) Semantic Role Labeling (SRL)
Knowledge related steps
(4) Verbs (roles) grouping
(5) Adjacency matrix construction
(6) Knowledge Extraction

Fig. 1. Structure of the DD-RDL Model.

As an input, the DD-RDL model gets a dataset of medical or life science-related abstracts of papers, or full texts. As a result, the model provides a set with triplets in the form (drug, verb, disease) with its relative frequency as an additional information. The model also generates a knowledge graph where nodes are disease and drugs, connected with links that represent the connecting verbs (roles) and the associated relative frequency. The entire process is available through a module that can be installed and started from a command-line interface, but we also provide a web service where the client can attach multiple scientific papers and process them with the DD-RDL model.

3.1 Named Entity Recognition (NER)

The first step is finding entities from the medical domain, such as disease, syndrome, pharmaceutical substance, amino acid, protein, therapy, etc. For this purpose, we use the MedCAT library [14] that is designed to be used for extracting information from electronic health records (EHRs) and link them to biomedical ontologies, such as UMLS. MedCAT offers a pre-trained model that automatically builds a pallet of various medical entities, such as pharmaceutical substances, diseases, treatments, amino acid, peptides, proteins, etc. The first step is very important, because the globally used NER tools extract a different pallet of words, since these domain-specific terms do not occur in their bag-of-words.

[4] https://gitlab.com/jovana.dobreva16/dd-rdl-model.git.

Therefore, we used MedCAT which is able to highlight the biomedical domain entities and extract them based on their category. In order to improve the drug[5] and disease[6] detection, we add an additional step that uses a simple lookup-based algorithm, tables are available on gitlab[7], that is trying to find additional drugs that were not identified by the MedCAT library.

3.2 Co-reference Resolution (CR)

The co-reference resolution (CR) step includes finding all linguistic expressions (mentions) in the text that refer to the related entity. After finding those mentions, they are resolved by replacing them with the associated entities. After this step, the pronouns ("that", "she", "his", etc.) are replaced with the appropriate entity making the sentences simpler and more appropriate for the next step of semantic role labeling. There are a couple of libraries like StanfordNLP [20], AllenNLP [6], and Neural-Coref [15,28] that provide an implementation of co-reference resolution algorithms. Due to the simplicity to use and the high accuracy, we use Neural-Coref. Even though these co-reference models are not the best possible solution for biomedical texts, given that they show lower performance results than the ones which are domain-specific, we want to have a model which is able to extract knowledge with the help of grammatical and semantic knowledge from general texts.

3.3 Semantic Role Labeling (SRL)

The goal of SRL is to identify the events in the sentence, such as discovering "who" did "what" to "whom", "where", "when", and "how". The predicates (typically verbs) are the central part of the SRL process, and they define "what" took place. The other sentence constituents express the participants in the event (such as "who" and "where"), as well as other event properties (such as "when" and "how"). SRL's major goal is to properly represent the semantic relationships that exist between a predicate and its related participants and attributes. These connections are chosen from a set of probable semantic roles for that predicate that has been pre-defined (or class of predicates) [18]. The SRL process is done by using a BERT-based algorithm for semantic role labeling [23]. When the input text passes through this step, the subject, verb, and object for each sentence is recorded. The subject and the object are then filtered to keep only those connected with the identified entities in the previous step. As a result of this step, we get a set of triples (subject, verb, object) associated with each input text.

[5] https://www.kaggle.com/arpikr/uci-drug.

[6] https://www.kaggle.com/priya1207/diseases-dataset.

[7] https://gitlab.com/jovana.dobreva16/dd-rdl-model/-/tree/master/data_storage.

3.4 Verbs Grouping

After processing the subject and object entities, the next step in the pipeline is verbs grouping. This step aims to lower the number of different verbs by replacing similar verbs with a single one that has the same meaning. First, all the verbs are converted to present tense, and then they are encoded using word2vec vectors from the NLTK library. A similarity matrix between all verbs is created, and the most similar verbs are grouped together. We use cosine similarity with a threshold to create the groups. When the groups are made, the most frequent verb is chosen as a group representative and is used to replace all other occurrences of the other verbs from the group. This step gives us a smaller range of verb-actions between the diseases and drugs, leading to decreased bias and variance.

3.5 Adjacency Matrix Construction

After the triplets are created, we use them to construct the adjacency matrix that represents the probability of the occurrence of the pair (verb, disease) with a given drug:

$$P(Verb, Disease|Drug)$$

We then prune the matrix in order to decrease the number of zero values in it. The triples that remain are the nodes and edges of the knowledge graph. The weight of each edge is a calculated relative frequency of the corresponding triple.

In order to generate the adjacency matrix, we first created two matrices where, the first one represents the probability of a given disease occurring with a known verb-drug pair:

$$P(Disease|Verb, Drug)$$

and the second one represents the probability of a given verb occurring together with a given drug:

$$P(Verb|Drug)$$

By inverting the first matrix and then multiplying it with the second matrix, we arrive to the above mentioned final matrix.

3.6 Knowledge Extraction

As a final step in the processing pipeline execution, from the adjacency matrix, we constructed a bipartite knowledge graph in which the nodes are drugs and diseases, and the links between them represent the identified action (verb). To each link, a relative frequency is also associated. An example of a knowledge graph constructed by the DD-RDL model is shown in Fig. 2.

As we showed in the previous steps, we added a dose of domain-specific knowledge into models that are globally used for processing English texts. We combined domain specific NER and SRL together with co-reference resolution based on the English grammar. In the final steps, we used a statistical and mathematical knowledge for building the matrices.

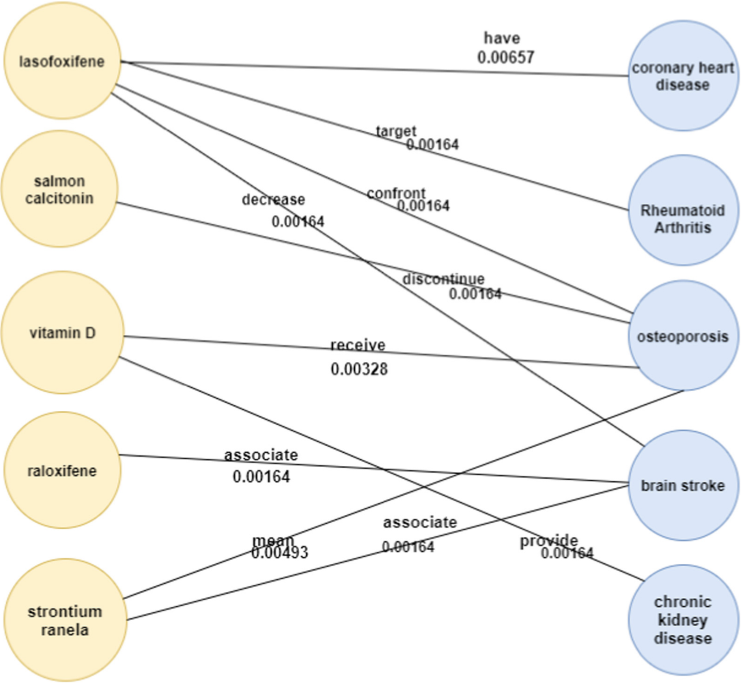

Fig. 2. An example knowledge graph of the knowledge extracted by the DD-RDL model.

4 Datasets

In order to compare our approach and its performance with existing systems, we used one of the gold standard datasets for drug-disease cases [22]. The dataset contains a pallet of different diseases out of which we selected two: osteoporosis and cardiovascular disease. The main reason of selecting just two diseases out of the corpus was because we wanted to test our model on a specific disease corpus, and see if the gold standard pairs were occurring in the model output, possibly along with new pairs. The gold standard dataset contains pairs of diseases and drugs that correlate based on previous medical research. Therefore, we needed to create a dataset that will represent a collection of abstracts where the given disease and drug occur. We collected a set of 203 abstracts from PubMed, which are related to the two selected diseases and the appropriate drugs paired with them from the drug-disease dataset. The selected papers refer to 27 different drug-disease pairs. The gold standard dataset [22] does not contain verb-actions for the drug-disease pairs, so we calculate the model's accuracy based only on the drug-disease coverage. In order to further test our model, we applied it on data from the COVID-19 Open Research Dataset (CORD-19) [26]. It represents a collection of scientific papers on COVID-19 and related historical coronavirus research. Even though it is composed of several features, we are simply interested in the title and the abstract of the papers. As we know, therapies for the SARS-CoV-2 virus are currently not defined. Although the result obtained using this dataset cannot be explicitly tested, we performed a manual check of the results.

5 Results

As we explained above, we use a set of 203 research paper abstracts from PubMed which contain 27 drug-disease pairs selected from the gold standard dataset for drug-disease cases [22]. The diseases in these pairs are osteoporosis and cardiovascular disease.

To show how the model extracts knowledge from the abstract, we will take a look at two sentences from two different abstracts, which are related to aspirin and cardiovascular disease. The sentences have the following representation:

- S1: "In 2016, the US Preventive Services Task Force recommended initiating aspirin for the primary prevention of both cardiovascular disease and colorectal cancer among adults ages 50 to 59 who are at increased risk for cardiovascular disease."
- S2: "Background: A polypill comprising statins, multiple blood-pressure-lowering drugs, and aspirin has been proposed to reduce the risk of cardiovascular disease."

From these two sentences the following SRL outputs were generated:

- S1: 'subject': 'the US Preventive Services Task Force', 'verb': 'recommend', 'object': 'initiating aspirin for the primary prevention of both cardiovascular disease and colorectal cancer among adults ages 50 to 59 who are at increased risk for cardiovascular disease'
- S2: 'subject': 'A polypill comprising statins, multiple blood-pressure-lowering drugs, and aspirin', 'verb': 'reduce', 'object': 'the risk of cardiovascular disease'.

The discovered (drug, verb, disease) triplet for the first sentence is (aspirin, recommend, cardiovascular disease) and for the second one is (aspirin, reduce, cardiovascular disease).

After running the abstracts through the pipeline of our model, the model extracted all pairs correctly, giving 100% coverage. Additionally, it discovered 18 new disease-drug pairs, which are correctly correlated. Given that the test dataset does not contain the action (verb), we calculated the accuracy based on the drug-disease pairs only, not the full triples. Some of the predicted pairs are shown in the Table 1, where we include the action verb to describe the drug-disease relation, and we show the calculated relative frequency. The relative frequency is calculated as the number of occurrences of the specified (drug, action, disease) triplet in the entire analyzed text corpus, divided by the total number of unique triplets extracted from the same corpus.

We performed the evaluation by comparing the (drug, disease) pairs we generated with the ones present in the gold standard dataset. We measured recall,

Table 1. A set of drugs, their actions on diseases and the relative frequency, extracted from the PubMed abstracts. We analyzed 203 paper abstracts from PubMed based on existing 27 drug-disease pairs, and managed to detect all of them with our model. This table presents 15 of them.

Drug	Action	Disease	Relative frequency
Calcium	Receive	Osteoporosis	0.00493
Vitamin E	Decrease	Cardiovascular Disease	0.00328
Strontium Ranelate	Present	Osteoporosis	0.00328
Ipriflavone	Prevent	Osteoporosis	0.00328
PTH-CBD	Prevent	Osteoporosis	0.00164
rhPTH	Evaluate	Osteoporosis	0.00164
Aspirin	Outweigh	Cardiovascular Disease	0.00164
Digitoxin	Block	Cardiovascular Disease	0.00164
Drug Combination	Underscore	Cardiovascular Disease	0.00164
Raloxifene	Decrease	Cardiovascular Disease	0.00164
Delapril	Ameliorate	Cardiovascular Disease	0.00164
Acarbose	Delay	Cardiovascular Disease	0.00164
Salmon Calcitonin	Discontinue	Osteoporosis	0.00164
Raloxifene	Prevent	Osteoporosis	0.00164
Lasofoxifene	Prevent	Osteoporosis	0.00164

which showed 100% accuracy. Other metrics, such as precision and F_1 score, cannot be utilized in our case, because we discover novel relations between drugs and diseases, relations which are not occurring in the gold standard dataset, and are not in the group of false positives.

Our model extracts 18 new drug-disease relations from the 203 PubMed abstracts. They are presented in Table 2, along with the extracted action, and the relative frequency of the triplet in the context of the entire analyzed corpus. The novel drug-disease pairs, shown in Table 2, are not part of the dataset we used. Because of that, we needed to manually check each pair in order to prove its correctness. We confirmed the model's predictions to be correct after searching each extracted pair in relevant biomedical literature. For instance, our model extracted a relation between vitamin D and osteoporosis (Table 2), and we can check and confirm that this pair has a proven relation, based on published papers available on the PubMed website[8].

6 Applying the DD-RDL Model on COVID-19 Related Papers

Given the global impact of the ongoing pandemic and the urgency to find solutions which will help mitigate the negative effects of COVID-19 on the public

[8] https://pubmed.ncbi.nlm.nih.gov/?term=%28%28Vitamin+D%29+AND+%28Osteoporosis%29%29+AND+%28receive%29.

Table 2. The novel drug-disease pairs and their actions, extracted from the PubMed abstracts. The analysis of the 203 PubMed paper abstracts resulted in the extraction of 18 novel drug-disease pairs, which we manually checked and confirmed to be correct in relevant biomedical literature.

Drug	Action	Disease	Relative frequency
Magnesium	Have	Coronary Heart Disease	0.00657
Antidiabetic	Represent	Cardiovascular Risk Factors	0.00493
Vitamin D	Receive	Osteoporosis	0.00328
Vitamin C	Suggest	Cardiovascular Disease	0.00328
Estrogen	Limit	Osteoporosis	0.00164
ACE inhibitors	Ameliorate	Diabetes Complications	0.00164
ACE inhibitors	Suggest	Cardiovascular Disease	0.00164
Atorvastatin	Contain	Cardiovascular Risk Factors	0.00164
Therapeutic Agents	Affect	Osteoporosis	0.00164
Vasoconstrictor	Stimulate	Hypertension	0.00164
Testosterone	Review	Hypogonadism	0.00164
Fluoride	Cause	Dental Fluorosis	0.00164
Vitamin D	Provide	Chronic Kidney Disease	0.00164
Calcium	Indicate	Atherosclerotic Disease	0.00164
Glucosidic Isoflavone	Prevent	Osteoporosis	0.00164
Simvastatin	Receive	Cardiovascular Disease	0.00164
Metformin	Reduce	Type 2 Diabetes	0.00164
Antidiabetic	Represent	Type 2 Diabetes	0.00164

health, we decided to apply our DD-RDL model on a set of COVID-19 related reseach papers. For that purpose, we used the CORD-19 dataset [26]. According to the results produced by our model (Table 3), we can conclude that selenium, zinc and magnesium as supplements have an effect on the maintenance of the immune system. Therefore, these supplements can be considered as positively correlated with the protection against the SARS-Cov-2 virus. On the other hand, a number of vaccines and conjugated vaccines have been developed or are in the process of development, so it is no wonder that they are very often mentioned in the abstracts. There are also protease inhibitors that, according to research, prevent the spread of the SARS-Cov-2 virus. The literature also mentions cyclohexapeptide diantin G, which occurs in the treatment of tumors. The results also indicate the appearance of copper, as new research in this field focuses on wearing masks with copper as a material that protects against the virus. We also detected that papers refer to corticosteroids as substances which can prevent the side-effects of COVID-19. Finally, the appearance of dissolved oxygen is part of every respirator, which is of great help to patients in whom this infectious disease is in its final stage. This description of each effect from the extracted drug-disease relations based on the CORD-19 dataset gives almost the same information as the daily news articles which cover the ongoing pandemic.

Table 3. The drug-action-disease triplets extracted from the CORD-19 dataset. The analysis of the abstracts resulted in the extraction of 11 triplets, which we manually checked and confirmed to be correct in relevant biomedical literature.

Drug	Action	Disease	Relative frequency
PCV13	Forestall	Community-Acquired	0.02362
Vaccine	Forestall	Influenza-like illness	0.01181
Protease Inhibitor	Depend	Irritable Bowel Syndrome	0.00393
Magnesium	Contain	Diphtheria	0.00393
Dianthin G	Block	Hypoglycemia	0.00393
Corticosteroids	Determine	Osteonecrosis of the jaw	0.00196
Dissolved Oxygen	Require	Respiratory	0.00196
Conjugate Vaccine	Present	Blood Stream Infection	0.00196
Zinc	Present	Zinc Deficiency	0.00196
Selenium	Consider	Ebola Virus Disease	0.00196
Copper	Exceed	Crohn Disease	0.00196

7 Conclusion and Future Work

This paper presented the DD-RDL model that can extract the diseases, drugs, and relations between them from medical and biological texts. In the proposed model, we apply an additional step based on SRL to discover the verbs connecting drugs and diseases. By discovering the verbs, we can add an additional label to the relation, thus allowing for the extraction of more precise knowledge from the texts. The proposed model is composed of a six-step pipeline, starting with NER, then CR, SRL as standard NLP tasks, which are then followed with the knowledge extraction steps, starting with verbs grouping, adjacency matrix construction, and at the end, the knowledge graph extraction. For all of the NLP tasks, the current state-of-the-art models based on transformers architectures are used. We evaluate the model based on a gold standard dataset for drug-disease relationships, and we demonstrate that it is possible to achieve similar results without the need of a large amount of annotated biological data or predefined semantic rules. The application of NLP in the medical domain has enormous potential. In our future work, we would like to extend the model to recognize the other important associations like disease-gene, drug-protein, and side-effects of Drugs. The second idea for future work is to implement different versions of this model for other business areas, where we will shift the focus to domain-specific knowledge extraction.

Acknowledgement. The work in this paper was partially financed by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje.

References

1. Abacha, A.B., Zweigenbaum, P.: Automatic extraction of semantic relations between medical entities: a rule based approach. J. Biomed. Seman. **2**(5), 1–11 (2011)
2. Aho, A.V., Corasick, M.J.: Efficient string matching: an aid to bibliographic search. Commun. ACM **18**(6), 333–340 (1975)
3. Dobreva, J., Jofche, N., Jovanovik, M., Trajanov, D.: Improving NER performance by applying text summarization on pharmaceutical articles. In: Dimitrova, V., Dimitrovski, I. (eds.) ICT Innovations 2020. CCIS, vol. 1316, pp. 87–97. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62098-1_8
4. Filannino, M., Uzuner, Ö.: Advancing the state of the art in clinical natural language processing through shared tasks. Yearbook Med. Inform. **27**(1), 184 (2018)
5. Fu, T.y., Lee, W.C., Lei, Z.: Hin2vec: explore meta-paths in heterogeneous information networks for representation learning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1797–1806 (2017)
6. Gardner, M., et al.: AllenNLP: a deep semantic natural language processing platform. In: Proceedings of Workshop for NLP Open Source Software (NLP-OSS) (2018)
7. Gottlieb, A., Stein, G.Y., Ruppin, E., Sharan, R.: Predict: a method for inferring novel drug indications with application to personalized medicine. Mole. Syst. Biol. **7**(1), 496 (2011)
8. Gu, J., Qian, L., Zhou, G.: Chemical-induced disease relation extraction with various linguistic features. Database **2016**, 042 (2016)
9. Henry, S., McInnes, B.T.: Literature based discovery: models, methods, and trends. J. Biomed. Inform. **74**, 20–32 (2017).https://doi.org/10.1016/j.jbi.2017.08.011,https://www.sciencedirect.com/science/article/pii/S1532046417301909
10. Jofche, N., Mishev, K., Stojanov, R., Jovanovik, M., Trajanov, D.: PharmKE: Knowledge extraction platform for pharmaceutical texts using transfer learning (2021)
11. Kadir, R.A., Bokharaeian, B.: Overview of biomedical relations extraction using hybrid rulebased approaches. J. Ind. Intell. Inf. **1**(3) (2013)
12. Khan, J.Y., et al.: COVID-19Base: a knowledgebase to explore biomedical entities related to COVID-19. arXiv preprint arXiv:2005.05954 (2020)
13. Kilicoglu, H., Rosemblat, G., Fiszman, M., Shin, D.: Broad-coverage biomedical relation extraction with Semrep. BMC Bioinform. **21**, 1–28 (2020)
14. Kraljevic, Z., et al.: MedCAT - Medical Concept Annotation Tool (2019)
15. Lee, K., He, L., Lewis, M., Zettlemoyer, L.: End-to-end neural co reference resolution. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 188–197. Association for Computational Linguistics, Copenhagen, Denmark, September 2017. https://doi.org/10.18653/v1/D17-1018, https://www.aclweb.org/anthology/D17-1018
16. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. CoRR abs/1812.09449 (2018), http://arxiv.org/abs/1812.09449
17. Liu, J., Abeysinghe, R., Zheng, F., Cui, L.: Pattern-based extraction of disease drug combination knowledge from biomedical literature. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI), pp. 1–7. IEEE (2019)
18. Màrquez, L., Carreras, X., Litkowski, K.C., Stevenson, S.: Semantic role labeling: an introduction to the special issue. Comput. Ling. **34**, 145–159 (2008)

19. Preiss, J., Stevenson, M., Gaizauskas, R.: Exploring relation types for literature-based discovery. J. Am. Med. Inform. Assoc **22**(5), 987–992 (2015). https://doi.org/10.1093/jamia/ocv002
20. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: a python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 101–108 (2020)
21. Rosario, B., Hearst, M.A.: Classifying semantic relations in bioscience texts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004), pp. 430–437 (2004)
22. Sang, S., Yang, Z., Wang, L., Liu, X., Lin, H., Wang, J.: SemaTyP: a knowledge graph based literature mining method for drug discovery. BMC Bioinform. **19**(1), 1–11 (2018)
23. Shi, P., Lin, J.: Simple BERT models for relation extraction and semantic role labeling (2019)
24. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Comput. Linguist. **27**(4), 521–544 (2001). https://doi.org/10.1162/089120101753342653, https://www.aclweb.org/anthology/J01-4004
25. Torfi, A., Shirvani, R.A., Keneshloo, Y., Tavaf, N., Fox, E.A.: Natural Language Processing Advancements By Deep Learning: A Survey (2020)
26. Wang, L.L., et al.: CORD-19: The COVID-19 open research dataset (2020)
27. Wang, P., Hao, T., Yan, J., Jin, L.: Large-scale extraction of drug-disease pairs from the medical literature. J. Assoc. Inf. Sci. Technol. **68**(11), 2649–2661 (2017)
28. Wolf, T., et al.: HuggingFace's transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 EMNLP (Systems Demonstrations), pp. 38–45 (2020)
29. Xia, Q., et al.: Syntax-aware neural semantic role labeling (2019)
30. Xu, R., Wang, Q.: Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinform. **14**(1), 1–11 (2013)
31. Xue, H., Li, J., Xie, H., Wang, Y.: Review of drug repositioning approaches and resources. Int. J. Biol. Sci. Int. J. Biol. Sci. **14**(10), 1232 (2018)
32. Yang, H., Swaminathan, R., Sharma, A., Ketkar, V., Jason, D.: Mining biomedical text towards building a quantitative food-disease-gene network. In: Learning Structure and Schemas from Documents, pp. 205–225. Springer, Cham (2011). https://doi.org/10.1007/978-3-642-22913-8
33. Zhou, R., Lu, Z., Luo, H., Xiang, J., Zeng, M., Li, M.: NEDD: a network embedding based method for predicting drug-disease associations. BMC Bioinform. **21**(13), 1–12 (2020)