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

Xin An[1], Mengmeng Zhang[1], Shuo Xu[2]
1 Beijing Forestry University 2 Beijing University of Technology

Introduction

◆**Background**: In December 2019, an outbreak of COVID-19 caused by SARS-CoV-2 broke out; During this 3-year period, and lots of scholarly articles are published; The entity recognition from these scientific publications can help identify the source of SARS-CoV-2.

◆**Research Target**: Building a high-quality manually annotated corpus in full-text articles in the COVID-19 field.

◆**Research significance**: A valuable resource for downstream analysis of COVID-19; Help for Text mining task: Entity/relation recognition and so on.

Materials and Methods

◆**Document selection**

- Source: CORD-19 dataset (COVID-19 Open Research Dataset)
- Selection: extracting 99 full-text articles according to the proportion of categories.

◆**Entity Types**

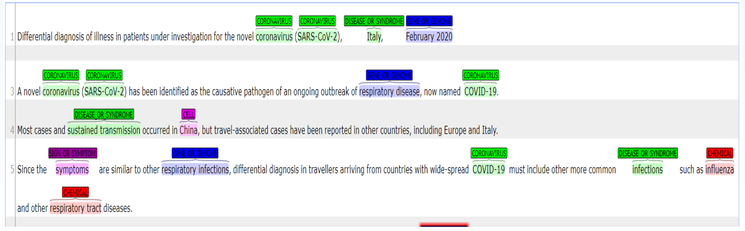
- Defining 18 types of entity after literature review and expert discussion.

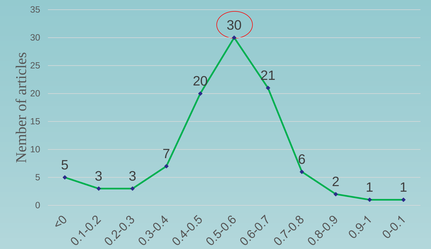
Entity types	Examples
GENE_OR_PROTEIN_OR_ENZYME	CSHGs, ACE2
CITATION	[1], (Wang et al)
NON_CORONAVIRUS	MERS
CHEMICAL	amoxicillin
DISEASE_OR_SYMPTOM	aseptic meningitis
LABORATORY_TECHNIQUE	qRT-PCR
REFERENCE	Fig.1, Table 1
BODY_ORGAN	heart
LABORATORY_ANIMAL	C57BL/6 mice
PERSON	people, children
BACTERIUM	enterococcus
BODY_SUBSTANCE	serum, urine
CORONAVIRUS	SARS-CoV-2
WILDLIFE	bat, monkey
LIVESTOCK	pig, sheep
OTHER_ANIMAL	parasites
MATERIAL	silver
PET	cat, dog

Annotation process and results

◆ **Annotation tool** : BRAT
◆ **Annotation team**: 6 annotators and 1 manager
◆ **Annotation process**: (two rounds)

- The first rounds：





Calculating the IAA score of each article. Mainly articles clustered between 0.4 and 0.6.

- **The second rounds: distribution of 18 entities**

Entity types	Num of Entity	%	Entity types	Num of Entity	%
GENE_OR_PROTEIN_OR_ENZYME	9,917	25.35	PERSON	761	1.95
CITATION	5,957	15.23	BACTERIUM	665	1.7
NON_CORONAVIRUS	4,128	10.55	BODY_SUBSTANCE	599	1.53
CHEMICAL	4,040	10.33	CORONAVIRUS	554	1.42
DISEASE_OR_SYMPTOM	3,319	8.48	WILDLIFE	543	1.39
LABORATORY_TECHNIQUE	2,754	7.04	LIVESTOCK	446	1.14
REFERENCE	1,856	4.74	OTHER_ANIMAL	425	1.09
BODY_ORGAN	1,594	4.07	MATERIAL	340	0.87
LABORATORY_ANIMAL	1,184	3.03	PET	36	0.09

Conclusion

➢ Creating a high-quality manual annotated corpus about COVID-19 using 99 full-text articles. It includes 18 categories of entities and 39,118 entities in total. On average, each document mentions about 395 entities.

➢ As a resource of COVID-19, this corpus can lay a foundation for subsequent related research.