English dictionaries, gold and silver standard corpora for biomedical natural language processing related to SARS-CoV-2 and COVID-19

Salma Kazemi Rashed[1], *Johan Frid*[2], *Sonja Aits*[1, 3, 4]

[1] Cell Death, Lysosomes and Artificial Intelligence Group, Department of Experimental Medical Science, Faculty of Medicine, Lund University, Lund, Sweden
[2] Humanities Lab, Lund University, Lund, Sweden
[3] Lund Institute for Advanced Neutron and X-Ray Science (LINXS), Lund , Sweden
[4] For correspondence contact: sonja.aits@med.lu.se

Abstract

Here we present a toolbox for natural language processing tasks related to SARS-CoV-2. It comprises English dictionaries of synonyms for SARS-CoV-2 and COVID-19, a silver standard corpus generated with the dictionaries and a gold standard corpus of 10 Pubmed abstracts manually annotated for disease, virus, symptom and protein/gene terms. This toolbox is freely available on github (on https://github.com/Aitslab/corona) and can be used for text analytics in a variety of settings related to the COVID-19 crisis. It will be expanded and applied in NLP tasks over the next weeks and the community is invited to contribute.

Introduction

The analysis of various types of text from scientific articles to tweets can be used to support research on SARS-CoV-2/COVID-19 and public health management of the pandemic. It can also be used for studying the social impact of the crisis.

One critical step in biomedical natural language processing (BioNLP), as for most types of text analysis, is the identification of relevant keywords and phrases as well as the detection of synonymous expressions. This so-called named entity recognition (NER) and disambiguation can be especially challenging with newly emerging diseases such as COVID-19 because no official name had been defined during the early phases of the outbreak. Instead, authors used a wide range of descriptive terms such as “Wuhan seafood market pneumonia”.

NER can be performed using dictionaries of keywords or with the help of models, which are often trained on annotated corpora in which keywords have been labelled manually by experts (gold standard) or automatically (silver standard) [1]. Dictionary and model-based methods can also be combined as the approaches are complimentary. Here, we have developed a toolbox for NLP related to SARS-CoV-2 and COVID-19. It includes English dictionaries of synonymous terms as well as an annotated gold and silver standard corpus. This work will be expanded over the next weeks but due the exceptional circumstances of the current health crisis we have chosen to release the tools already at this stage so that other researchers can use them and contribute to their improvement. Updated versions will be published on https://github.com/Aitslab/corona.

Methods

Generation of SARS-CoV-2 and COVID-19 dictionaries

Synonyms and biomedical identifiers for COVID-19 and SARS-CoV-2 were identified by reviewing a variety of databases and text sources including: NCBI Taxonomy database, wikidata, the International Classification of Diseases (v. 10 and v. 11), Disease Ontology (https://disease-ontology.org), Medical Subject Headings (MeSH), medical literature, twitter feeds and newspaper websites.

From this manually curated list, hyphens were stripped and letters changed to lower case. Then variants for virus names and disease names were generated as follows:

Virus names:

- Adding '2019', '2019novel', '2019new', '2019 novel', '2019 new' as prefixes or 2019 as suffix (only if '19' was not in the virus name)
- interchanging corona virus and coronavirus
- removing virus names containing the same word twice or containing both 'new' and 'novel'

Disease names:

- interchanging corona virus and coronavirus
- interchanging the virus name with any of its synonyms, the term 'Wuhan' or 'Hubei'
- interchanging the terms disease, disorder, syndrome, pneumonia, infection
- adding the adjectives acute, severe and respiratory alone or in combination
- removing disease names containing the same word twice or containing both 'new' and 'novel'

Duplicate dictionary entries as well as automatically generated entries with duplicate or semantically similar words (e.g. containing 2019 in beginning and end, containing novel and new) were removed.

Production of Corona silver standard corpus

SARS-Cov-2/COVID-19 related studies were identified on Pubmed with the search term (((COVID-19 OR SARS-CoV-2 OR (Wuhan AND virus) OR 2019-nCoV OR (Wuhan AND pneumonia)) AND English[lang])) AND ("2019/12"[Date - Create] : "2020"[Date - Create]). The 987 detected records were downloaded in text format (even those containing only a heading and no abstract). The text was stripped of hyphens and '/' replaced by a blank space. Tokenization (chopping up text into individual words/terms), and detection of dictionary terms was performed by spacy (https://spacy.io/) with a spacy-lookup module in a capitalization-independent manner. A second tokenization step was performed by splitting the terms matching the dictionaries on blank spaces. Annotations were stored in IOB2 format where the first token of an entity is labelled 'B', subsequent tokens belonging to the entity are labelled 'I' and all other tokens are labelled 'O'. When matches were found in both dictionaries, the correct one was selected manually.

Production of Corona gold standard corpora

Using BioQRator (http://www.bioqrator.org/), English abstracts related to COVID-19/SARS-CoV-2 and published between Dec 2019 and March 6 2020 were identified on PubMed with the following search query: (((COVID-19 OR SARS-CoV-2 OR (Wuhan AND virus) OR 2019-nCoV OR (Wuhan AND pneumonia)) AND English[lang])) AND ("2019/12"[Date - Create] : "2020"[Date - Create]). From the abstracts that loaded into BioQRator without error, 10 were randomly selected and subsequently annotated for the following concepts:

- Virus_SARS-CoV-2: for terms representing SARS-CoV-2
- Virus_other: for terms representing a specific virus other than SARS-CoV-2 (e.g. MERS, SARS-CoV)
- Virus_family: for terms representing more than one virus (e.g. coronaviruses) or all viruses; for this concept unique identifiers from the NCBI taxonomy database were also indicated

- Protein: for terms representing specific proteins or genes but not a protein family (e.g. IL-2 is annotated under this concept but not protease, IL-1 family members or interleukins); for this concept unique identifiers from the UniProt database were also indicated
- Disease_COVID-19: for terms representing COVID-19
- Disease_other: for terms representing a disease other than COVID-19 (e.g. Zika virus infection) or general terms of disease (e.g. infection)
- Symptom: for terms representing disease symptoms (e.g. fever, cough)

"Pneumonia" was annotated as Disease_other or Symptom depending on the context.
Abbreviations were annotated independently, e.g. in the expression novel coronavirus (2019-nCoV) both "novel coronavirus" and "2019-nCoV" were annotated as Virus_SARS-CoV-2.
Co-references (e.g. pronouns referring to the entities) were not annotated but this is planned for future updates of the corpus.
The corpus was exported from BioQRator as csv and BioC xml file and converted to BioC json format (Supplemental_file4, Supplemental_file5, Supplemental_file6) using the script of the BioC-JSON tool from the NLM/NCBI BioNLP Research Group (https://github.com/ncbi-nlp/BioC-JSON) with minor changes (Supplemental_file7).

Results

Production of SARS-CoV-2 and COVID-19 dictionaries

In this early stage, we have focused on synonyms for SARS-CoV-2 (virus) and COVID-19 (disease). Synonyms were collected from a variety of biomedical and general public text sources. The dictionaries were then expanded by artificially generating variants. The final dictionaries contain 215 virus synonyms (Supplemental_file1) and 12906 disease synonyms (Supplemental_file2).

Production of Corona silver standard corpus

To generate the Corona silver standard (computationally annotated) corpus, 987 Pubmed records, including abstracts when available, were extracted as text file. After tokenization with spacy, the corpus was annotated in IOB2 format by comparing tokens against the SARS-CoV-2/COVID-19 terms in our dictionaries (Supplemental_file3).

Production of Corona gold standard corpus

To produce the gold standard corpus, 10 randomly selected Pubmed abstracts related to SARS-CoV-2/COVID-19 were annotated in BioQRator by a biomedical expert who labelled virus, disease, symptom and protein/gene terms. BioQRator does not allow for overlapping annotations which was problematic when annotating the expression "respiratory, enteric and systemic infections". In this case, the words respiratory and enteric were annotated as individual entities of the Disease_other category. In total, the corpus contains 155 annotations across the 6 concepts (Table1). The corpus is available in BioQRator csv, BioC xml and BioC json format (Supplemental_file4, Supplemental_file5, Supplemental_file6) and will be improved and expanded further over the next weeks.
Updated files will be published on https://github.com/Aitslab/corona.

Discussion

The corona NLP toolbox can be used in a variety of text analytics settings, e.g. to extract information from scientific and news articles or to identify social media posts related to SARS-

CoV-2/COVID-19. Even though the toolbox is designed for use with the English language, the dictionaries can also be applied for NLP in other languages because many of the terms for the virus and disease are highly similar across many languages.
Updated versions of the corpora and use cases will be published on https://github.com/Aitslab/corona and the research community is invited to comment and contribute. The tools can also be combined with other corona-related BioNLP resources which have just been published [2, 3].

Acknowledgement

Our work is performed in the AI Lund environment (http://ai.lu.se/), which connects academic and industrial artificial intelligence researchers in the Lund region.

We thank Swe-Clarin for support and the NLM/NCBI BioNLP Research Group for making their BioC-JSON tool publicly available.

We thank the following funders who support our research: the Swedish Research Council, the Swedish Research Council for Sustainable Development, the Segerfalk Foundation, the Swedish Brain Foundation, the Craafoord Foundation and the Royal Physiographic Society.

References

1. Cook HV, Jensen LJ. *Methods Mol Biol* 2019 1939:73-89, doi: 10.1007/978-1-4939-9089-4_5
2. Chen Q, Allot A, Lu Z. *Nature* 2020 579(7798):193, doi: 10.1038/d41586-020-00694-1, https://www.ncbi.nlm.nih.gov/research/coronavirus/
3. COVID-19 Open Research Dataset (CORD-19), https://pages.semanticscholar.org/coronavirus-research

Table1. Characteristics of the Corona gold standard corpus

PMID	31986264	31991541	31992388	31996494	32007643	32013309	32015508	32020836	32029004	32036774	Total
Disease_COVID-19	3		2		1		2		6		14
Disease_other	5				3	5	5	3	2	3	26
Protein	8					2		1			11
Symptom	16					7	3	2			28
Virus_family		1		3		5	3	4		3	18
Virus_other	1	1						3		4	9
Virus_SARS-CoV-2	5	4	2	7	6	1	5	5	3	10	48
Total	**38**	**6**	**4**	**10**	**10**	**20**	**18**	**18**	**11**	**20**	**155**

Supplemental files

Supplemental_file1
SARS-CoV-2 dictionary, version 1

Supplemental_file2
COVID-19 dictionary, version 1

Supplemental_file3
Corona silver standard corpus version 1

Supplemental_file4
Corona gold standard corpus version 1 in BioC xml format,

Supplemental_file5
Corona gold standard corpus version 1 in BioC json format

Supplemental_file6
Corona gold standard corpus version 1 in BioQRator csv format

Supplemental_file7
Jupyter notebook with BioC-json conversion script