A corpus for entity recognition in COVID-19 full-text literature*

Xin An
School of Economics and Management
Beijing Forestry University
Beijing 100083, P.R. China
anxin@bjfu.edu.cn

Mengmeng Zhang
School of Economics and Management
Beijing Forestry University
Beijing 100083, P.R. China
zhangmm@bjfu.edu.cn

Shuo Xu†
College of Economics and Management
Beijing University of Technology
Beijing 100124, P.R. China
xushuo@bjut.edu.cn

ABSTRACT

To support the development of entity recognition tools, this study manually annotates 99 full-text articles about COVID-19. Each article is annotated by 6 annotators through two rounds. 18 types of entity are involved, including genes, diseases, chemicals, coronaviruses and so on. We also calculate the inter-annotator agreement (IAA) scores in term of multi-κ measure to ensure the quality of the annotations. In the end, 39,118 entity mentions are manually annotated in total.

KEYWORDS

COVID-19, Entity Annotation, Corpus, Full Text

ACM Reference format:

Xin An, Mengmeng Zhang and Shuo Xu. 2022. A corpus for entity recognition in COVID-19 full-text literature. 3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022). At the ACM/IEEE Joint Conference on Digital Libraries 2022 (JCDL2022), Cologne, Germany and Online, 2 pages.

1 Introduction

In December 2019, an outbreak of COVID-19 caused by SARS-CoV-2 broke out, and it still threatens people's lives and health up to now. During this 3-year period, researchers did their best to study the infection, symptoms and diagnosis of COVID-19, and lots of scholarly articles are published. These achievements have played an important role in curbing the spread of SARS-CoV-2. What's more, the entity recognition from these scientific publications can help identify the source of SARS-CoV-2, discover the chain of infection and so on.

To the best of our knowledge, multiple entity annotation corpora are used in the biomedical field [1][2], but there is no manually annotated corpus specifically for COVID-19. Furthermore, majority of previous corpora only annotated entities mentioned in the title and abstract of each article or patent [1][3]. Therefore, we built a high-quality manually annotated corpus in full-text articles from the COVID-19 field. The corpus focuses not only on basic biological entities (genes, diseases, chemicals mutations, etc.) but also on new types of entities associated with COVID-19 (e.g., coronaviruses and viral proteins), making it a valuable resource for downstream analysis of COVID-19.

† Corresponding author

2 Materials and Methods

2.1 Document selection

Our full texts come from the CORD-19 dataset (COVID-19 Open Research Dataset) (2020-3-13 version)[4]. The dataset has 29,500 publications, in which 13,200 articles are attached with full texts. All publications are further divided into several categories in terms of data source, access license, and availability of full text. Then, we randomly select 80 articles according to the category distribution. It is noteworthy that the round-up operation is utilized in this study to ensure at least one article per category. In the end, 99 articles are chosen.

2.2 Entity Types

We define 18 types of entity after literature review and expert consultation. The entity types and the resulting examples are shown in Table 1. Note that DISEASE, CHEMICAL, BACTERIUM and GENE are also contained in the task of BioNLP-OST 2019. In addition, the citation and reference are also defined as two types of entities.

Table 1: Types of entity

ID	Entity types	Examples
1	GENE_OR_PROTEIN_OR_ENZYME	CSHGs, ACE2
2	CITATION	[1], (Wang et al)
3	NON_CORONAVIRUS	MERS
4	CHEMICAL	amoxicillin
5	DISEASE_OR_SYMPTOM	aseptic meningitis
6	LABORATORY_TECHNIQUE	qRT-PCR
7	REFERENCE	Fig.1, Table 1
8	BODY_ORGAN	heart
9	LABORATORY_ANIMAL	C57BL/6 mice
10	PERSON	people, children
11	BACTERIUM	enterococcus
12	BODY_SUBSTANCE	serum, urine
13	CORONAVIRUS	SARS-CoV-2

14	WILDLIFE	bat, monkey
15	LIVESTOCK	pig, sheep
16	OTHER_ANIMAL	parasites
17	MATERIAL	silver
18	PET	cat, dog

2.3 Annotation guidelines

Before annotating, an annotation guideline is created, which includes 3 parts: (a) what should be annotated as an entity, (b) general guidelines for annotating entity mentions, (c) what should not be annotated. This guideline clarifies the entity types by referring to the entity definitions in the UMLS and Wikipedia. In the meanwhile, for ease of understanding, several examples from our corpus are provided. The second part illustrates general rules. For example, "(Fig.1)" is marked as "Fig.1" (only words in parentheses matter), and "[1]" is marked as a whole (the words and square brackets matter). The span of each mention should be the shortest continuous text that identifies the entity. The abbreviations should be annotated separately from their long forms. The third part points out the cases we don't need to annotate. General concepts were excluded from the annotation process, such as protein(s), cell(s), and human. The entities that are embedded in other words, even if they are co-incidentally the same set of characters, should not be annotated. Each regulation in the guideline is attached with the corresponding examples for the convenient comprehension and follow-up applications.

3 Annotation process and results

3.1 Annotation process

We use the visual annotation tool BRAT[5]for entity annotation. Our team consists of one administrator and six annotators. Each annotator is assigned an account. Entity annotation is composed of two rounds. During the first round, 6 annotators annotate all 99 articles independently. They are allowed to refer to the classification of entities on the UMLS Semantic Network and Internet. But discussions are forbidden during the annotation process to learn the difficulty and consistency. After the completion of the first round, all annotations for each document are merged by the administrator to identify the agreements and disagreements. Then, all annotators discuss the ambiguous annotations until consensus is reached.

3.2 Annotation results

We calculate the IAA score[1] of each article after the first round, as shown in Figure 1. Among the 99 articles, the IAA score of each article mainly clustered between 0.4 and 0.6. Especially, the scores of 30 articles range from 0.5 to 0.6. After in-depth analysis, the following two types of errors can be found: (1) Several entity mentions are labeled by an annotator, but missed by the others; (2) The span of a focal mention is different amongst six annotators. Let's take the mention "Broad-Spectrum Coronavirus Protease Inhibitor" as an example. Some annotate the entire mention as a mention, but the others only annotate "Coronavirus" as an entity. After extensive discussion, the IAA score of each article reaches 100% in the second round.

Table 2: The distribution of entity mentions

ID	Num of Entity	%	ID	Num of Entity	%
1	9917	25.35	10	761	1.95
2	5957	15.23	11	665	1.7
3	4128	10.55	12	599	1.53
4	4040	10.33	13	554	1.42
5	3319	8.48	14	543	1.39
6	2754	7.04	15	446	1.14
7	1856	4.74	16	425	1.09
8	1594	4.07	17	340	0.87
9	1184	3.03	18	36	0.09

Table 2 lists the distribution of 18 entities. GENE_OR_PROTEIN_OR_ENZYME has the highest numbers (9,917) among all entity types and accounts for 25.35%. There are 5,957 CITATION entities, which is the 2nd highest percentage at 15.23%. NON_CORONAVIRUS and CHEMICAL entities have 4,128 (10.55%) and 4040 (10.33%), which are in the third and fourth place, followed by DISEASE_OR_SYMPTOM (8.48%) and LABORATORY_TECHNIQUE (7.04%). Compared to the top 6 entity types, the other entities only account for less than 5%. This indicates that these entities are not so related to COVID-19.

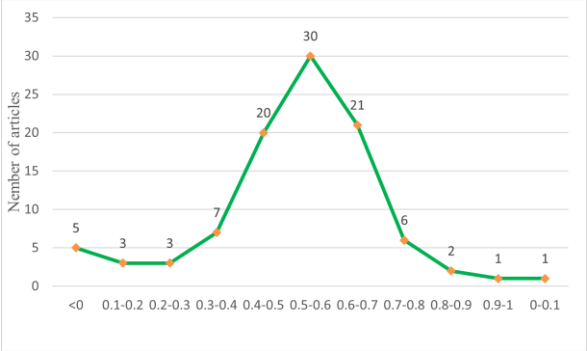

Figure 2: The IAA scores of 99 articles in the first round

4 Discussion

We create a high-quality manual annotated corpus about COVID-19. It includes 18 categories of entities, focusing not only on entities such as genes, proteins and chemicals that are widely present in the biological field, but also on entities related to COVID-19. We annotate 39,118 entities in total from 99 full-text articles. On average, each document mentions about 395 entities. Compared with several other corpora of biomedical field in recent years (such as NLM-gene[6], BC5CDR[7] and NLM-Chem[8]), the scale of our corpus is

relatively large in term of the number of full-text documents. As a resource of COVID-19, this corpus can lay a foundation for subsequent related research. In the future, we will train a model on the basis of this corpus to automatically identity the entities from the other publications. This corpus will be opened to the community in the near future.

ACKNOWLEDGMENTS

This research received the financial support from the National Natural Science Foundation of China under grant number 72004012 and 72074014.

REFERENCES

[1] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez and David Salgado. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of Cheminformatics, 7 (Suppl 1), S2. DOI: 10.1186/1758-2946-7-S1-S2.

[2] Xu S, An X, Zhu L, Zhang Y, and Zhang H. 2015. A CRF-based System for Recognizing Chemical Entity Mentions (CEMs) in Biomedical Literature. Journal of Cheminformatics, 7 (Suppl 1): S11. DOI: 10.1186/1758-2946-7-S1-S11

[3] Chen L, Xu S, Zhu L, Zhang J, Lei X, and Yang G. 2020. A Deep Learning based Method for Extracting Semantic Information from Patent Documents. Scientometrics, 125(1): 289-312.

[4] Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Burdick D, et al. 2020. CORD-19: The COVID-19 Open Research Dataset. Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, arXiv:2004.10706v2. PMID: 32510522.

[5] Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta and Sophia Ananiadou. 2012. BRAT: A Web-based tool for NLP-Assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 102 – 107.

[6] Islamaj R, Wei C, Cissel D, et al. 2021. NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. Journal of biomedical informatics, 118:103779. DOI: 10.1016/j.jbi.2021.103779.

[7] Li J, Sun Y, Johnson R, et al. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford), baw068. DOI: 10.1093/database/baw068.

[8] Islamaj R, Leaman R, Kim S, et al. 2021. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. Scientific data, 8(1):91. DOI: 10.1038/s41597-021-00875-1.