Comprehensive Named Entity Recognition on CORD-19 with Distant or Weak Supervision

Xuan Wang[1], **Xiangchen Song[1]**, **Bangzheng Li[1]**, **Yingjun Guan[2]**, **Jiawei Han[1]**

[1]Department of Computer Science, University of Illinois at Urbana-Champaign
[2]School of Information Sciences, University of Illinois at Urbana-Champaign
[1,2]{xwang174,xs22,yingjun2,bl17@,hanj}@illinois.edu

Abstract

We created this CORD-NER dataset with comprehensive named entity recognition (NER) on the COVID-19 Open Research Dataset Challenge (CORD-19) corpus (2020-03-13). This CORD-NER dataset covers 75 fine-grained entity types: In addition to the common biomedical entity types (e.g., genes, chemicals and diseases), it covers many new entity types related explicitly to the COVID-19 studies (e.g., coronaviruses, viral proteins, evolution, materials, substrates and immune responses), which may benefit research on COVID-19 related virus, spreading mechanisms, and potential vaccines. CORD-NER annotation is a combination of four sources with different NER methods. The quality of CORD-NER annotation surpasses SciSpacy (over 10% higher on the F1 score based on a sample set of documents), a fully supervised BioNER tool. Moreover, CORD-NER supports incrementally adding new documents as well as adding new entity types when needed by adding dozens of seeds as the input examples. We will constantly update CORD-NER based on the incremental updates of the CORD-19 corpus and the improvement of our system.

1 Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease was first identified in 2019 in Wuhan, Central China, and has since spread globally, resulting in the 20192020 coronavirus pandemic. On March 16th, 2020, researchers and leaders from the Allen Institute for AI, Chan Zuckerberg Initiative (CZI), Georgetown University's Center for Security and Emerging Technology (CSET), Microsoft, and the National Library of Medicine (NLM) at the National Institutes of Health released the COVID-19 Open Research Dataset (CORD-19)[1] of scholarly literature about COVID-19, SARS-CoV-2, and the coronavirus group.

Named entity recognition (NER) is a fundamental step in text mining system development to facilitate COVID-19 studies. There is a critical need for NER methods that can quickly adapt to all the COVID-19 related new types without much human effort for training data annotation. We created this **CORD-NER dataset**[2] with comprehensive named entity annotation on the CORD-19 corpus (2020-03-13). This dataset covers 75 fine-grained named entity types. CORD-NER is automatically generated by combining the annotation results from four sources. In the following sections, we introduce the details of CORD-NER dataset construction. We also show some NER annotation results in this dataset.

2 CORD-NER Dataset

2.1 Corpus

The input corpus is generated from the 29,500 documents in the CORD-19 corpus (2020-03-13). We first merge all the meta-data (all_sources_metadata_2020-03-13.csv) with their corresponding full-text papers. Then we create a tokenized corpus (CORD-NER-corpus.json) for further NER annotations.

The input corpus is a combination of the "title", "abstract" and "full-text" from the CORD-19 corpus. We first conduct automatic phrase mining and tokenization on the input corpus using AutoPhrase (Shang et al., 2018a). Then we do a second round of tokenization with Spacy[3] on the phrase-replaced

---

[1]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
[2]https://xuanwang91.github.io/2020-03-20-cord19-ner/
[3]https://spacy.io/api/annotation#

	Gene			Chemical			Disease			Total		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
SciSpacy (BIONLP13CG)	**91.48**	**82.06**	**86.51**	64.66	39.81	49.28	8.11	2.75	4.11	76.36	53.59	62.98
SciSpacy (BC5CDR)	-	-	-	**86.97**	51.86	64.69	**80.31**	59.65	68.46	**82.40**	54.57	65.66
Ours	82.14	74.68	78.23	82.93	**75.22**	**78.89**	75.73	**68.42**	**71.89**	81.29	**73.65**	**77.28**

Table 1: Performance comparison on three major biomedical entity types in COVID-19 corpus.

corpus. We found that keeping the AutoPhrase results will significantly improve the distantly- and weakly-supervised NER performance.

2.2 NER Methods

CORD-NER annotation is a combination of four sources with different NER methods:

1. Pre-trained NER on 18 general entity types from Spacy using the model "en_core_web_sm".

2. Pre-trained NER on 18 biomedical entity types from SciSpacy[4] using the models "en_ner_bionlp13cg_md" and "en_ner_bc5cdr_md".

3. Knowledgebase (KB)-guided NER on 127 biomedical entity types with our distantly-supervised NER methods (Wang et al., 2019; Shang et al., 2018b). We do not require any human-annotated training data for the NER model training. Instead, We rely on UMLS [5] as the input KB for distant supervision.

4. Seed-guided NER on nine new entity types (specifically related to the COVID-19 studies) with our weakly-supervised NER method. We only require several (10-20) human-input seed entities for each new type. Then we expand the seed entity sets with CatE (Meng et al., 2020) and apply our distant NER method for the new entity type recognition.

We reorganized all the entity types from the four sources into one entity type hierarchy (CORD-NER-types.xlsx). Specifically, we align all the types from SciSpacy to UMLS. We also merge some fine-grained UMLS entity types to their more coarse-grained types based on the corpus count. Our entity type hierarchy covers 75 fine-grained

entity types: In addition to the common biomedical entity types (e.g., genes, chemicals and diseases), it covers many new entity types related explicitly to the COVID-19 studies (e.g., coronaviruses, viral proteins, evolution, materials, substrates and immune responses), which may benefit research on COVID-19 related virus, spreading mechanisms, and potential vaccines.

Then we conduct named entity annotation on the 75 fine-grained entity types with the four sources of NER methods. After we get the NER annotation result with each method, we merge the results into one NER annotation file (CORD-NER.json). The conflicts are resolved by giving priority to different entity types annotated by different methods according to their annotation quality. Finally, we merge all the related information (meta-data, full-text corpus and NER results) into one file (CORD-NER-full.json) for users' convenience. The size of the dataset is about 1.2GB.

3 Results

3.1 NER Annotation Results

In Table 1, we show the performance comparison between our annotation and the SciSpacy models. BIONLP13CG is the model in SciSpacy that covers the most entity types (18 entity types). BC5CDR is another model in SciSpacy that has the best performance on two entity types (chemicals and diseases). We manually annotated more than 1000 sentences for evaluation. Then we calculate the precision, recall and F1 scores on three major biomedical entity types: gene, chemical and disease. We can see that our annotation has worse performance on the gene type but much better performance on the chemical and disease types. In summary, the quality of our annotation surpasses SciSpacy by a large margin (over 10% higher on the F1 score). Moreover, SciSpacy requires human effort for training data annotation and covers only 18 types. Our NER system supports incrementally adding new documents as well as adding new entity types when needed by adding dozens of seeds as the input examples.

Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 **CORONAVIRUS** receptor: molecular mechanisms and potential therapeutic target. SARS-CoV-2 **CORONAVIRUS** has been sequenced [3 **CARDINAL**] . A phylogenetic **EVOLUTION** analysis [3 **CARDINAL** , 4 **CARDINAL**] found a bat **WILDLIFE** origin for the SARS-CoV-2 **CORONAVIRUS** . There is a diversity of possible intermediate hosts for SARS-CoV-2 **CORONAVIRUS** , including pangolins **WILDLIFE** , but not mice **EUKARYOTE** and rats **EUKARYOTE** [5 **CARDINAL**] . There are many similarities of SARS-CoV-2 **CORONAVIRUS** with the original SARS-CoV **CORONAVIRUS** . Using computer modeling , Xu et al . [6 **CARDINAL**] found that the spike proteins **GENE_OR_GENOME** of SARS-CoV-2 **CORONAVIRUS** and SARS-CoV **CORONAVIRUS** have almost identical 3-D structures in the receptor binding domain that maintains Van der Waals forces **PHYSICAL_SCIENCE** . SARS-CoV spike proteins **GENE_OR_GENOME** has a strong binding affinity to human ACE2 **GENE_OR_GENOME** , based on biochemical interaction studies and crystal structure analysis [7 **CARDINAL**] . SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** share identity in amino acid sequences and ……

(a) CORD-19 corpus

The U.S. **GPE** now leads the world in confirmed coronavirus **CORONAVIRUS** cases.

Scientists **GROUP** warned that the United States **GPE** someday would become the country hardest hit by the coronavirus **CORONAVIRUS** pandemic. That moment arrived on Thursday **DATE** . In the United States **GPE** , at least 81,321 **CARDINAL** people are known to have been infected with the coronavirus **CORONAVIRUS** , including more than 1,000 **CARDINAL** deaths **ORGANISM_FUNCTION** — more cases than China **GPE** , Italy **GPE** or any other country has seen, according to data gathered by The New York Times **ORG** . With 330 million **CARDINAL** residents, the United States **GPE** is the world's third **ORDINAL** most populous nation, meaning it provides a vast pool of people who can potentially get Covid-19 **CORONAVIRUS** , the disease caused by the virus. And it is a sprawling, cacophonous democracy, where states set their own policies **INTELLECTUAL_PRODUCT** and President Trump **PERSON** has sent mixed messages **INTELLECTUAL_PRODUCT** about the scale of the danger and …… a failure to take the pandemic seriously even as it engulfed China **GPE** , a deeply flawed effort to provide broad testing for the virus that left the country blind to the extent of the crisis, and a dire shortage of masks **MANUFACTURED_OBJECT** and protective gear to protect doctors **GROUP** and nurses **GROUP** on the front lines, as well as ventilators **MANUFACTURED_OBJECT** to keep the critically ill alive.

(b) New York Times corpus

Figure 1: Examples of the annotation results with CORD-NER system.

In Figure 1a, we show some examples of the annotation results in CORD-NER. We can see that our distantly- or weakly supervised methods achieve high quality recognizing the new entity types, requiring only several seed examples as the input. For instance, we recognized "SARS-CoV-2" as the "CORONAVIRUS" type, "bat" and "pangolins" as the "WILDLIFE" type and "Van der Waals forces" as the "PHYSICAL_SCIENCE" type. This NER annotation can help downstream text mining tasks in discovering the origin and the physical nature of the virus. Also, our NER methods are domain-independent that can be applied to the corpus in different domains. We show another example of NER annotation on the New York Times corpus with our system in Figure 1b.

In Figure 2, we show the comparison of our annotation with existing fully-supervised NER/BioNER systems. In Figure 2a, we can see that our method

can identify "SARS-CoV-2" as a coronavirus. In Figure 2b, we can see that our method can identify many more entities such as "phylogenetic" as an evolution term and "bat" as a wildlife term. In Figure 2c, we can also see that our method can identify many more entities such as "racism" as social behavior. In summary, our distantly- and weakly-supervised NER methods are reliable for high-quality entity recognition without requiring human effort for training data annotation.

3.2 Top-Frequent Entity Summarization

In Table 2, we show some examples of the most frequent entities in our annotated corpus. Specifically, we show the entity types, including both our new types and some UMLS types that have not been manually annotated before. We find our annotated entities very informative for the COVID-19 studies. For example, the most frequent enti-

Spacy (General NER):

> Angiotensin-converting enzyme 2 **CARDINAL** (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target.

SciSpacy (Biomedical NER):

> Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 receptor **GENE_OR_GENOME** : molecular mechanisms and potential therapeutic target.

Ours:

> Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 **CORONAVIRUS** receptor: molecular mechanisms and potential therapeutic target.

(a)

Spacy (General NER):

> A phylogenetic analysis [3 **CARDINAL** , 4 **CARDINAL**] found a bat origin for the SARS-CoV-2.

SciSpacy (Biomedical NER):

> A phylogenetic analysis [3, 4] found a bat origin for the SARS-CoV-2 **SIMPLE_CHEMICAL** .

Ours:

> A phylogenetic **EVOLUTION** analysis [3 **CARDINAL** , 4 **CARDINAL**] found a bat **WILDLIFE** origin for the sars_cov_2 **CORONAVIRUS** .

(b)

Spacy (General NER):

> The covid-19 pandemic , rapid spread and magnitude unleashed panic and episodes of racism against people of asian **NORP** descent .

SciSpacy (Biomedical NER):

> The covid-19 **GENE_OR_GENE_PRODUCT** pandemic , rapid spread and magnitude unleashed panic and episodes of racism against people **ORGANISM** of asian descent .

Ours:

> The covid-19 **CORONAVIRUS** pandemic , rapid spread and magnitude unleashed panic and episodes of racism **SOCIAL_BEHAVIOR** against people **ORGANISM** of asian **NORP** descent **GROUP** .

(c)

Figure 2: Annotation result comparison with other NER methods.

ties for the type "SIGN_OR_SYMPTOM behavior" includes "cough" and "respiratory symptoms" that are the most common symptoms for COVID-19. The most frequent entities for the type "INDIVIDUAL_BEHAVIOR" include "hand hygiene", "disclosures" and "absenteeism", which indicates that people focus more on hand cleaning for the COVID-19 issue. Also, the most frequent entities for the type "MACHINE_ACTIVITY" include "machine learning", "data processing" and "automation", which indicates that people focus more on automated methods that can process massive data for the COVID-19 studies. This type also includes "telecommunication" as the top results, which is quite reasonable under the current COVID-19 situation. More examples can be found in our dataset.

4 Conclusion

CORD-NER will be constantly updated based on the incremental updates of the CORD-19 corpus and the improvement of our system. We will also build text mining systems based on the CORD-NER dataset with richer functionalities. We hope this dataset can help the text mining community build downstream applications for the COVID-19 related tasks. We also hope this dataset can bring insights for the COVID-19 studies on making scientific discoveries.

Acknowledgment

CORONAVIRUS	EVOLUTION	WILDLIFE	PHYSICAL SCIENCE
sars	mutation	bat	positively charged
cov	phylogenetic	wild birds	negatively charged
mers	evolution	wild animals	force field
covid-19	recombination	fruit bats	highly hydrophobic
sars-cov-2	substitutions	pteropus	van der waals interactions
LIVESTOCK	**MATERIAL**	**SUBSTRATE**	**IMMUNE_ RESPONSE**
pigs	air	blood	immunization
poultry	plastic	urine	immunity
calves	fluids	sputum	immune cells
chicken	copper	saliva	innate immune
pig	silica	fecal	inflammatory response
SIGN_OR_SYMPTOM	**SOCIAL_BEHAVIOR**	**INDIVIDUAL_BEHAVIOR**	**THERAPEUTIC_OR _PREVENTIVE _PROCEDURE**
cough	collaboration	hand hygiene	detection
respiratory symptoms	sharing	disclosures	vaccination
diarrhoea	herd	absenteeism	isolation
vomiting	mediating	compliance	stimulation
wheezing	adoption	empathy	inoculation
DIAGNOSTIC_PROCEDURE	**RESEARCH_ACTIVITY**	**EDUCATIONAL_ACTIVITY**	**MACHINE_ACTIVITY**
imaging	rt-pcr	health education	machine learning
immunohistochemistry	sequencing	workshops	data processing
necropsy	screening	nursery	automation
scanning	diagnosis	medical education	deconvolution
biopsy	prevention	residency	telecommunication

Table 2: Examples of the most frequent entities annotated in CORD-NER.

References

Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020 (WWW20)*.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018a. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837.

Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018b. Learning named entity tagger using domain-specific dictionary. In *EMNLP*. ACL.

Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019. Distantly supervised biomedical named entity recognition with dictionary expansion. In *BIBM*.