CORD-19 Processing Details

COVID-19 Open Research Dataset (CORD-19) LINK to the CORD-19 dataset

Version 1 (at the time of Hackathon) :
Metamap 2018 was used. 7 semantic types were extracted.

Version 2 :

- Metamap 2020 was used to extract entities and Semrep 2016 was used to extract semantic relations from the publications.
- The annotations included many subject and object arguments incorrectly normalised to CUI in UMLS.
- Eg: Mentions of mask in a covid related paper were linked to a Gene in UMLS.
  Phrase: mask
  Meta Mapping (1000):
  1000 C1538279:MASK (ANKHD1 gene) [Gene or Genome]

  Mentions of probability values in the text incorrectly identified as medical entities:
  Phrase: (P < .001
  Meta Mapping (861):
  861 C0369773:PNOS (P Blood group antibodies) [Amino Acid, Peptide, or Protein,Immunologic Factor]

Version 3:

- NIH has made available Semrep tool results on CORD-19 dataset. We use this results to extract both the entities and relations from the publications.

  SemRep Options Used for Processing:
  semrep -A -N -n -S -F -Z 2020AA
  A: Anaphora resolution
  N: use_generic_domain_extension
  n: use_generic_domain_modification
  S: generic_processing
  F: full_fielded_output
  Z 2020AA: use 2020AA data

  We used Semrep processed CORD-19 release dated 02/15/2021. LINK to the dataset

  Number of publications processed: 97,441

Processing the Semrep Results

For Entities:

To prune the annotations for our research goal, we include biomedical entities from the publications that belong to only one or more of the following UMLS Semantic Types:

CHEM|Chemicals & Drugs|T116|Amino Acid, Peptide, or Protein|aapp
CHEM|Chemicals & Drugs|T195|Antibiotic|antb
CHEM|Chemicals & Drugs|T123|Biologically Active Substance|bacs
CHEM|Chemicals & Drugs|T103|Chemical|chem
CHEM|Chemicals & Drugs|T200|Clinical Drug|clnd
CHEM|Chemicals & Drugs|T126|Enzyme|enzy
CHEM|Chemicals & Drugs|T197|Inorganic Chemical|inch
CHEM|Chemicals & Drugs|T114|Nucleic Acid, Nucleoside, or Nucleotide|nnon
CHEM|Chemicals & Drugs|T109|Organic Chemical|orch
CHEM|Chemicals & Drugs|T121|Pharmacologic Substance|phsu
DISO|Disorders|T020|Acquired Abnormality|acab
DISO|Disorders|T049|Cell or Molecular Dysfunction|comd
DISO|Disorders|T047|Disease or Syndrome|dysn
DISO|Disorders|T037|Injury or Poisoning|inpo
DISO|Disorders|T191|Neoplastic Process|neop
DISO|Disorders|T046|Pathologic Function|patf
DISO|Disorders|T048|Mental or Behavioral Dysfunction|mobd
DISO|Disorders|T184|Sign or Symptom|sosy
GENE|Genes & Molecular Sequences|T087|Amino Acid Sequence|amas
GENE|Genes & Molecular Sequences|T028|Gene or Genome|gngm
GENE|Genes & Molecular Sequences|T088|Carbohydrate Sequence|crbs
GENE|Genes & Molecular Sequences|T085|Molecular Sequence|mosq
GENE|Genes & Molecular Sequences|T086|Nucleotide Sequence|nusq
LIVB|Living Beings|T005|Virus|virs

For Semantic Relations:

We excluded a subset of predicate types that were not helpful to understand toxicities in COVID related treatments such as PART _ OF and PROCESS _ OF. The predicate types we used are:

AFFECTS , ASSOCIATED _ WITH , AUGMENTS , CAUSES , COEXISTS _ WITH , COMPLICATES , DISRUPTS , INHIBITS , INTERACTS _ WITH , MANIFESTATION _ OF , PREDISPOSES , PREVENTS , PRODUCES , STIMULATES , USES, PRECEDES and TREATS .

We also included only those relations in which the subject or the object belongs to one or more of the semantic types mentioned above for entities.