Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 165-176 recibido 31-03-2022 revisado 11-05-2022 aceptado 13-05-2022

An Overview of Drugs, Diseases, Genes and Proteins in the CORD-19 Corpus

Una visión general de los Fármacos, Enfermedades, Genes y Proteínas en el corpus CORD-19

Carlos Badenes-Olmedo, Álvaro Alonso, Oscar Corcho
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
{carlos.badenes, oscar.corcho}@upm.es
{alvaro.alonsoc}@alumnos.upm.es

Abstract: Several initiatives have emerged during the COVID-19 pandemic to gather scientific publications related to coronaviruses. Among them, the COVID-19 Open Research Dataset (CORD-19) has proven to be a valuable resource that provides full-text articles from the PubMed Central, bioRxiv and medRxiv repositories. Such a large amount of biomedical literature needs to be properly managed to facilitate and promote its use by health professionals, for example by tagging documents with the biomedical entities that appear on them. We created a biomedical named entity recognizer (NER) that normalizes (NEN) the drugs, diseases, genes and proteins mentioned in texts with the codes of the main standardization systems such as MeSH, ICD-10, ATC, SNOMED, ChEBI, GARD and NCBI. It is based on fine-tuning the BioBERT language model independently for each entity type using domain-specific datasets and an inverse index search to normalize the references. We have used the resultant BioNER+BioNEN system to process the CORD-19 corpus and offer an overview of the drugs, diseases, genes and proteins related to coronaviruses in the last fifty years.
Keywords: ner, normalization, bioentities, document retrieval.

Resumen: Durante la pandemia del COVID-19 han surgido varias iniciativas para recopilar publicaciones científicas relacionadas con el coronavirus. Entre ellos, el conjunto de datos de investigación abierta sobre COVID-19 (CORD-19) ha demostrado ser un recurso valioso que proporciona el texto completo de artículos extraídos de los repositorios PubMed Central, bioRxiv y medRxiv. Una cantidad tan grande de literatura biomédica debe gestionarse adecuadamente para facilitar y promover su uso por parte de los profesionales de la salud, por ejemplo, etiquetando documentos con las entidades biomédicas que aparecen mencionadas. Hemos creado un reconocedor biomédico de entidades nombradas (NER) que normaliza (NEN) los fármacos, enfermedades, genes y proteínas mencionados en textos con los códigos de los principales sistemas de estandarización como MeSH, ICD-10, ATC, SNOMED, ChEBI, GARD y NCBI. Se basa en afinar el modelo de lenguaje BioBERT de forma independiente para cada tipo de entidad utilizando conjuntos de datos específicos de dominio y una búsqueda de índice inverso para normalizar las referencias. Hemos utilizado el sistema BioNER+BioNEN resultante para procesar el corpus CORD-19 y ofrecer una visión general de los fármacos, enfermedades, genes y proteínas relacionados con el coronavirus en los últimos cincuenta años.
Palabras clave: identificación de entidades, normalización, bio-entidades, recuperación de documentos.

1 Introduction

Several initiatives have emerged during the COVID-19 pandemic to gather scientific publications related to coronaviruses. The COVID-19 Data Portal[1], maintained by the EU, or the Humandata[2], focused on COVID-

[1] https://www.covid19dataportal.org
[2] https://data.humdata.org/event/covid-19

ISSN 1135-5948. DOI 10.26342/2022-69-14

19 cases around the world, are some examples. The Allen Institute for Artificial Intelligence created the COVID-19 Open Research Dataset (CORD-19)(Wang et al., 2020). It is a continuously growing corpus with all publicly available COVID-19 and coronavirus-related research (e.g. SARS, MERS, etc.) published during the last fifty years, with a huge increase in the last two years. This dataset provides full-text research papers in PDF and JSON format, which can be used as a source of information to extract knowledge related to the infection and disease. At the time of this study (January 2022), it is composed of 334,572 scientific articles retrieved from PubMed Central, a corpus maintained by the World Health Organization (WHO), bioRxiv and medRxiv pre-prints.

Such a large amount of biomedical literature needs to be properly managed to facilitate and promote its use by health professionals. Natural Language Processing (NLP) facilitates document analysis through the extraction of key information from the underlying texts and turning them into structured knowledge that can be understood by humans (Pyysalo et al., 2007). One of the main NLP tasks is the recognition of relevant entities found in texts, what is commonly known as Named Entity Recognition (NER)(Nadeau and Sekine, 2007). This task enables the exploration of texts guided by key terms and the discovery of relationships between them. The NER task identifies meaningful terms in a domain, called *named entities*, and classifies them into predefined entity classes (Li et al., 2020). In the biomedical domain, these entities are medical concepts such as drugs, diseases, or gene mutations, and the task is more specifically known as BioNER. Entities can also be classified according to existing taxonomies to avoid ambiguity, in a task that is commonly known as Entity Linking or Named Entity Normalization (NEN) and, when applied in the biomedical domain, BioNEN (Campos, Matos, and Oliveira, 2012).

The main objective in BioNEN is to use controlled and curated biomedical vocabularies such as Medical Subject Headings (MeSH) [3] codes or the Anatomical Therapeutic Chemical (ATC)[4] classification system, to reduce ambiguities and to extend the information about the entities. Once the entity recognition and normalization tasks are applied in biomedical literature, a set of normalized concepts can be used by Information Retrieval processes, such as the creation of efficient search algorithms, content classification, or Knowledge-Graph construction among others (Chatterjee et al., 2021). These processes play a key role in subsequent NLP tasks such as Question-Answering, Relation Extraction, Knowledge-base population, or Semantic search (Nadeau and Sekine, 2007).

However, the biomedical language entails some challenges in identifying entities (Zhou et al., 2004): (1) *highly specialized terms* (i.e. most of the terms are exclusive of these kinds of texts, making it difficult to reuse general domain knowledge to identify and classify specific domain concepts), (2) *sharing of nouns* (e.g. "5kb and 17kb viruses" refers to "5kb viruses and 17kb viruses"), and (3) *non-standardized naming convention* (e.g. "$N-acetyl-\beta-D-glucosamine$", "$N-Acetylglucosamine$", and "$C_{18}H_{15}NO_6$" refers to the same concept).

This article describes how we performed BioNER and BioNEN tasks on the CORD-19 corpus, and our analysis of the presence of diseases, drugs, genes and proteins in their texts. Our main contributions are:

- A BioNER+BioNEN system based on independently fine-tuned BioBERT models to identify diseases, drugs and genes/proteins from technical texts.[5]
- A collection of scientific texts tagged with normalized terms and codes of diseases, drugs, and genes/proteins[6]. (Badenes-Olmedo, Alonso, and Corcho, 2022)
- A statistical analysis of the presence of biomedical entities in the January 2022 edition of the CORD-19 corpus.

The paper is structured as follows: Section 2 review the state-of-the-art methods to identify biomedical entities and present our approach. The normalization process that we have followed is described in section 3. Section 4 details how the CORD-19 corpus has been processed, and show and discuss the results. Final remarks and future work are presented in section 5.

[3] https://www.nlm.nih.gov/mesh

[4] https://www.whocc.no

[5] https://github.com/drugs4covid/bio-ner

[6] https://doi.org/10.5281/zenodo.6532473

2 *Biomedical Named Entity Recognition*

NER tasks usually follow the pipeline showed in Fig. 1. The text is firstly pre-processed depending on the requirements of subsequent processes (e.g. word cleaning, stemming, verb tense normalization, etc). Afterwards, a representation of the words which compose a text span is made, what serves as an input to a NER model which performs the classification of these features to assign tags to the words. Sometimes, in order to refine the results, a post-processing step is also required to extend or group the entities. Biomedical-NER (BioNER) specializes the NER classification task for the medical domain and, sometimes, particularizes the techniques used to characterize texts. A BioNER method, depending on the type of technique used to classify terms, can be organized into: *Rule/Dictionary-based* (i.e requires domain knowledge to define patterns of the different sorts of named entities to characterize them), *Machine Learning-based* (i.e. discovers rules through automatic patterns and reduces the need for domain knowledge) and Hybrid approaches (i.e. combines methods to leverage benefits from different approaches) (Li et al., 2020) (Perera, Dehmer, and Emmert-Streib, 2020)(Yadav and Bethard, 2019).

2.1 Our Approach

We have created a hybrid system for the recognition and normalization of biomedical entities based on state-of-the-art methods. Our model specializes the BioBERT (Lee et al., 2020) model pretrained with millions of scientific and biomedical articles, with additional training corpora to extend the BioNER task and to cover the BioNEN task from multiple external standardization databases.

The entity classes considered for our system were the most widely used classes in BioNER modelling and the ones with a higher number of corpora available for a fine-tuning process (see Table 1). It is important to note that these biomedical entities can be mentioned in different ways and this further makes it more difficult to achieve a correct recognition and normalization. Variations can be trivial names (e.g. *water*), technical names (.e.g *lung infection with Mycoplasma pneumoniae* to refer to *Bacterial Pneumonia*), brands (e.g. *Veklury®*), systematic IUPAC names (e.g. *2,5,5-trimethyl-2-hexene*),

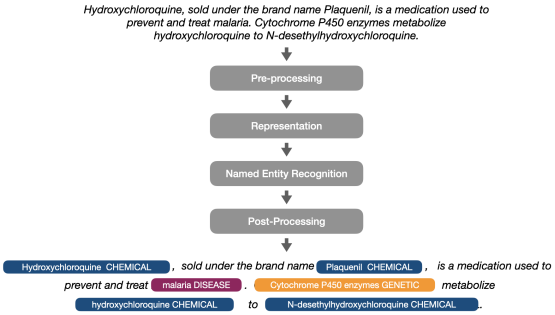

Figure 1: NER pipeline.

generic names (e.g. *Benzenes*), molecular formulas (e.g. CH_3), abbreviated forms (e.g. *DMA* for *dimethylacetamide*) and identifiers of curated databases such as ChEBI[7] (e.g. *145994*).

For each entity class (i.e. disease, drugs and gene/proteins), a different BioNER model was created (see Fig. 2). One model was fine-tuned to recognize disease entities, another for chemical (i.e. drugs) entities and another for genes/proteins. We adopted this strategy because it has proven to behave better for fine-tuned tasks than combining several entity classes in the same task in only one model. The more specific the model is, the better results will be usually obtained for a specific task (Gururangan et al., 2020). Separate models capture better patterns within each of the entity classes allowing to maximize its tagging performance, resulting in a system with the better model possible for each of the entities. Our system offers slightly lower performance than BioBERT model because we jointly use several datasets for fine-tuning. The aim is to increase the ability to identify as many entities as possible, even at the cost of penalizing the accuracy of the model, since our pipeline incorporates an additional normalization step where the entities will be filtered out. The post-processing tasks are based on an inverse index search. The architecture of the system is described in Fig. 2 and further details about each of the components are revised along the following sections.

2.2 Datasets

We have added an untrained fully-connected layer on top of a BioBERT model to perform the fine-tuning. At least three fine-tuning processes have been done to cover the

[7] `https://www.ebi.ac.uk/chebi`

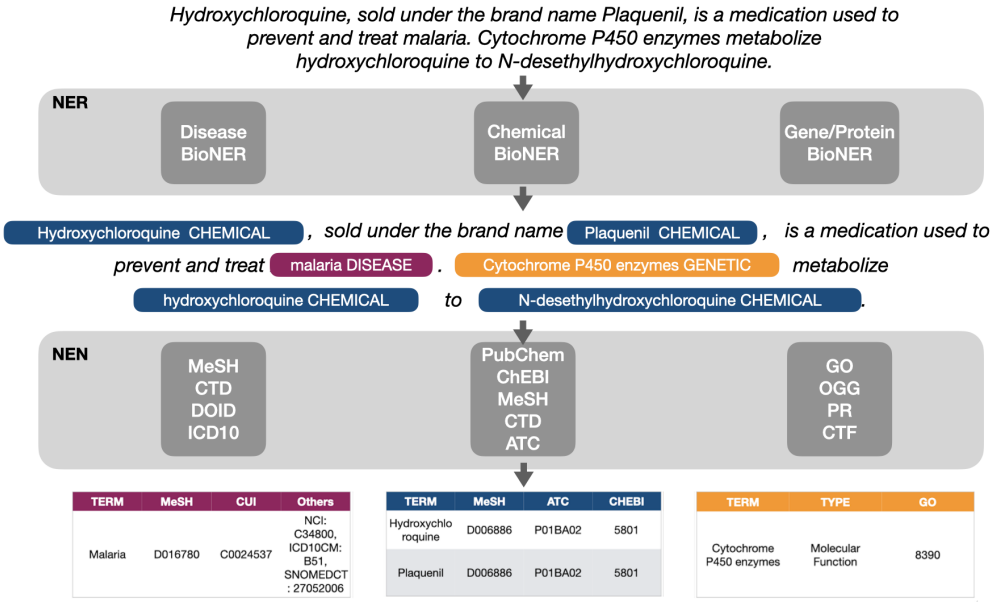

Figure 2: Overview of our BioNER+BioNEN architecture.

three different entities (i.e diseases, drugs and genes/proteins). The corpora used for each training process were selected from existing datasets (see Table 1) according to criteria based on data volume and quality. Furthermore, since the techniques used to identify entities slightly differ between the existing datasets despite being the same target entity, we use two corpora for each entity class to supply the model with a better generalization capacity in situations where a never-seen text is used (Lee et al., 2020).

2.2.1 Diseases Dataset

The BC5CDR-Diseases dataset (Li et al., 2016), with around 13,000 annotations, and the NCBI-Diseases dataset (Doğan, Leaman, and Lu, 2014), with almost 7,000, were the corpora used to recognize diseases. These datasets are the most widely used in BioNER tasks for disease entities and most models provide results for each of them, including BioBERT which obtained an F1-score of 89.71 for NCBI-Diseases and 87.15 for B5CDR-Diseases. Our model, created by combining both datasets during the fine-tuning process, offers a slightly lower performance with a F1-score of 87.4 and 85.8, respectively. This is likely because of the hyperparameter search intensity and because the number of epochs done is lower.

2.2.2 Drugs Dataset

For chemical entities, the two selected datasets were BC4CHEMD (Krallinger et al., 2015) and BC5CDR-Chemicals (Li et al., 2016) with around 80,000 and 15,000 entities respectively. The largest annotated corpus, BioSemantics (Akhondi et al., 2014), was not considered since it is based on patents which could slightly differ from biomedical articles that are the kind of texts in which our system is focused on. The selected datasets were also the most widely adopted corpora for NER tasks in chemical entities and most models provide performance results for them. BioBERT obtained state-of-the-art results in BC4CHEMD with an F1-score of 92.36 and the second best result for BC5CDR with 93.47, which is almost the same than the state-of-the-art result obtained by BlueBERT (Peng, Yan, and Lu, 2019) which was 93.5 . Our system obtained F1 results of 91.7 for BC4CHEMD and 92.99 for BC5CDR-Chemicals.

2.2.3 Gene and Proteins Dataset

Gene and protein entities were jointly considered since they belong to similar semantic types. This consideration is widely adopted in most existent corpus, which consider them together (Goyal, Gupta, and Kumar, 2018). The pair of selected datasets were JNLPBA (Kim et al., 2004) and BC2GM (Smith et al., 2008), which offer around 35,000 and 25,000 annotations respectively. The CRAFT corpus (Bada et al., 2012), which is the largest Gene/Protein NER corpus, was discarded since most models report results

Year	Reference	Corpus Name	Entities	# Annotations	# Tokens
2004	(Kim et al., 2004)	JNLPBA	Genes/Proteins	35460	597333
			Cell Lines	4332	
2008	(Smith et al., 2008)	BC2GM	Genes/Proteins	24583	508257
2012	(Bada et al., 2012)	CRAFT	Chemicals	8137	560000
			Genes/Proteins	49961	
			Species	7449	
			Cell Lines	5760	
2013	(Pafilis et al., 2013)	Species-800	Species	3646	195197
2013	(Ohta et al., 2013)	BioNLP13CG	Species	21683	129878
			Anatomy		
			Genes/Proteins		
2013	(Segura Bedmar, Martínez, and Herrero Zazo, 2013)	SemEval2013 - DrugBank	Chemicals	15745	≈ 65000
2013	(Segura Bedmar, Martínez, and Herrero Zazo, 2013)	SemEval2013 - Medline	Chemicals	2746	≈ 20000
2013	(Pyysalo et al., 2015)	BioNLP13PC	Genes/Proteins	15901	108356
			Chemicals		
2014	(Bagewadi et al., 2014)	mi-RNA	Genes/Proteins	1006	65998
			Species	726	
			Diseases	2123	
2014	(Akhondi et al., 2014)	BioSemantics	Chemicals	386110	5690518
2014	(Pyysalo and Ananiadou, 2014)	AnatEM	Anatomy	13000	250000
2014	(Doğan, Leaman, and Lu, 2014)	NCBI Disease	Diseases	6881	174487
2015	(Goldberg et al., 2015)	LocText	Species	276	22550
2015	(Krallinger et al., 2015)	BC4CHEMD	Chemicals	79842	2235435
2016	(Li et al., 2016)	BC5CDR	Diseases	12694	323281
			Chemicals	15411	
2016	(Kaewphan et al., 2016)	CLL	Cell Lines	341	6547
		Gellus		640	278910
2020	(Legrand et al., 2020)	PGxCorpus	Diseases	635	≈ 35000
			Chemicals	1718	
			Genes/Proteins	1708	

Table 1: Corpora with biomedical entities.

based on those corpus and a comparison between them can be established. BioBERT reported state-of-the-art results on BC2GM results with a F1 of 84.72 and in JNLPBA results (77.59) were slightly worse than state-of-the-art which were reported by PubMedBERT with a F1 of 80.06. Results from our fine-tuning model were a bit worse with 83.0 and 76.0 for BC2GM and JNLPBA respectively. Results on this joint entity class are significantly worse than other entity classes, perhaps due to the broad range of subentity classes which take part within this class. This makes the amount of linguistic variability larger, and hence harder to capture than the former entity classes.

Once the models have been fine-tuned, we require some additional steps before having a homogeneous representation of the entities (see Fig. 2). The following section details the entity normalization process and the additional tasks required in our NER+NEN pipeline (Fig. 1).

3 Entity Normalization

The normalization process has been addressed through an inverted index search. Each entity is associated with a set of related terms extracted from external coding systems. Once the medical term is recognized, we search for entities that contain that term in any of their related fields, and we sort that set of candidates based on the BM25 ranking function (Robertson et al., 1994). Those with fewer related terms will have greater relevance. Each type of entity has its own database (i.e index). This way, indexes can be built separately with curated and related terms that helps to map concepts with terms and codes (see Table 2). Multiple sources were taken into account in each of the entity classes, mainly from BioPortal ontologies[8], but also from the Comparative Toxicogenomics Database[9] and PubChem[10].

[8] `http://bioportal.bioontology.org`

[9] `http://ctdbase.org`

[10] `https://pubchem.ncbi.nlm.nih.gov`

Type	Entities	Codes	Sources
Diseases	126573	5	4
Drugs	344238	7	5
Genes	946584	3	4

Table 2: Resources used for normalization.

For each entity, regardless of whether it is a drug, disease or gene/protein, the following information was collected: (1) a *term* or *description* of the underlying concept (e.g. "*Hydroxychloroquine*") ; (2) a list of *synonyms* that holds all possible related words present for a given term (e.g. '*Oxichloroquine*', '*Polirreumin*'); (3) a *semantic type* (e.g. '*Pharmacologic Substance*') and (4) a list of *identifiers* based on MeSH, CUI, ATC, or any other more specific database cross references (e.g. *mesh_id:D006886*, *cid:3652*, *atc:P01BA02*). The range of possibilities to refer to the same element (i.e by code, term or synonym) allow choosing the one with the higher score between different search criteria (e.g. terms or synonyms; strict or similar matches) and filtering criteria (e.g. based on word order, or single terms). The result with the higher score is considered.

3.1 Diseases

Four different sources were merged in the same index to normalize disease terms based on the mappings between their codes and the medical terms used to represent them.

MeSH - Diseases: Medical Subject Headings[11] is a thesaurus with hierarchical and controlled vocabulary produced by the National Library of Medicine (NLM). This thesaurus includes thousands of terms regarding to several semantic types with disease-related terms among them. BioPortal includes an ontology version of this thesaurus from which we have extracted disease-related terms attending to the UMLS Semantic Type each term belongs to.

CTD - Diseases: CTD's MEDIC disease vocabulary is a modified subset of the "Diseases" branch of the NLM's MeSH, combined with genetic disorders from the Online Mendelian Inheritance in Man[12] (OMIM) database. These terms have been merged with the previous ones through an outer join on MeSH IDs.

DOID: The Human Disease Ontology (Schriml et al., 2012) is a comprehensive knowledge base of inherited, developmental and acquired human diseases. It integrates terms from a wide range of medical vocabularies such as MeSH, SNOMED, NCI, or OMIM, and has been used to extend terms which were not previously captured by the other sources. The way this was done is through an outer join on MeSH IDs.

ICD-10-CM: The International Classification of Diseases is a hierarchical classification listed by the World Health Organization (WHO), in which are encoded a wide range of signs, symptoms, abnormal findings, causes of damage, diseases, and/or other disease-related terms. The ICD-10-CM is the 10th version of this classification with a Clinical Modification of the source. Since this classification is used in its proper BioPortal ontology, further mapping concepts are added, which is the case of Unified Medical Language System identifiers (CUIs). The way this source extends the previous sources is through this CUI since not MeSH IDs are included. For that purpose, an outer join on this id was done.

3.2 Drugs

Five sources were considered to merge chemical terms in a shared index. The main objective was to capture the wide range of possible chemical mentions that this entity class can support

PubChem: PubChem is the world largest chemistry open database maintained by the National Institute of Health (NIH). Among the classification systems offered to organize the chemical entities, we used the MeSH hierarchy for our database. Approximately 130000 terms were considered which is expected to have the most widely adopted chemical terms within all the collection.

ChEBI: ChEBI is a chemical database mainly focused on small chemical components of molecular entities and therefore it complements other types of terms considered in the rest of sources. Any biological or synthetical component present in biological organisms is aimed to be captured on this database. An outer join on *InChIKey* was used for connecting these terms with the ones present in the previous source. *InChIKey* is a hashed key of *InChI*, an International Identifier for chemicals, which offers an IUPAC identifier for an standardized codification of

[11] https://www.nlm.nih.gov/mesh

[12] https://www.omim.org

chemicals.

MeSH - Chemicals: MeSH also includes thousands of terms regarding to chemical-related terms. The ontology version in BioPortal has been used to extract chemical-related terms attending to the UMLS Semantic Type each term belongs to. Since PubChem already includes MeSH terms, this source has been just used to add MeSH IDs and extend information from the previous terms. This source was combined with the previous ones through checking if the term is found either on term field or on the synonyms list. If it is not found, it has been appended to chemical terms.

CTD - Chemicals: Database that incorporates terms from multiple chemical sources and therefore it has been used for complementing previously existent processed terms. It also helps to extend the retrieved information about previously considered terms. Non previously found terms have been appended from this source.

ATC: Classification of pharmacological substances organized in therapeutic levels. The ontology version of BioPortal has been the source considered for ATC since it incorporates further information and relations with other terms. Information regarding ATC level and ATC code was added to the previously considered terms. If the term is not present, it has been appended.

3.3 Genetics

This entity class is composed of a broad semantic type since it includes both gene and proteins-related terms. They are close semantic types and even in some occasions the use of the same expressions is diffuse. This has led to a wide range of terms within this entity class in which four large and complementary sources were merged in the same index to cover the biggest amount of entity variability possible.

GO: The knowledgebase underlying the Gene Ontology (Ashburner et al., 2000) is the largest source for the functions of genes and therefore it has been used aiming to capture terms related to genetic mechanisms.

OGG: The Ontology of Genes and Genomes (He, Liu, and Zhao, 2014) collects genes and genomes of certain organisms such as humans, virus and bacteria. Mappings to multiple sources are found in the BioPortal ontology.

	Entities	Coverage (%)	Normalization (%)
Diseases	18,355	49.4	4.2
Drugs	55,120	15.1	22.3
Genes/Proteins	79,063	16.6	9.1

Table 3: CORD-19 statistics (January-2022). Total number of appearances (*Entities*), diversity (*Coverage*) and standardization (*Normalization*) ratio.

PR: The Protein Ontology (Natale et al., 2017) contains a wide range of protein-related entities along with relations between them. This source contains a large amount of terms that covers the protein part.

CTD - Genes: It contains a vocabulary retrieved from multiple sources with a great variety of genes in multiple species. It has been used to extend the gene terms which were not previously captured, appending non-retrieved genes.

4 CORD-19 Entities

The BioNER+BioNEN system described in this paper was used to identify and normalize the drugs, diseases and genetic-related terms mentioned in the CORD-19 corpus (January 2022 Edition). The recognition process was time consuming (approximately 48 days) in a server composed by a 32 CPU-cores Intel Xeon with 256GB RAM. The lack of GPUs made the process considerably slower (i.e. 1173hours at a rate of 0,4s/task) since it requires matrix computation for the transformer-based language models, one for each biomedical concept. The source code is publicly available [13].

Entity recognition and normalization was done for each paragraph of the scientific article. A first group of labels is created to identify the medical terms as they appear in the text (i.e. *diseases_ss*, *chemicals_ss*, *genetics_ss*), and in a standardized way (i.e. *disease_terms_ss*, *chemical_terms_ss*, *genetic_terms_ss*). In the case of diseases and genes/proteins, a predefined category is also established during the normalization process (i.e. *disease_types_ss*, *genetic_types_ss*). The following group of labels contains the codes for each of the classification systems described in Section 3 (i.e. *mesh_codes_ss*, *atc_codes_ss*, *cid_codes_ss*, *doid_codes_ss*, *cui_codes_ss*, *icd10_codes_ss*, *icd9_codes_ss*,

[13] https://github.com/drugs4covid/cord-19

gard_codes_ss, *snomed_codes_ss*, *nci_codes_ss*, *ncbi_codes_ss*, *uniprot_codes_ss*). The suffix *_ss* in all tags indicates that the format is a textual list (i.e. string sequence).

Table 3 shows some statistics about entity classes once the corpus was processed. As expected, almost half of the paragraphs contain at least one mention of a disease or symptom (see column *Coverage*), while drugs, genes or proteins appear less frequently. This is strongly influenced by the criteria used by Allen AI to create the CORD-19 corpus, as they filter articles that contain coronavirus-related terms in their title or abstract. This guide the content of the article and also explains why the variety of disease and symptom entities (see column *Entities*) is far inferior to drugs and genetic information. However, what is striking is the high rate of standard terms (according to our model) used to refer to drugs, with respect to the rest of biomedical entities. Column *Normalization* shows the ratio of entities mentioned in the text using any of the terms extracted from the classification systems described in section 3. We think that there is more flexibility in scientific texts to refer to symptoms or diseases than to drugs or active ingredients, with respect to the standards (e.g. ATC, MeSH, ICD-10 or SNOMED mainly). Regarding genetic information, perhaps the cause lies in the precision in the recognition of the boundary that defines the entity, being sometimes eliminated part of the chemical expression of the entity itself.

Table 4 shows the most widely captured entities according to the following classification systems: ICD-10, MeSH, ChEBI, ATC, MedGen (CUID), GARD, NCBI and SNOMED. Jointly with the code and description of the entity, the occurrences of these words are given (column *Ratio*). This allows us to have an idea about the relevance of the concept in the corpus with respect to the rest of the concepts of the same classification system. In top positions we can find general concepts related to respiratory difficulties. As we go down in the top, more specific terms begin to appear. In the systems that cover diseases such as MeSH or ICD-10, we can find as the most relevant concept the COVID-19 disease, as expected, and the related symptoms (e.g. U07.1 in ICD-10, D000086382 in MeSH or C5203670 in MedGen). The systems more oriented to chemicals identify substances related to respiratory disorders (e.g. Dioxyegn in ChEBI or Oxygen in ATC). And the systems focused on genetic and protein information show, with similar relevance, the pathways of the coronavirus (e.g. Angiotensin converting enzyme 2, Interleukin-6 or Interferon in NCBI).

Thanks to the normalization process that we incorporate in our entity recognition system, we can use the hierarchies defined in the underlying classification system to establish more or less general labels. For example, the Anatomical Therapeutic Chemical (ATC) classification system, which is supported by the World Health Organization (WHO) and widely used in hospital pharmacies to identify drug components, organizes active substances according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties. Drugs are classified into groups at five different levels. The first one corresponds to main groups, the second one to pharmacological or therapeutic subgroups, the third and the fourth one are chemical-pharmacological-therapeutic subgroups and the last one is the chemical substance. Once the code of a drug has been identified in this classification system, we can extend the labels of the text with those groups of the hierarchy, enabling additional ways of exploiting the results of the annotation process.

In the following experiment we want to take advantage of the labels generated by our system to find evidence about the anatomical behavior of drugs used to treat coronavirus. We do not know a priori which groups of drugs are related in this domain, and we assume that an evidence implies the joint presence of several groups in the same paragraph. Since the ATC classification system is hierarchical and establishes 14 anatomic groups at the first level of drug organization, we can create a matrix with the paragraphs where drugs are mentioned and the anatomic groups to which they belong. Figure 3 shows the correlation between each of these anatomical groups based on analysis of mentions of drugs in texts. It can be seen how the highest correlation exists between drugs associated with *Sensory organs* and *Anti-infectives for systemic use*. This may be due to the fact that many of the anti-infective active substances used systemically (i.e. orally or intravenously) are

		Entity	
	Ratio	**Code**	**Description**
ICD-10	51.1	U07.1	COVID-19
	8.0	J12.81	Pneumonia due to SARS-associated coronavirus
	3.7	J11.1	Influenza due to unidentified influenza virus with other respiratory manifestations
	3.3	A79.0	Trench fever
	3.2	F53.0	Postpartum depression
MeSH	34.0	D000086382	COVID-19
	11.1	D000085343	Latent Infection
	4.5	D018352	Coronavirus Infections
	3.9	D003643	Death
	3.7	D045169	Severe Acute Respiratory Syndrome
ChEBI	8.8	15379	Dioxygen
	5.2	33708	Amino-acid Residue
	3.8	172234	TG(14:1(9Z)/22:5(7Z,10Z,13Z,16Z,19Z)/24:1(15Z))
	3.2	30879	Alcohol
	2.9	5801	Hydroxychloroquine
ATC	14.6	V03AN01	Oxygen
	6.4	V04CA02	Glucose
	4.9	P01BA02	Hydroxychloroquine
	3.0	A12AA	Calcium
	2.8	V03AN04	Nitrogen
MedGen (CUI)	40.6	C5203670	COVID-19
	13.2	C0872054	Latent Infection
	6.4	C1175175	Severe acute respiratory syndrome
	4.3	C3714514	Infection
	3.0	C0003467	Anxiety
GARD	26.9	9237	SARS
	11.7	5698	Acute respiratory distress syndrome
	5.9	6427	Farmer's lung
	4.1	2035	Lymphatic filariasis
	2.9	6254	Dengue fever
NCBI	6.6	59272	Angiotensin converting enzyme 2
	5.5	100628202	Interleukin-6
	4.5	100304604	Interferon
	4.3	101180090	Immunoglobulin G level
	4.0	7124	tumor necrosis factor
SNOMED	57.4	840539006	Disease caused by 2019 novel coronavirus (disorder)
	6.3	398447004	Severe acute respiratory syndrome (disorder)
	4.2	155559006	Influenza (disorder)
	4.1	266391003	Pneumonia and influenza or pneumonia (disorder)
	3.8	82214002	Trench fever (disorder)

Table 4: Presence (*Ratio*) of the most frequent entities (*Code*) organized by coding system.

also categorized within the sensory organs group, for example the Ciprofloxacin, since they can also be administered by the otic or ophthalmic route. Thanks to the tags created by our system, it is sufficient to filter the paragraphs labeled with the ATC codes 'S' (i.e *Sensory organs*) and 'J' (i.e. *Anti-infectives*) to find the candidates for evidence. The other most notable correlation is between *Anti-infectives for systemic use* and *Anti-parasitic products*. It could be explained because the anti-infective drugs used for parasites are classified as anti-parasitic products, and the active substances most used experimentally for the treatment of coronavirus were found within these categories, such as Lopinavir/Ritonavir (anti-infective) and Hydroxychloroquine (anti-parasitic). Again, we can take advantage of the tags in our system to find texts in the articles that help us validate this assumption.

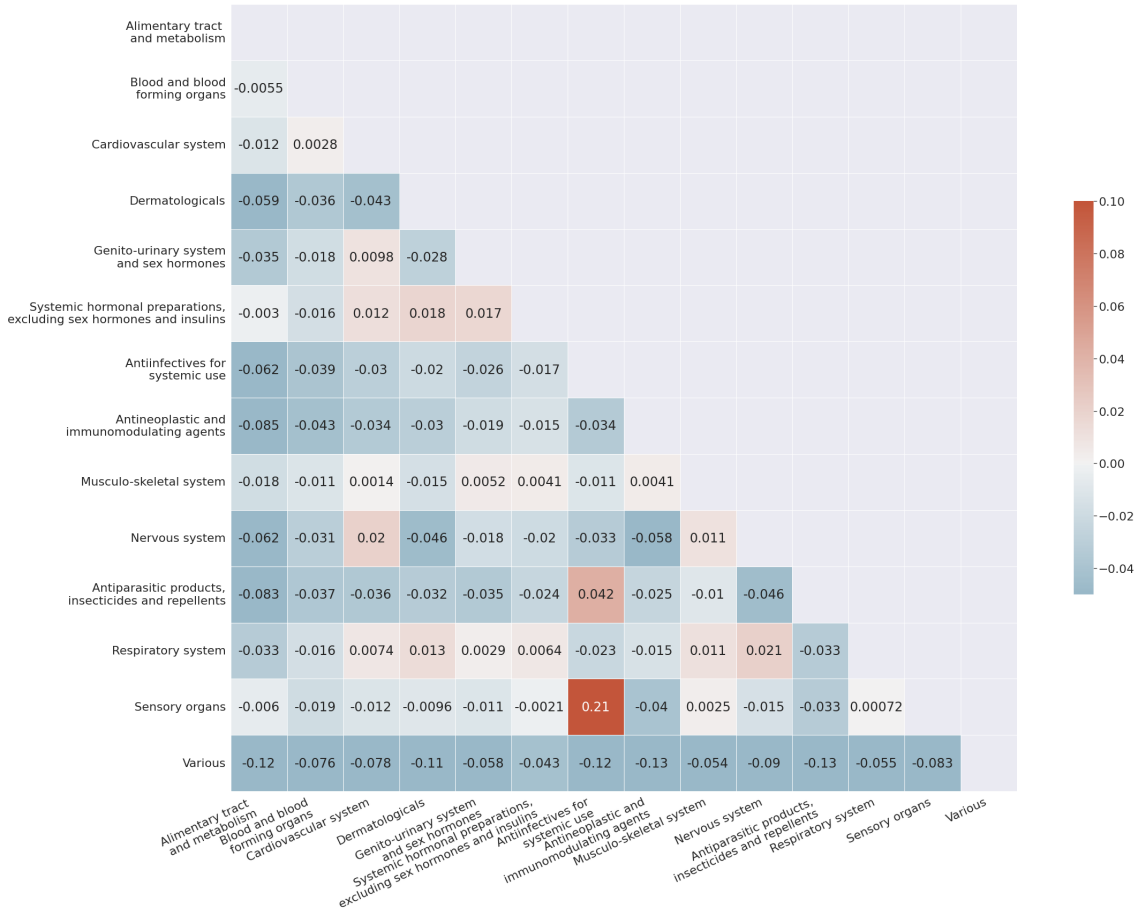

Figure 3: Correlation matrix of ATC data at Anatomical group level.

5 *Conclusions*

We have created a corpus with the diseases, drugs, genes, and proteins mentioned in the paragraphs of the articles in the January edition of the CORD-19 corpus. It contains not only the biomedical entities, but also their normalized references based on several curated databases such as MeSH, ICD-10, ATC, ChEBI or SNOMED. The generated corpus is publicly available and is updated periodically to take up changes in the CORD-19 dataset.

An analysis has been carried out on this corpus to measure the presence and degree of normalization of each type of biomedical entity. As expected, practically half of the paragraphs contain some reference to a disease or symptom. However, only 4% of them were mentioned using any of the standard codes or alias. The behavior in genes and proteins is similar although much lower in terms of presence. Drugs are the least present and most varied type of entity in the corpus. The correlation between the anatomical groups of the drugs has also been measured to value the usefulness of the tags created. The procedure to easily extract the evidence, i.e. paragraphs where the groups are mentioned, is also described.

Our biomedical named entity recognizer created to produce the tags is also described. It is based on the pre-trained BioBERT language model and combines three different models each of them specialized in the recognition of a different biomedical entity: disease, drug and gene/protein. In the future we want to explore the ability of the tags to produce knowledge, either to organize entities or to discover relationships that may arise between them, and to take advantage of the knowledge acquired to create a Spanish BioNER+BioNEN model.

Acknowledgments

Work supported by the *DRUGS4COVID++* project , financed by *Ayudas Fundación BBVA a equipos de investigación científica SARS-CoV-2 y COVID-19*.

References

Akhondi, S. A., A. G. Klenner, C. Tyrchan, A. K. Manchala, K. Boppana, D. Lowe, M. Zimmermann, S. A. Jagarlapudi, R. Sayle, J. A. Kors, et al. 2014. Annotated chemical patent corpus: a gold standard for text mining. *PloS one*, 9(9):e107477.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Bada, M., M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, K. B. Cohen, K. Verspoor, J. A. Blake, et al. 2012. Concept annotation in the craft corpus. *BMC bioinformatics*, 13(1):1–20.

Badenes-Olmedo, C., A. Alonso, and O. Corcho. 2022. Drugs, Diseases, Genes and Proteins in the CORD-19 Corpus, March.

Bagewadi, S., T. Bobić, M. Hofmann-Apitius, J. Fluck, and R. Klinger. 2014. Detecting mirna mentions and relations in biomedical literature. *F1000Research*, 3.

Campos, D., S. Matos, and J. L. Oliveira. 2012. Biomedical named entity recognition: a survey of machine-learning tools. *Theory and Applications for Advanced Text Mining*, 11:175–195.

Chatterjee, A., C. Nardi, C. Oberije, and P. Lambin. 2021. Knowledge graphs for covid-19: An exploratory review of the current landscape. *Journal of Personalized Medicine*, 11(4).

Doğan, R. I., R. Leaman, and Z. Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Goldberg, T., S. Vinchurkar, J. M. Cejuela, L. J. Jensen, and B. Rost. 2015. Linked annotations: a middle ground for manual curation of biomedical databases and text corpora. In *BMC proceedings*, volume 9, pages 1–3. BioMed Central.

Goyal, A., V. Gupta, and M. Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43.

Gururangan, S., A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

He, Y., Y. Liu, and B. Zhao. 2014. Ogg: a biological ontology for representing genes and genomes in specific organisms. In *ICBO*, pages 13–20. Citeseer.

Kaewphan, S., S. Van Landeghem, T. Ohta, Y. Van de Peer, F. Ginter, and S. Pyysalo. 2016. Cell line name recognition in support of the identification of synthetic lethality in cancer from text. *Bioinformatics*, 32(2):276–282.

Kim, J.-D., T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer.

Krallinger, M., O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D. M. Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Legrand, J., R. Gogdemir, C. Bousquet, K. Dalleau, M.-D. Devignes, W. Digan, C.-J. Lee, N.-C. Ndiaye, N. Petitpain, P. Ringot, et al. 2020. Pgxcorpus, a manually annotated corpus for pharmacogenomics. *Scientific data*, 7(1):1–13.

Li, J., Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Li, J., A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Nadeau, D. and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Natale, D. A., C. N. Arighi, J. A. Blake, J. Bona, C. Chen, S.-C. Chen, K. R. Christie, J. Cowart, P. D'Eustachio, A. D. Diehl, et al. 2017. Protein ontology (pro): enhancing and scaling up the representation of protein entities. *Nucleic acids research*, 45(D1):D339–D346.

Ohta, T., S. Pyysalo, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, S. Ananiadou, and J. Tsujii. 2013. Overview of the pathway curation (pc) task of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 67–75.

Pafilis, E., S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390.

Peng, Y., S. Yan, and Z. Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Perera, N., M. Dehmer, and F. Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, 8:673.

Pyysalo, S. and S. Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Pyysalo, S., F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen, and T. Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):1–24.

Pyysalo, S., T. Ohta, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, J. Tsujii, and S. Ananiadou. 2015. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. *BMC bioinformatics*, 16(10):1–19.

Robertson, S. E., S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at trec-3. In *TREC*.

Schriml, L. M., C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. 2012. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946.

Segura Bedmar, I., P. Martínez, and M. Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Smith, L., L. K. Tanabe, R. J. nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):1–19.

Wang, L. L., K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al. 2020. CORD-19: The Covid-19 Open Research Dataset. *arXiv preprint arXiv:2004.10706*.

Yadav, V. and S. Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

Zhou, G., J. Zhang, J. Su, D. Shen, and C. Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190, May.