information

MDPI

Article

Building Bio-Ontology Graphs from Data Using Logic and NLP

Theresa Gasser † and Erick Chastain *

Department of Mathematics, University of Dallas, Irving, TX 75062, USA; tgasser@udallas.edu
* Correspondence: echastain@udallas.edu
† Current address: Patrick E. Haggerty Science Center, Basement Floor, 46 1845 E Northgate Drive, Irving, TX 75062, USA.

Abstract: In this age of big data and natural language processing, to what extent can we leverage new technologies and new tools to make progress in organizing disparate biomedical data sources? Imagine a system in which one could bring together sequencing data with phenotypes, gene expression data, and clinical information all under the same conceptual heading where applicable. Bio-ontologies seek to carry this out by organizing the relations between concepts and attaching the data to their corresponding concept. However, to accomplish this, we need considerable time and human input. Instead of resorting to human input alone, we describe a novel approach to obtaining the foundation for bio-ontologies: obtaining propositions (links between concepts) from biomedical text so as to fill the ontology. The heart of our approach is applying logic rules from Aristotelian logic and natural logic to biomedical information to derive propositions so that we can have material to organize knowledge bases (ontologies) for biomedical research. We demonstrate this approach by constructing a proof-of-principle bio-ontology for COVID-19 and related diseases.

Keywords: natural language processing; natural logic; ontology; knowledge graphs; relation extraction

Citation: Gasser, T.; Chastain, E. Building Bio-Ontology Graphs from Data Using Logic and NLP. *Information* **2024**, *15*, 669. https://doi.org/10.3390/info15110669

Academic Editor: Ryutaro Ichise

Received: 27 June 2024
Revised: 15 October 2024
Accepted: 18 October 2024
Published: 25 October 2024

1. Introduction

Ours is an era marked by two major interests and practical areas of focus: big data and natural language processing. With new language-based technologies proving their usefulness in many different practical fields and businesses, perhaps one could apply natural language processing to an older application: making a knowledge base for biological data. The field of bio-ontology seeks to combine sequencing data with phenotype information and other sources of information to develop new tools for biomedical applications [1]. Bio-ontologies take biological concepts and tie them to data, thus making for a principled approach to cross-link different biomedical data-sources. Furthermore, one can make a logical inference in bio-ontologies to derive new facts from existing data sources. Bio-ontologies can also be used to answer questions about various biomedical concepts, thus being useful for practical biomedical applications. Bio-ontologies rely on the use of logical propositions to give relations between different concepts in the ontology. Logical propositions are true statements about concepts which are asserted to be related in some way.

A serious practical obstacle to the construction of vast and broadly applicable bio-ontologies is the dependence on human input to the creation of various structures. There are, however, many publicly accessible datasets which provide a great deal of text describing diseases, biological entities, conditions, and gene expression data [2,3]. It is, however, quite difficult to translate natural language into first-order logic or other formal logic in standard use for logical inference in ontologies (it is an ongoing area of research [4]). In fact, philosophers of logic and language P. F. Strawson [5] and Bertrand Russell [6] famously agreed that there was no logic in natural language. Russell preferred to work directly in first-order logic. One idea of how to overcome this is to use a logic that works in natural language [7–9], which, following the pioneering work of Montague [10–12] and Sommers [13], has become once again a thriving area of research in logic. There is an older

tradition, the Aristotelian logic tradition, which reasons in natural language [14,15]. We combine tools from both traditions and grammatical rules in order to give a new approach to obtaining propositions for an ontology from natural language data. We then build a proof-of-principle bio-ontology for COVID-19 and related diseases from these propositions. Existing work in bio-ontologies for COVID is based on two principle approaches: the first is manual entry of all relations and entities into existing ontology frameworks [16–18], and the second is a more automatic and NLP-based approach to building a COVID bio-ontology (based on named-entity recognition approaches [19]). Our approach in contrast uses a novel ontology approach based on Aristotelian logic, and an accompanying methodology for implementing inference in that logic to obtain relations and entities for a COVID bio-ontology from the text of scientific papers using dependency parsing (with minimal manual post-processing).

In this paper, we will show in the Results section how to use tools from the Aristotelian logic tradition [14,15], grammar, and natural logic [20] to derive some propositions from the CORD-19 dataset [21] of COVID-19 papers (as a proof-of-principle demonstration). In the same section, we develop logic rules which can be used to derive propositions from natural language. We also developed a new proposition tool, justified in part by the logic rules, which gathers propositions from text documents. We give some of the outputs of this tool on various biomedical Wikipedia pages. In sum, we show the usefulness of the logic rules for gathering propositions from text documents. Finally, we apply all of these tools and more to deriving new propositions to CORD-19 data from a subset of PubMed, building a proof-of-principle COVID-19 bio-ontology. Besides the novel collation of Aristotelian logic rules to the extraction of propositions for the purpose of building ontologies, in this work, we introduce a unique combination of dependency parsing to implement the logic rules and regular expressions to match Aristotelian-based logical propositions. This is the first such implementation of these Aristotelian logic rules using state-of-the-art NLP techniques. Previous implementations of all these rules in Aristotelian logic inference were carried out manually.

Before describing the Methods and Results, we will give a general overview of Aristotelian logic.

A Primer on Aristotelian and Natural Logics

For natural and Aristotelian logic, propositions of the form "Every canis lupus is a grey wolf" are called universal affirmative (type A) propositions because they affirm that the concept "grey wolf" applies universally to the species canis lupus. Propositions in general are statements which can be either true or false, rather than questions or exclamations. The general form of propositions has two terms, a subject S and a predicate P, the latter of which we affirm or deny of the former in some respect (called the quantity). So, our type-A proposition has the predicate "a grey wolf" and the subject "canis lupus". The first word in our type-A proposition determines how we affirm the predicate of the subject; in our case, we do so universally. There are three other types of propositions: "Some S is a P" (type I), "Some S is not a P" (type O), "No S is a P" (type E). Type I propositions affirm P of S in particular, that is, for some conceivable subject S, P applies to it. Besides these more abstract kinds of terms, there are also singular terms like Socrates, which are of singular quantity. We will see that when we discuss reasoning (inference rules), we can treat propositions like "Socrates is a rational animal" as if it had universal quantity [14].

Words like "Socrates" refer to a particular man living in Athens at a particular time. The way in which words refer to individuals, if at all, is called the medieval "supposition theory". Supposition determines what we can say about the terms and in what way we can say it. So, for example, in "Every dog is an animal", the term dog stands in for (supposits for) individual dogs, rather than just the universal nature (general concept) of dogs. When the term stands for individuals, we say it has personal supposition. In contrast, if we say "Canis familiaris is a species of animal", then canis familiaris stands only for the universal nature of dogs, rather than individual dogs. When only the universal

nature is supposited for, we call this "simple supposition". We must mind these two kinds of supposition; otherwise, we would have a silly consequence like "Fido is a species of animal". When the term has personal supposition and stands for "all the individual subjects" to which the word applies, we call it common universal complete supposition (universal, for the sake of brevity). One can also have a personal supposition of a term which is individual (only applying to an individual man) or more generally particular determinate (only applying to "a certain determinate few" of what it signifies). If a term T has a particular determinate supposition (which generalizes individual supposition), we give the term a set I of individuals for which it supposits, denoting it as T_I. There are many other modes of supposition than we have described here, but we are only discussing those that will be useful for our applications [14].

Once we have propositions in the kinds of forms we have above, we can reason our way to other propositions. Logical inference works by taking the propositions and combining them to infer other propositions. The simplest form of argument in Aristotelian logic is called the syllogism, which has the following form:

Q Major is Middle (Major premise)
Q Middle is Minor (Minor premise)
Therefore, Q Major is Minor (Conclusion) where Q is a quantity (Every, Some, No), and Major, Minor, and Middle are terms. The most powerful form of valid (that is, correct) syllogism is called Barbara (a Latin moniker that denotes three propositions of type A), giving a Major, Minor, and Conclusion that are all of type A. So, for example:

Every dog is an animal
Every animal is a creature
Therefore, Every dog is a creature.

2. Materials and Methods

2.1. Datasets

For the dataset we used to generate the propositions/ontology in the Results section, we primarily used the CORD-19 dataset of COVID-19-related papers assembled by the Allen Institute [21] from PubMed. CORD-19 also was used where indicated to generate the proof-of-principle results for the rules of logic in the Results. In addition, the NCBI Taxonomy database [22] was used for the ontology so as to give an access point to associated sequencing and gene-expression datasets for entries.

2.2. Dependency Analysis

We use the spaCy toolkit [23] for automatic dependency analysis, and displaCy [23] for dependency analysis carried out by hand (for the latter, the proof-of-principle results in the exposition of the rules).

2.3. Proposition Tool

We will discuss in this subsection the code and approach used for the proposition tool used in the Results. The code can be found in `biomedical_text_processing\biomedical_processing.py` in the Supplementary Files, in the zip file `biomedical_text_processing.zip`. All other code is there as well. The general flow for the approach is diagrammed in Figure 1, which visualizes the process taken to process each sentence.

The function `map_nouns_and_preps_v3` is the main output function in `biomedical_processing.py`. Given a list of sentences, the goal of this function is to break down each sentence into different propositions that contain the most important information.

A few things that happen in this function in order to obtain the output. There are three nested for-loops. Also, there is an important variable `true_subj` that allows for a "carry over" of a given subject if a newly named/specific subject is not found in a sentence. This is justified by rules DA, SA, and PC in the Results section.

The first for-loop simply goes through all of the sentences in the list of sentences that you give it. Within that first for-loop is a second for-loop that goes through the noun chunks within the current sentence. Finally, within the second for-loop is a third loop that constructs the output sentences. Importantly, due to an earlier edit, the second loop is not truly going through all of the noun chunks in a sentence, but instead is only taking the first noun chunk it sees and going from there.

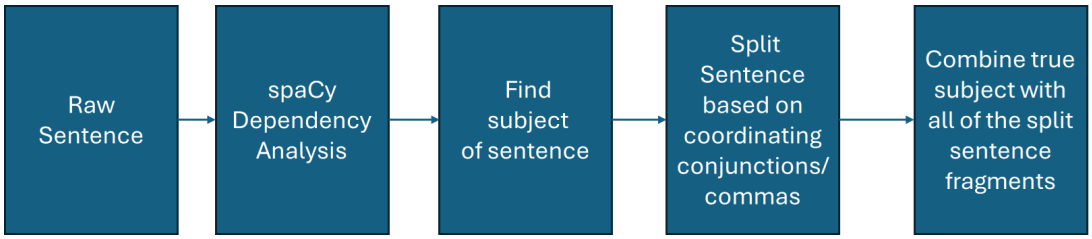

Figure 1. A flow chart which gives the pipeline for each sentence put through the proposition tool.

Besides holding the second for-loop, the first for-loop also does the preliminary processing of the string using the spaCy language model. (e.g., doc = nlp(sentence)).

The second for-loop goes through a few steps looking at the dependencies within the sentence (using the spaCy dependency parser). First, it checks whether or not the current noun chunk contains a subject for the sentence via the `is_subj` function. If it is a subject, then it is assigned to the `true_subj` variable. We then continue through the loop. After this, the function `get_all_right_children` is used to create a string of all of the words in the right subtree of the root verb (or the verb that the subject is the immediate child of). Finally, this new string is given to the `split_sent` function. This `split_sent` function "splits" up the sentence by coordinating conjunctions, commas and ending parentheses (which does assume that parentheses hold additional information). To go into some detail, given a sentence, it copies over each word into a new sentence segment until it sees one of the above-mentioned conjunctions or commas, at which point it adds the current segment to an output list and begins to build the next segment from the sentence. These steps are justified by rules CC and CD with the uncertain variant (excluding the word "possibly" in parentheses) from the Results section. The resulting list of sentence segments is then passed to the third for-loop. The third for-loop brings together the `true_subj` and the segments obtained from the `split_sent`. The word "is" was inserted into the string.

```
cur_subj =
true_subj + ‘‘is’’ + ‘‘ ’’.join([str(i) for i in cur_child_sent])
else:
cur_subj = true_subj + ‘‘is’’ + i
```

Above is given some code from biomedical_processing.py.

2.4. Keywords and Relations: Filtering and Ontology

In order to filter out many of the propositions, we narrowed down to just those that contained keywords of interest (various disease names and related words) and particular kinds of relations between them. The filtering is carried out in `biomedical_text_processing/filterData.py` and `biomedical_text_processing/findPatterns.py`.

For keywords, we chose various disease names, symptoms, and classification terms of interest based on CORD-19, Bio-GPT [24] and Wikipedia. This is the keywords list we used: ['Mycoplasma pneumoniae', 'Mycoplasma pneumonia', 'leprosy', 'MERS', 'Middle East respiratory syndrome', 'SARS-CoV-2', 'COVID-19', 'Spanish flu', 'Zika virus', 'MERS-CoV', 'SARS-CoV-1', 'Coronavirus', 'bacterial pneumonia', 'viral pneumonia', 'bacterial infection', 'SARS', 'Pneumonia', 'Influenza A virus', 'Swine Flu', 'Dengue virus', 'Dengue fever', 'Influenza virus', 'Leishmaniosis', 'viral infection', 'flu', 'parasitic disease', 'Molli-

cutes', 'Herpesviridae family', 'Equine herpes virus type 1', 'EHV-1', 'disease', 'disorder', 'infection', 'pandemic', 'symptom', 'virus', 'bacteria', 'pathogen'].

The relations (copulae) we are interested in are based on the Hearst patterns [25] and Aristotle to find hypernymy (relations between A and B where A is a kind of B), synonymy, and causation for ontology. From Aristotelian considerations on genus, species, and causation, we have the following: "A defined as B", "A classified as B", "A also known as B" (for synonymy), "A caused by B", "A causes B", From the literature on patterns known to be useful for building ontologies in the natural language processing community [25], we have: "A is B", "A part of B", "A member of B", "A which is called B", "A which is (example | class | kind) of B". Based on these patterns, combined with the keywords, we filtered all of the propositions to those just containing at least two keywords and the relation. Then, to make the ontology, we needed a final list of relations. Besides the relations given above, we added some others which were based on what was in the propositions selected from the dataset post-sampling: ['caused by', 'defined as', 'classified as', 'also known as', 'causes', 'initiated', 'is', 'part of', 'such as', 'member of', 'which is called', 'which is an example of', 'which is a class of', 'which is a kind of', 'grouped', 'has', 'of family', 'indicates', 'complication of', 'involves', 'family of', 'belongs to'].

2.5. Sub-Sampling of Short Propositions

The proposition-generating process was based on finding a subject and then splitting up the remaining clauses by branching on coordinating conjunctions. It stands to reason that this could fail to produce the information content of a sentence in the fragments so derived. This insight can be formed into a probabilistic model.

If the probability of a bad split per word in the split sentence is ρ, then if one tries to figure out the probability of a good split with k words, it ends up being $(1-\rho)^k$. The probability of a good split decreases very quickly with increasing k (the number of words). Therefore, if you want a very high probability of a good split, you must choose split sentences with a small number of words.

The assumption that the accuracy decreases in k, and does so in an exponentially decreasing manner is supported by the following experiment on the outputs of the proposition tool. We ran an experiment to characterize errors in gathering propositions using dependency parsing. The output of the proposition tool then is grouped according to length, from 4 words of output to 11 words (as there were only empty outputs with less than 4 words). We took a random set of 45 words sampled without replacement from each group. The number of outputs for each length group is 46 for 4 words, 175 for 5 words, 221 for 6 words, 361 for 7 words, 477 for 8 words, 467 for 9 words, 631 for 10 words, and for 11 words. Out of the 45 sampled outputs, we only kept those which were distinctively informative and which made sense. We did not count something as distinctively informative when it just had generic pronouns like "the virus" or "two patients" as the subject. We also excluded overly vague propositions or those which looked nonsensical or unlike a proposition (not having a subject, verb, or object form in some broad sense). The number of propositions which we kept as valid outputs for each group is nvo = [41, 32, 38, 27, 23, 29, 32, 26] (for the groups k = [4, 5, 6, 7, 8, 9, 10, 11]). The accuracy was estimated by the nvo list divided by 45 (the sample size). Least-squares fits for the accuracy as a function of k were found of exponential, quadratic, cubic, logarithmic, and linear models using WolframAlpha [26]. The best fitting models had equations $1.03678e^{(-0.0556514k)}$, $0.488095k^2 - 8.94048k + 68.0357$, $-0.0757576k^3 + 2.19264k^2 - 21.0238k + 94.7403$, $54.9423 - 12.1917log(k)$, $0.95873 - 0.0359788k$. All of these fitted functions have an accuracy decreasing in k. Furthermore, the best fitting model (in terms of R^2 value) is the exponential fit, which had $R^2 = 0.982$ (in contrast with a R^2 of 0.577, 0.591, 0.492, and 0.423 for the quadratic, cubic, logarithmic, and linear model fits).

As such, we only use the outputted (split) propositions that are very short. This was implemented in `biomedical_text_processing/sample_props.py`, which samples all propositions less than or equal to 7 in length.

2.6. How the Final Proposition Output Was Made

In order to prepare the final propositions that were given as input to the ontology builder to make them of better quality and more machine-readable, we carry out the following:

First, we clean up miscellaneous grammatical and words that are not of interest. For example, words like "the", punctuation or other symbols, and those which seem unrelated to the true subject or to the true predicate term. For instance, we remove the strange abbreviation "DCs" in the proposition "DCs MERS-CoV causes infection".

Second, we fill in missing context/words based on what makes sense or correct misspelled words. For example, "MERS is disease Qatar" becomes "MERS is disease from Qatar".

Third, we split conjunctions using rule CC. For example, "Chikungunya virus is arbovirus arthritis" becomes the two propositions "Chikungunya virus is arbovirus" and "Chikungunya virus causes arthritis" (incorporating the correct word for the relation).

Fourth, we normalize the order after the relations "is" and "causes" of terms and their modifiers. In particular, the order within the predicate was standardized to "term modifier". For example, "COVID-19 is disease infectious". If the term was actually the "modifier term" itself, then that order was retained by putting a dash in between the two in the target ordering. For example, "COVID-19 is respiratory-disease".

Fifth, we reorder the predicate so that all of the words are combined in the right order (combined with a "-" as before).

2.7. Obtaining the Subject and Predicate for the Ontology

We can use regular expressions to obtain the subject and predicate for the ontology by taking a relation from the list of relations and applying the following pattern: (.*) relation (.*). We gather the two terms subject and predicate from the two capture groups. Then, we use the NLTK [27] word tokenizer on the predicate to give the predicate term as follows:

1. if "is" or "causes" are the relation, then make the first word of the predicate the entry for the ontology, and the following words be modifiers for the relation. For example, "Equine herpesvirus causes infection perinatal foal" takes the entry to be infection and the relation to be "causes perinatal foal".
2. If other relations are there, the entry is the entire predicate.

These are implemented in `biomedical_text_processing/buildOntology.py`.

2.8. Estimating the Accuracy of the Infectious Disease Ontology

Accuracy was based on the outcome of the relationship outlined actually happening rather than being exclusively true (so for instance, if there are two chairs in a room, then it is also true that there is one chair in the room). The estimate given is an underestimation of the accuracy, as there were some cases that were uncertain (which we counted as an error). The sources used for verifying the accuracy of the bio-ontology entries are WHO, StatPearls, CDC, Merck Veterinary manual, Nature group, NCBI Bookshelf, and various scholarly publications.

3. Results

In order to answer queries about biomedical subjects and phenotypes and their associated gene sequences, one can hope to combine massive natural language datasets like PubMed with information from various freely available sources online (like Gene Expression Omnibus [3] or Wikipedia). One can then use this approach to make a data-driven bio-ontology, which can link biological concepts and phenotypes with sequence data [1]. We show later in this section how to build a bio-ontology for infectious diseases based on a subset of PubMed (CORD-19, containing papers on COVID-19 and related diseases), linking entities with their NCBI Taxonomy IDs [22]. The latter allows easy lookup in the Gene Expression Omnibus or the NCBI Nucleotide [28] database.

Bio-ontologies, as we have seen, are an approach to logically analyzing queries about various biological concepts and associated phenotypes/sequences. One of the primary challenges of working with ontologies (including bio-ontologies) is the difficulty of building a sufficiently general database, since usually ontologies need to be generated by programmers working in tandem with experts to create the various propositions about concepts. Our novel approach towards making a more general bio-ontology is to use a data-driven approach that derives propositions in the ontology from various freely available natural language datasets and then uses logical inference to infer connections between the various concepts described in the ontology. The results outlined in this section develop a set of tools and techniques that can be used to derive propositions from natural language data. These tools and techniques will be used, for instance, in building a bio-ontology.

One of the primary difficulties with using logical inference on natural language datasets is the difficulty of translating English sentences into logical propositions—a prerequisite for obtaining concepts in an ontology. The difficulty can stem from apparent disparities between the form of logical propositions in formal logics, which come in the form of mathematical formulas using operators and the form of English sentences. Our approach to reducing the difficulty of translation into propositions is to make use of natural logic (generalized quantifiers [20]) and older forms of formal logic—medieval [15] and modern forms of Aristotelian logic [14]. The forms of logic we will apply and combine here operate in natural language primarily, not using the kinds of mathematical formalism more common in mathematical logic. The opportunity with such approaches is that we have no need to translate a sentence like "Every canis lupus is a grey wolf" into a formula in order to make it into a proposition. We merely recognize it as a proposition of the forms of logic we combine here. Even though English sentences are rarely in such a nice form, one can also distill the deep structure of the sentence into several propositions of that form [29]. We will show methods using logic and natural language processing (NLP) to obtain the deep structure of sentences.

In this section, as a whole, we will show a proof-of-principle set of results that verifies the value of the new approach after giving some background information on the logic. The proof-of-principle results illustrate different instances of our approach used to both obtain logical propositions about epidemiology from natural language data (including CORD-19 [21], Wikipedia) and infer new propositions as well.

There are other forms of inference rules we will describe as needed in the following sections, but we will focus on particular rules and gleanings from the natural and Aristotelian logic literature that we will apply to real sentences from datasets to give us results. Finally, we will then combine all of the rules and other natural language processing (NLP) tools to give us the rest of the results.

3.1. Particular Rules

First, we will cover particular rules from the Aristotelian logic and grammar literature that we will find of use, along with some proof-of-principle results showing that these rules can help in building bio-ontologies. All of the rules are visualized and summarized in Figure 2.

3.1.1. Anaphora

In many sentences of the English language, including those about biological entities, there are many uses of pronouns to stand in for the subject of a sentence. Anaphora in linguistics refers to the use of pronouns to stand in for the subject of a sentence. For example:

Socrates sees his donkey.

In this case, we have the pronoun "his" standing in for "Socrates". In order to derive propositions for a bio-ontology we should find some way to handle anaphora. It turns out that the Aristotelian logic tradition has a way of making the kinds of substitutions one needs to resolve anaphora. We call such a pronoun a relative of identity. Terence Parsons [15] formalizes the rule as follows (based on Peter of Spain and other sources):

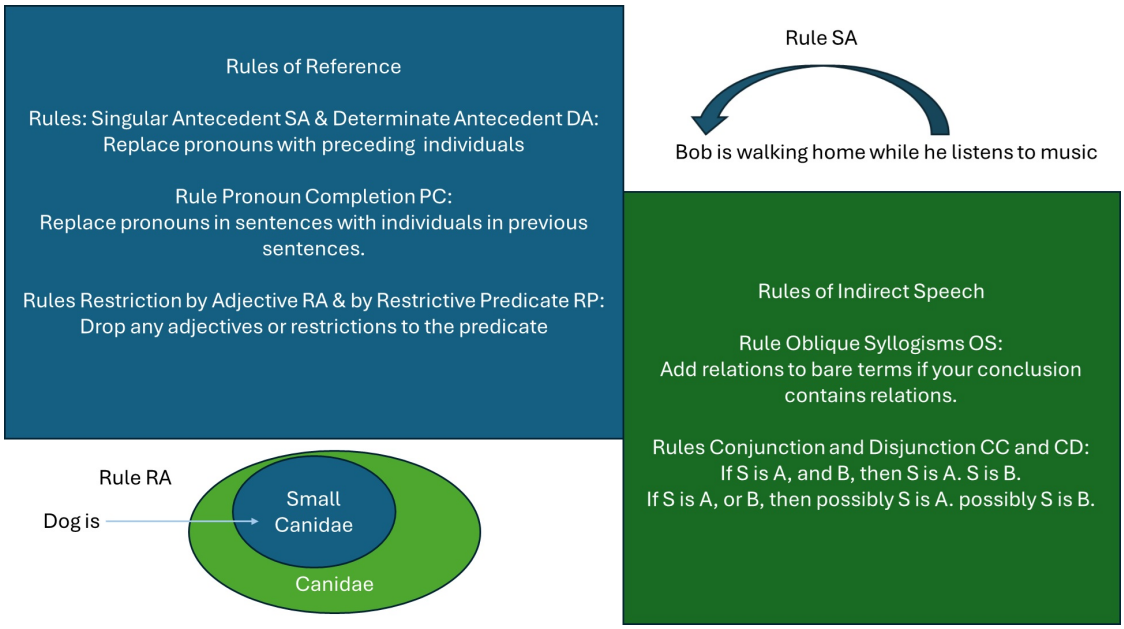

Figure 2. A diagram summarizing all the logic rules presented in this section.

Rule SA (Singular Antecedents):

A proposition containing a relative of identity with a singular antecedent is equivalent to the proposition that results from replacing that relative with its antecedent.

Of course Rule SA can be applied to pronouns like "It" as well. A counterexample is given by Walter Burley [15] to the rule's application in the case of non-singular terms without particular determinate supposition. The example is "A man runs and he argues", which is not equivalent to "A man runs and a man argues". The latter proposition is more general, since we can imagine it to be true when one man runs and another argues, rather than just when the same man does. So, if we have non-singular terms, they must have a particular determinate supposition in order to use the above rule (in our case, the particular man who supposits for "a man" must be the same man as in the antecedent). This results in the following variant of Rule SA.

Rule DA (Particular Determinate Antecedents):

A proposition containing a relative of identity that has an antecedent with a particular determinate supposition is equivalent to the proposition that results from the substitution of the relative with its antecedent (with the same supposition as the antecedent).

Of course, the new proposition will have to have a mark to denote the set of individuals I which goes with the particular determinate supposition. Let $\cdot$ mean indefinite quantification, e.g., "a man". Using this notation, for instance, the marking of the particular determinate supposition would be "$[\cdot man]_{\{m\}}$ runs and he argues", where $\cdot man$ stands for a particular individual man m so $I = \{m\}$. Using rule DA thus results in "$[\cdot man]_{\{m\}}$ runs and $[\cdot man]_{\{m\}}$ argues". The second proposition unambiguously identifies that m is the one arguing, rather than some other man; thus, Burley's critique of this case does not apply. Importantly, in our use of DA, if there is no notation given, it is assumed that the same set is given for the substituted term as in the original term.

An example for SA: The proposition "Bob walked home and he listened to music" is equivalent to "Bob walked home and Bob listened to music" by Rule SA (since Bob is a singular term).

An example for DA: The proposition "Some professor taught his morning lecture and he ate lunch afterwards" is equivalent by rule DA to "Some professor taught his morning lecture and the same professor ate lunch afterwards" (since some professor refers to a particular professor).

Finally, we can apply these rules across multiple sentences by taking sentences with a pronoun as a subject to be incomplete propositions. In order to make sentences with pronouns into a proposition, we use the following rule based on ordinary English grammar [30].

Rule PC (pronoun completion):

To make a sentence containing a pronoun (an incomplete proposition) into a proposition, combine the sentence with all previous incomplete propositions or propositions until the resulting sentence is a proposition (e.g., has a non-pronoun subject).

Example: "Bob decided to teach his class in the most exciting way possible. He brought in a circuit board and did some demos" is equivalent by rule PC to "Bob decided to teach his class in the most exciting way possible. Bob brought in a circuit board and did some demos".

We can now give a derived result using these rules. Consider the following sentence from [31]:

"Compared with other pathogens, M pneumoniae is atypical in many ways: it is one of the smallest self-replicating organisms, has a reduced and highly stable genome (0.8 Mbp), lacks a cell wall, grows slowly (generation time 6 h), requires close contact for transmission, and has a distinct disease presentation (atypical pneumonia), the pathogenesis of which might involve host cell-mediated immunity".

To identify the antecedent for the pronoun "it", we find the subject of the first part of the sentence (preceding "it") by using spaCy's dependency parser with noun chunks. We find that "M Pneumoniae" is the antecedent. As it is a common noun, it is not singular, and so we must use rule DA (assuming there are determinate particular entities referred to by the term M Pneumoniae in the context, which is easy to see).

Using rule DA, we obtain the following: Compared with other pathogens, M pneumoniae is atypical in many ways: M pneumoniae is one of the smallest self-replicating organisms, has a reduced and highly stable genome (0.8 Mbp), lacks a cell wall, grows slowly (generation time 6 h), requires close contact for transmission, and has a distinct disease presentation (atypical pneumonia), the pathogenesis of which might involve host cell-mediated immunity.

The rule PC will be used as part of our general strategy for getting subjects to carry over to following sentences that start with pronouns in Wikipedia articles on diseases. We will cover this in the final section combining all of the rules to obtain propositions from articles.

3.1.2. Generalized Quantifiers, Monotonicity and Restriction

In many sentences about biological entities there are many clauses and deeply nested prepositional phrases and adjectival phrases. In order to derive propositions about biological entities from these very dense sentences, we need a tool from logic that can handle very sophisticated predicates and break them down into components.

For example, we can take the sentence "Every dog is a small canidae". The predicate here contains an adjective phrase due to the adjective small. The medieval theory of supposition allows us to say that if we have an adjectival phrase which reduces the number of individuals that the predicate applies to, this is called a restriction of supposition [14]. We introduce a rule based on restriction:

Rule RA (Restriction by Adjective):

From a proposition which has a predicate including an adjective phrase, one can derive a proposition without the adjectives in that adjective phrase.

For example, using rule RA, we can derive "Every dog is a canidae" from "Every dog is a small canidae".

Another example: From "Every bad day comes to a good end" we can derive "Every bad day comes to an end" using rule DA.

We can further generalize this rule using the theory of generalized quantifiers [20] as follows:

Rule RP (Restriction by Restrictive Predicate):

Let P be the predicate of a universal proposition T, with $S(P)$ being the set of individuals that satisfy T. Then we can derive from T any proposition which shares the same subject as T, but has a predicate P' such that $S(P) \subseteq S(P')$. Note that for this rule to be valid, only predicates which restrict the supposition of the term in the following ways are possible: (1) compound sentences with coordinating conjunctions, (2) prepositional phrases like "...which", (3) modifiers or modifying clauses.

Note that the above rule is called monotonicity of the "Every" quantifier. Note that even though the theory of generalized quantifiers that justifies this rule does not depend on grammar and the theory of supposition, we can also derive the same rule from the theory of supposition. To see this, we note that one can speak of restriction of supposition as with the case of the adjective phrase, but more generally discuss all of the possible individuals for which a term supposits (stands in for) as being reduced. If one could carry this out using grammatical forms, then this would take away some of the artificiality of the pure set theory version of the generalized quantifier. For example, this kind of restriction of supposition can be obtained by every condition added on after by a prepositional phrase like "...which does Y" or various kinds of modifiers or modifying clauses.

Now we will use rule RP to obtain some results on the following sentence from [21]:

"Jena virus, a bovine norovirus, is a member of the Caliciviridae family of positive sense RNA viruses and was first isolated from the diarrhoeic stools of newborn calves."

Now to use rule RP, note that we can take the predicate to be more restricted by having the additional condition "was first isolated from the diarrhoeic stools of newborn calves". We also assume that it is implicitly a universal affirmative proposition. Therefore, we can use rule RP to derive the following proposition (confirming that it is a compound sentence with a coordinating conjunction for its validity):

Jena virus, a bovine norovirus, is a member of the Caliciviridae family of positive sense RNA viruses.

We can apply rule RP again on this sentence:

Chronic obstructive pulmonary disease (COPD) is a respiratory disease characterized by an airflow limitation and inflammation of the lower airways.

Again we assume this is implicitly a universal affirmative proposition, and note that "characterized by an airflow limitation and inflammation of the lower airways" restricts the set of things satisfying the predicate. That is also a kind of prepositional phrase. So, we may use rule RP validly to obtain the following:

Chronic obstructive pulmonary disease (COPD) is a respiratory disease.

And now using rule RA, (as respiratory is an adjective) we obtain the following:

Chronic obstructive pulmonary disease (COPD) is a disease.

Another example: Chronic stress is a cause of many cardiac diseases which involve a heightened heart rate. Using Rule RP, one can derive from this proposition the one that follows: Chronic stress is a cause of many diseases.

3.1.3. Oblique Propositions and Syllogisms

On occasion, in the description of different biological conditions or entities, we see a mode of speech in which we do not simply ascribe some kind of concept or term to the subject. In fact, sometimes, it becomes sufficiently complex that we have propositions of the form S is R of P, where R is the word indicating a kind of relation that S has to P. We could also have S is R to P, S is R by P, or other common prepositional phrases involving P; or, most generally, we can use any set of words indicating that S is related to P via relation R. All of these are propositions that are involved in oblique syllogisms [14]. We will call such propositions oblique propositions of the form S is $Obl_{R,PP}$ P, where $Obl_{R,PP}$ gives the translation of R with the preposition PP into the appropriate form. For example, *James* is $Obl_{Father,of}$ *Vanessa* is translated into *James* is Father of *Vanessa*. Now, we are ready to state the classic rule for oblique syllogisms, as elucidated by [14]:

Rule OS (Oblique Syllogism):

The following syllogism:

S is $Obl_{R,PP}$ P
P is Q
Therefore, S is $Obl_{R,PP}$ Q

Is equivalent to the syllogism:

S is $Obl_{R,PP}$ P
$Obl_{R,PP}$ P is $Obl_{R,PP}$ Q
Therefore, S is $Obl_{R,PP}$ Q

We can see immediately the usefulness of this rule when we are trying to characterize biological conditions which have relations to biological entities. Here is a result that exactly raises this kind of scenario (sentences in the syllogism from the Mycoplasma Pneumonia and Mycoplasma Pneumoniae Wikipedia pages [32,33]). In the following, we want to show that this syllogism holds, so that we may derive the conclusion:

M. Pneumonia is a form of bacterial infection caused by M. Pneumoniae
M. Pneumoniae is a very small bacteria in the class Mollicutes
Therefore, M. Pneumonia is a form of bacterial infection caused by a very small bacteria in the class Mollicutes.

To carry this out, we use rule OS which gives us that the above syllogism is equivalent to the syllogism

M. Pneumonia is a form of bacterial infection caused by M. Pneumoniae
A form of bacterial infection caused by M. Pneumoniae is a form of bacterial infection caused by a very small bacteria in the class Mollicutes
Therefore, M. Pneumonia is a form of bacterial infection caused by a very small bacteria in the class Mollicutes.

Which is a valid syllogism by Barbara (assuming these are all universal affirmative propositions).

Another example:

Sue is the mother of Bob
Bob is the department chair
Sue is the mother of the department chair

by rule OS is equivalent to:

Sue is the mother of Bob
The mother of Bob is the mother of the department chair
Sue is the mother of the department chair

3.1.4. Coordinating Conjunctions

In real English sentences, especially in scientific literature, one finds long sentences with many adjectives separated by commas using a coordinating conjunction. For example, we can have the sentence "A dog is a warm-blooded, furry, and terrestrial animal". The coordinating conjunction is "and" [30]. One can also have a comma followed by a coordinating conjunction when combining two independent clauses. Now, if we focus only on the coordinating conjunctions "and" and "or", there is evidence that such sentences correspond to their logical meaning [34]. What does "and" mean logically? If we have the propositions P and Q, then from the sentence "P and Q", one can derive that both P and Q are true [14,15]. What does "or" mean logically? If we have the propositions P and Q, then from the sentence "P or Q", it is true that either P is true, that Q is true, or both [15]. Inspired by these ideas, we have a rule to derive sentences from sentences with coordinating conjunctions and commas:

Rule CC (Commas and Conjunctions):

The sentence "S is P_1,..., and P_2" is equivalent to the true propositions "S is P_1", ..., "S is P_1"

An example: The dog is small, loud, and adorable. Using rule CC this is equivalent to the following propositions: The dog is small. The dog is loud. The dog is adorable.

Rule CD (Commas and Disjunctions):

The sentence "S is P_1,..., or P_2" is equivalent to some subset of the propositions "S is P_1", ..., "S is P_1" being true (though all of them could be true, as well).

Note that even though rule CD gives us only some of them as true, we will in practice allow for all of them to be true as a possibility, and so when it is unknown which subset of them is true we will assume all of them to be true (for example, in our process to resolve text into propositions). So this uncertainty means we will insert a "(possibly)" before the predicate. This we call the "Uncertain variant" of rule CD.

An example: My students will be stopping by my office today or tomorrow. Using rule CD, we can conclude: My students will be stopping by my office (possibly) today. My students will be stopping by my office (possibly) tomorrow.

Let us apply the uncertain variant of the rule to an actual sentence from the CORD-19 dataset to obtain a proof-of-principle result:

Clinical manifestations of influenza infections range from illness with asymptomatic, atypical (i.e., gastro-intestinal), or oligosymptomatic disease to severe toxic progression resulting in death.

Now, from this, we can derive, using the uncertain variant of rule CD, the following:

- Clinical manifestations of influenza infections range from illness with (possibly) asymptomatic disease to severe toxic progression resulting in death.
- Clinical manifestations of influenza infections range from illness with (possibly) atypical (i.e., gastro-intestinal) disease to severe toxic progression resulting in death.
- Clinical manifestations of influenza infections range from illness with (possibly) oligosymptomatic disease to severe toxic progression resulting in death.

3.1.5. Gathering Propositions from Text Data

We developed a tool to automatically generate propositions from text. The details of how it works are given in the Methods section. The approach used principally is justified by the rules CD with its uncertain variant (omitting the possibly), CC, PC, DA, and SA. In particular, the proposition tool breaks sentences with commas and the coordinating conjunctions "or"/"and" into several propositions with the same subject as the first clause with commas. We further post-processed the propositions by filtering them for keywords and relations relevant to logic and building ontologies. Finally, as described in the Materials and Methods (Section 2.5), the proposition tool produces more erroneous results for longer outputs and short outputs were sub-sampled and were not large in number (especially compared to the original set of all propositions).

Propositions promising for inclusion in the ontology were then identified for the ontology based on accuracy, interest, and size considerations. No probability considerations were involved at this stage (besides the size). Instead, propositions were chosen to be more accurate based on general background knowledge and general knowledge resources like Wikipedia or results of google searches. Propositions were chosen for interest when they were relevant to the concerns of infectious disease researchers, including entities that are of interest (such as COVID-19). The size considerations include both the length of the proposition (as discussed before) and the number of results included in the ontology. Finally, a revised version of the propositions was prepared in machine-readable format by hand, providing missing context and using methods that can in part be justified using the above rules. More details are provided in the Methods section. The workflow for preparing the final propositions is given in Figure 3.

3.1.6. Building an Infectious Diseases Ontology

Using the first 10 percent of the papers from the CORD-19 dataset, built from the COVID-19 papers in PubMed, we gathered final propositions (using our process above) about infectious diseases. The code for this is given in `biomedical_text_processing/`

`getCordData.py`. Some examples of the infectious disease proposition outputs in machine-readable format are given in Figure 4. The full set of propositions is given in `biomedical_text_processing/triplets/finalPropsOntoIn.txt`, and the original output of the computer code before manual adjustment is in `biomedical_text_processing/triplets/finalProps.txt`.

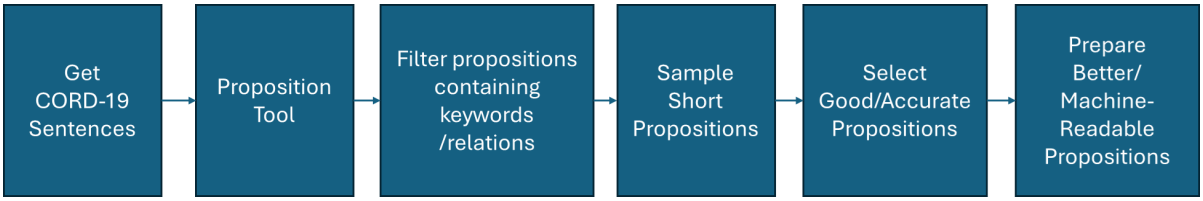

Figure 3. The final proposition generation workflow pipeline from the CORD-19 dataset to final proposition.

SARS-CoV-2 member of Coronavirus
Coronavirus disease is pandemic
Equine Coronavirus classified as betacoronavirus
Coronavirus disease is infection
SARS-CoV-2 is virus
Coronavirus is pathogen
Coronavirus disease is pandemic
Coronavirus is virus RNA
Coronavirus disease is disease contagious
Coronavirus is virus enveloped
Coronavirus disease initiated pandemic
Coronavirus disease is disease
SARS-CoV-2 is Coronavirus
rhinovirus is virus
rhinovirus is picornavirus
MERS is disease from Qatar
SARS-Cov-2 causes death
Sialodacryoadenitis virus is Coronavirus
Coronavirus disease is disease
rhinovirus is virus
SARS-CoV is virus enveloped

Figure 4. Examples of final propositions derived from the CORD-19 dataset.

Using the propositions, we made a list of organisms (either species, genus, or family). The resulting organism list is ['SARS-CoV-2', 'Equine coronavirus', 'picornavirus', 'MERS-CoV', 'Sialodacryoadenitis virus', 'SARS-CoV', 'Zika virus', 'Dengue virus', 'EHV-1', 'Sendai virus', 'Cytomegalovirus', 'Respiratory syncytial virus', 'Mimivirus', 'murine coronavirus', 'Sialodacryoadenitis virus', 'West Nile virus', 'Chikungunya virus', 'Influenza A virus', 'Japanese encephalitis virus', 'Equine herpesvirus sp.', 'Feline calicivirus', 'Parainfluenza virus'].

The list of organisms was put through the BioPython package [35] combined with Entrez to obtain the NCBI Taxonomy [22] database ID for each element. This is accomplished by `biomedical_text_processing/taxIDFinder.py`. The Taxonomy database ID is cross-linked with the NCBI databases Nucleotide [28] (a source of DNA sequences), GEO (a source of gene-expression data), and more, so that users of the ontology can gather data from those sources for the organisms. Using the Taxonomy ID and BioPython/Entrez, a user of the database can directly gather relevant sequence data for a given organism.

The final ontology is given by the output of an ontology-builder program, which takes the propositions and runs a simple regular-expression-based tool to gather the subject, predicate, and relation. The relation is taken from a list of known relations, some derived based on Aristotelian logic, others based on useful terms from the natural language processing community and some based on the data (see Methods). The entries in the ontology are automatically gathered from the propositions by taking the subject and some subset of the predicate (see Methods for details).

All entries in the ontology are made nodes in a NetworkX [36]-directed multi-graph, with each multi-edge being a possible relation between entries. A directed multi-graph was used so as to indicate both the direction of the relation (especially important in relations like "family of") and also to allow there to be multiple relations between the same two entries. For example, the machine-formatted proposition "SARS-CoV-2 is Coronavirus

novel" gives entries SARS-CoV-2 and Coronavirus, made into nodes, and the multi-edge SARS-CoV-2 -"is novel"-> Coronavirus contains an annotation saying "is novel" to indicate the relation between the two (read left to right). One could also add another relation between SARS-CoV-2 and Coronavirus by using another multi-edge.

Each node also contains a NCBI Taxonomy ID (under dictionary key "xlabel") if it was previously gathered using BioPython. The final result is converted into a dot file and visualized using force-directed placement (fdp) in GraphViz [37] in Figures 5–11. Note that in these figures, the Taxonomy ID of an entry is drawn near the node (if it is in the ontology). The ontology is built in `biomedical_text_processing/buildOntology.py` and can be modified or accessed in that file as well.

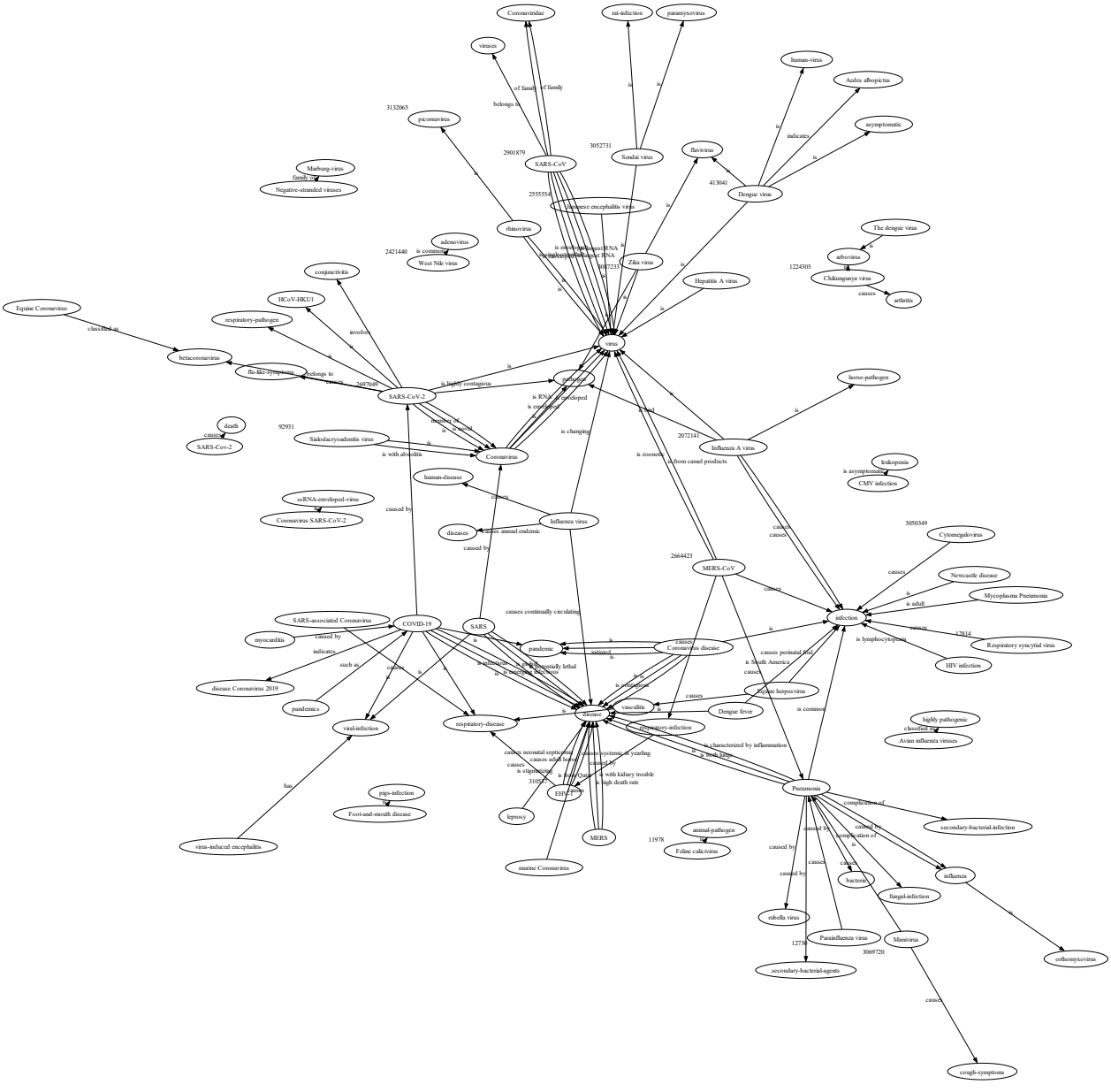

Figure 5. Visualization of the complete multi-graph for the infectious disease ontology. Note that detail panels from this figure are given in Figures 6–11. The entities in this Figure are SARS-CoV-2; Coronavirus; Coronavirus disease; pandemic; Equine Coronavirus; betacoronavirus; infection; virus; pathogen; disease; rhinovirus; picornavirus; MERS; SARS-Cov-2; death; Sialodacryoadenitis virus; SARS-associated Coronavirus; respiratory-disease; SARS-CoV; COVID-19; viral-infection; respiratory-pathogen; Zika virus; flavivirus; Dengue virus; asymptomatic; Dengue fever; EHV-1; Pneumonia; influenza; CMV infection; leukopenia; Sendai virus; leprosy; MERS-CoV; Cytomegalovirus; Respiratory

syncytial virus; bacteria; fungal-infection; Mimivirus; SARS; viruses; Newcastle disease; virus-induced encephalitis; HIV infection; Coronaviridae; murine Coronavirus; disease Coronavirus 2019; West Nile virus; adenovirus; myocarditis; pandemics; Chikungunya virus; arbovirus; arthritis; respiratory-infection; Influenza virus; diseases; Influenza A virus; human-disease; Hepatitis A virus; Mycoplasma Pneumonia; rubella virus; cough-symptoms; conjunctivitis; Foot-and-mouth disease; pigs-infection; secondary-bacterial-infection; Coronavirus SARS-CoV-2; ssRNA-enveloped-virus; human-virus; Aedes albopictus; Japanese encephalitis virus; The dengue virus; Equine herpesvirus; vasculitis; flu-like-symptoms; Avian influenza viruses; highly pathogenic; orthomyxovirus; paramyxovirus; rat-infection; Parainfluenza virus; horse-pathogen; Feline calicivirus; animal-pathogen; secondary-bacterial-agents; Negative-stranded viruses; Marburg-virus; HCoV-HKU1.

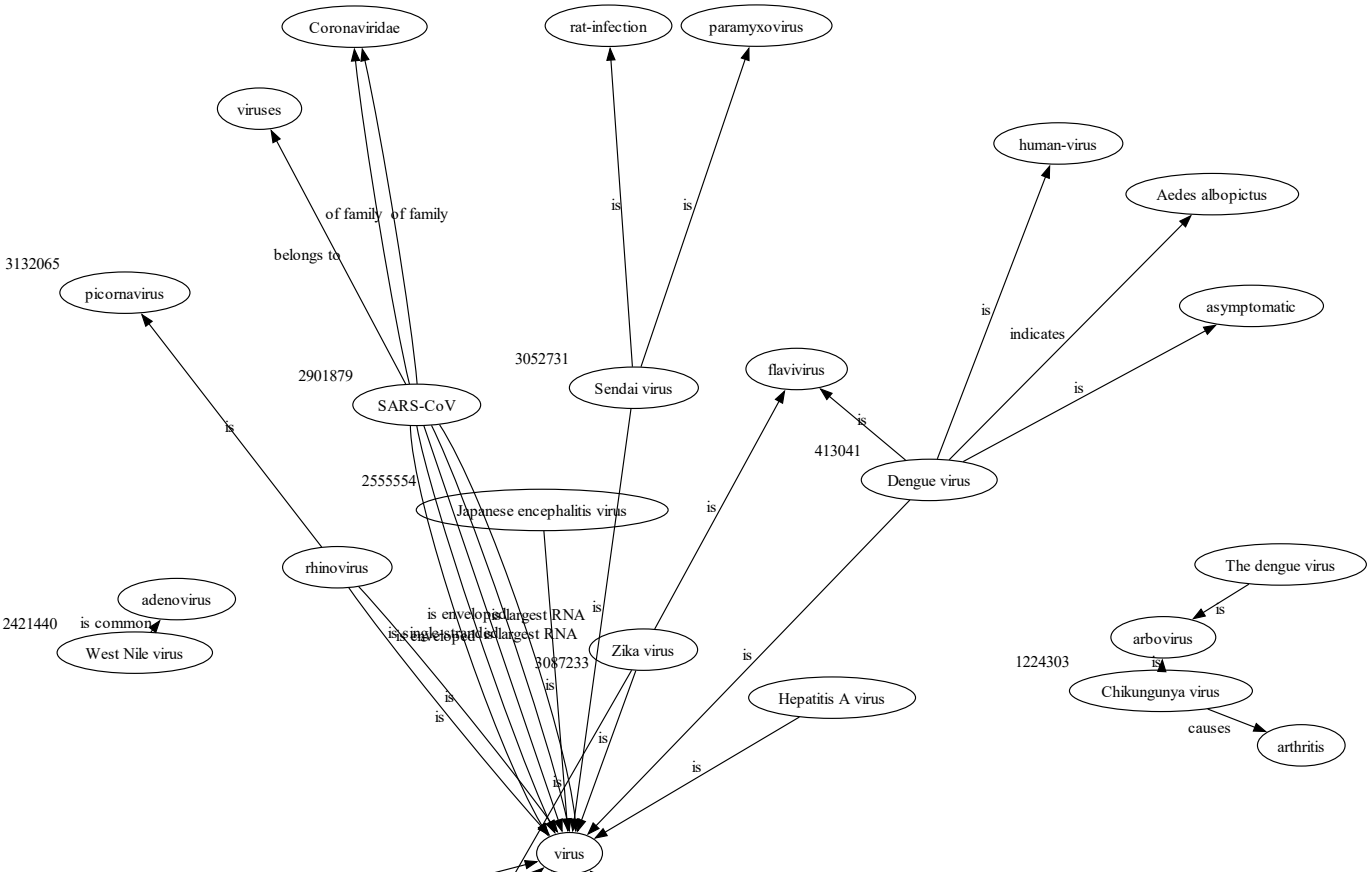

Figure 6. Visualization of multi-graph for the infectious disease ontology panel one. Edge labels obscured in the arrow from SARS-CoV to virus say "is largest RNA", "is single-stranded", "is enveloped".

As outlined in the Materials and Methods, we estimated the accuracy of the resulting bio-ontology based on manually scoring the discovered entities and relationships. The accuracy of the bio-ontology is 117/125 (93.6 percent) with an error rate of 8/125 (6.4 percent). We found the accuracy of our ontology, achieving an accuracy of 93.6 percent (error rate 6.4 percent).

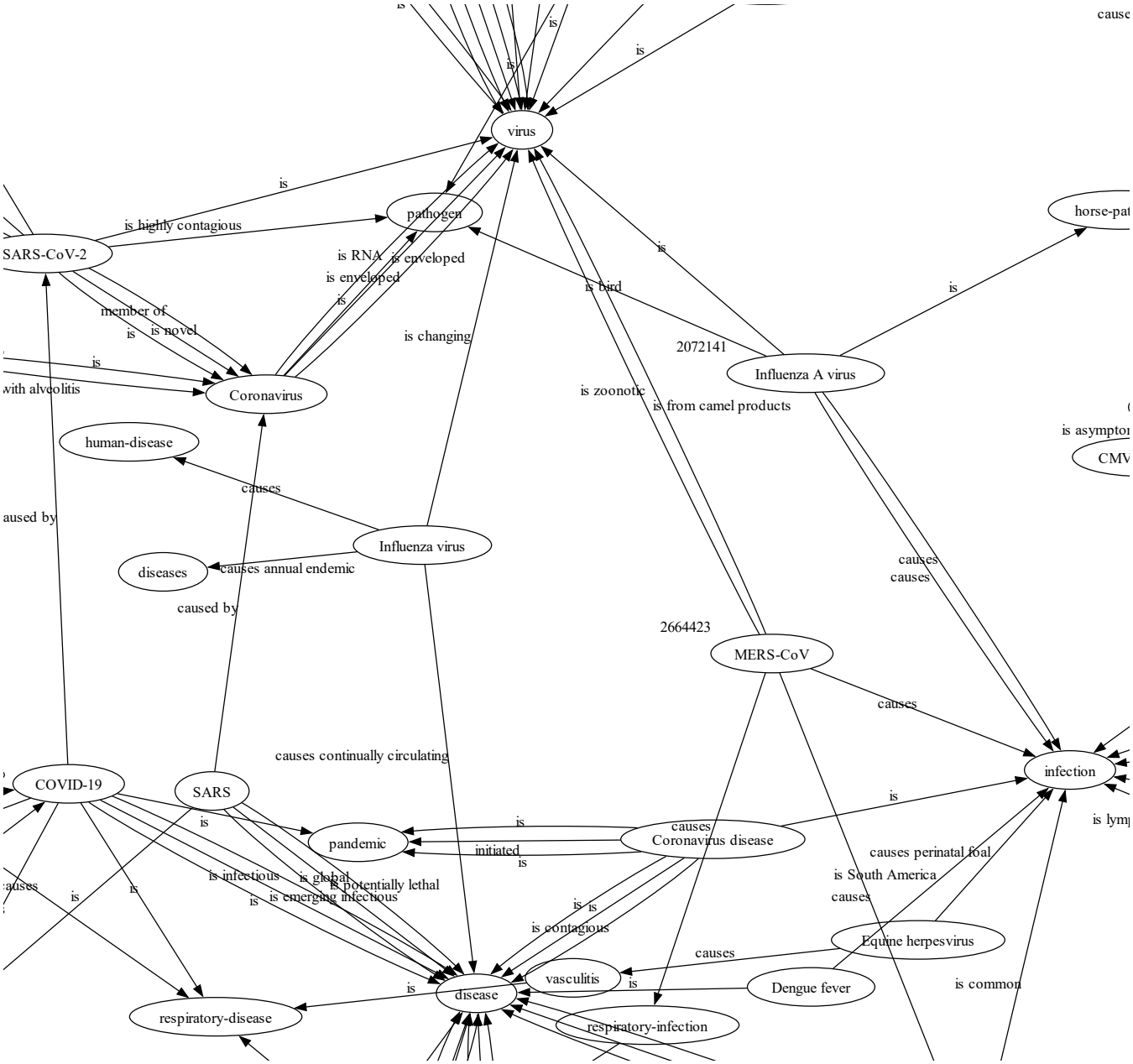

Figure 7. Visualization of multi-graph for the infectious disease ontology panel two.

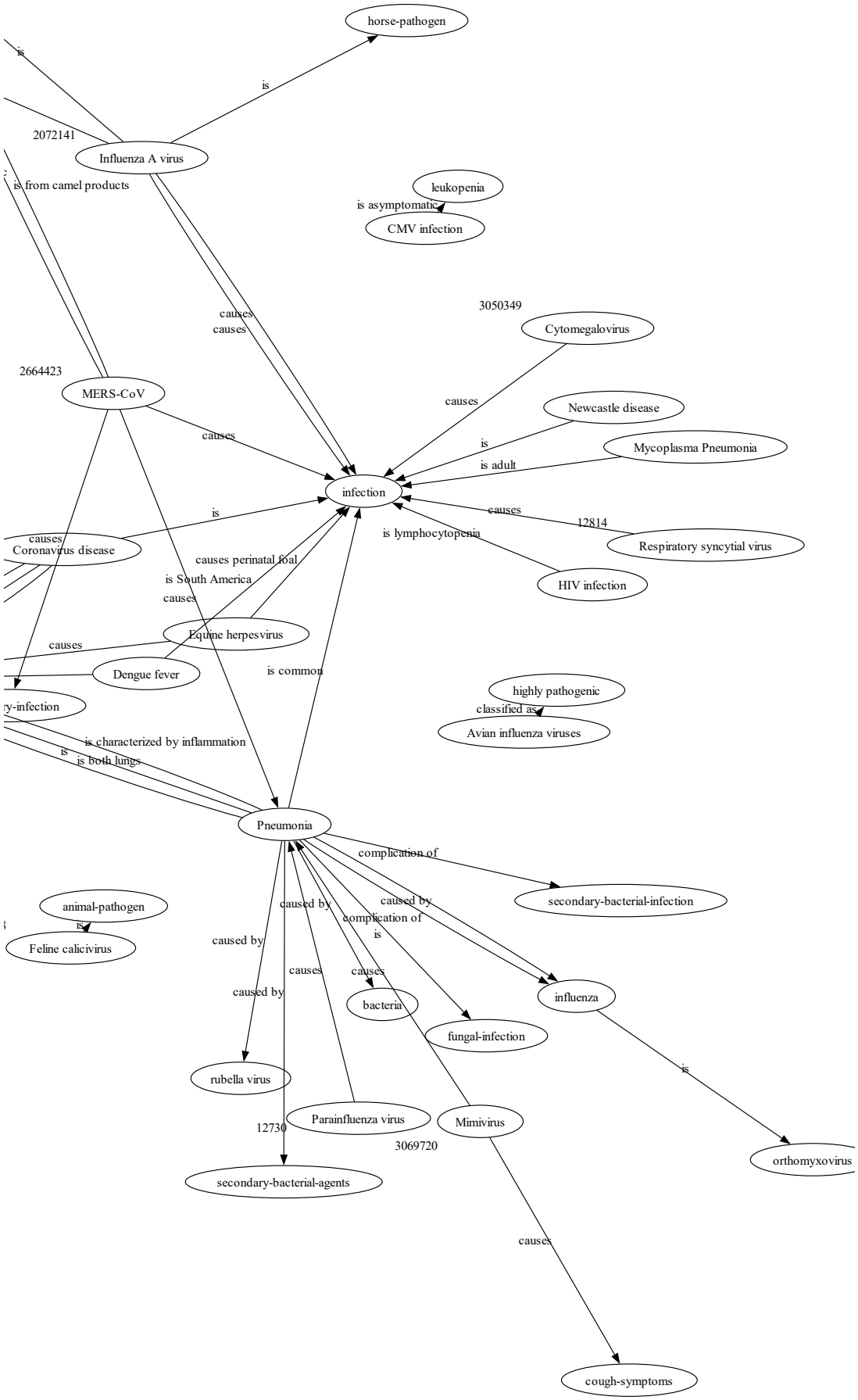

Figure 8. Visualization of multi-graph for the infectious disease ontology panel three.

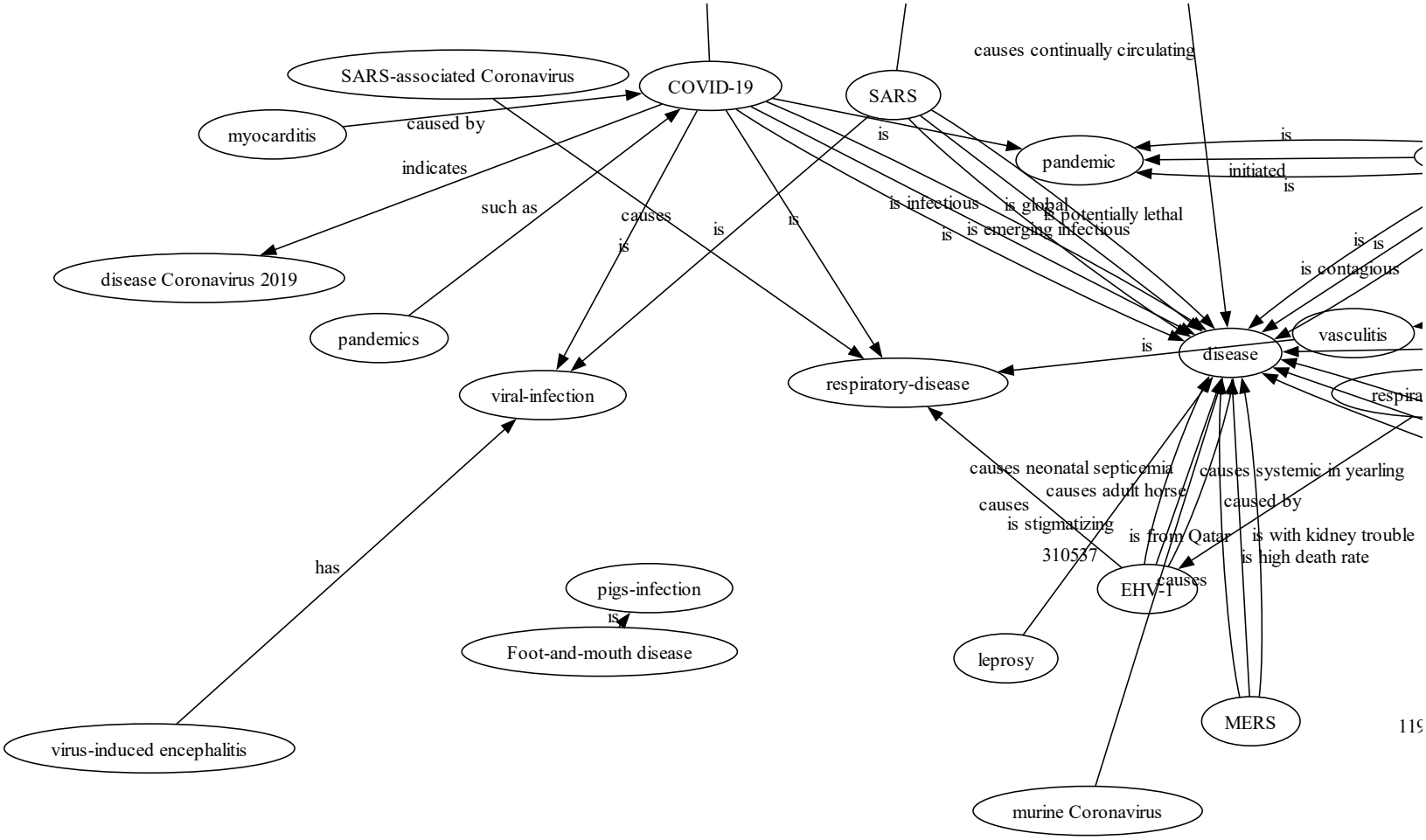

Figure 9. Visualization of multi-graph for the infectious disease ontology panel four.

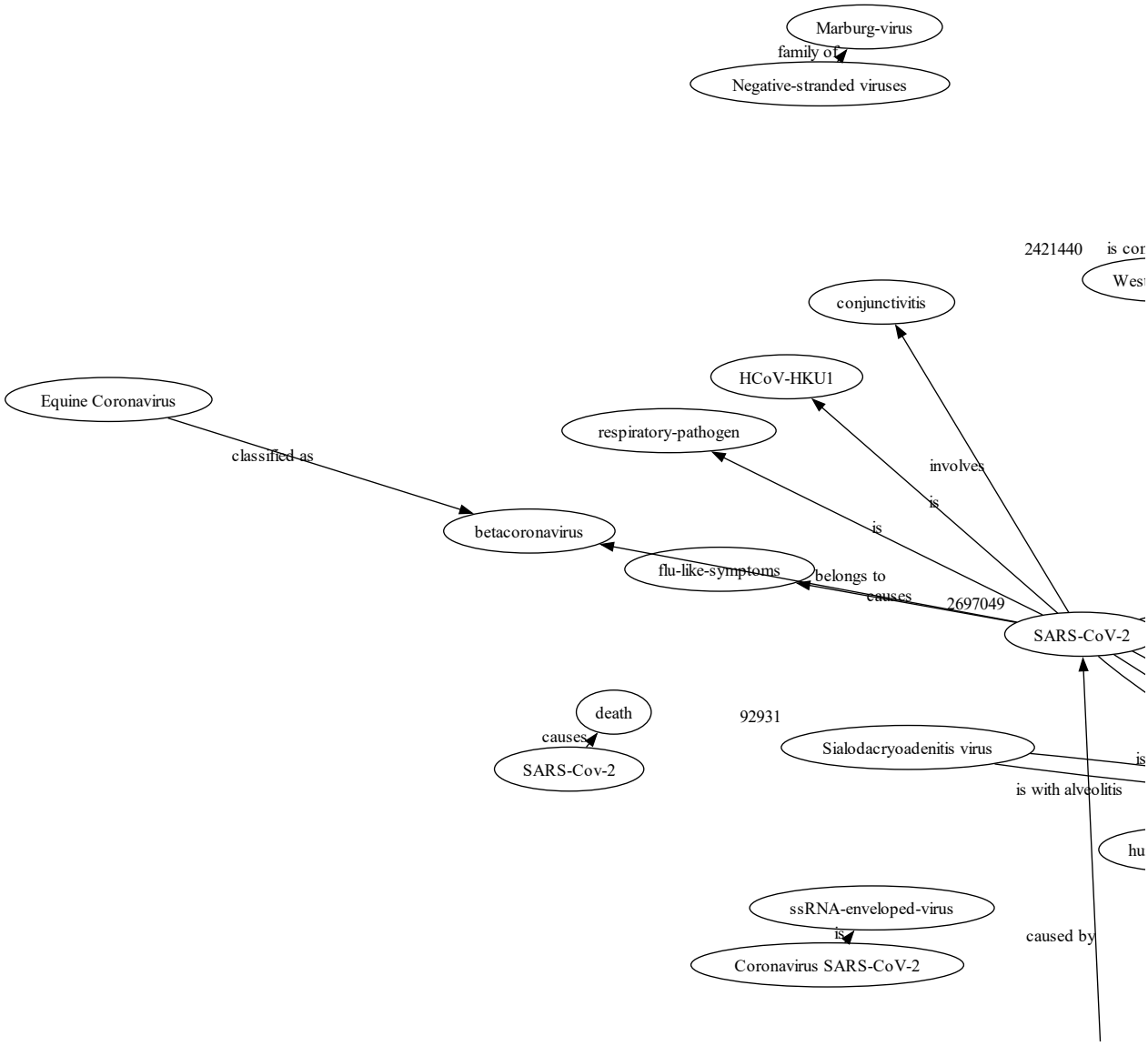

Figure 10. Visualization of multi-graph for the infectious disease ontology panel five.

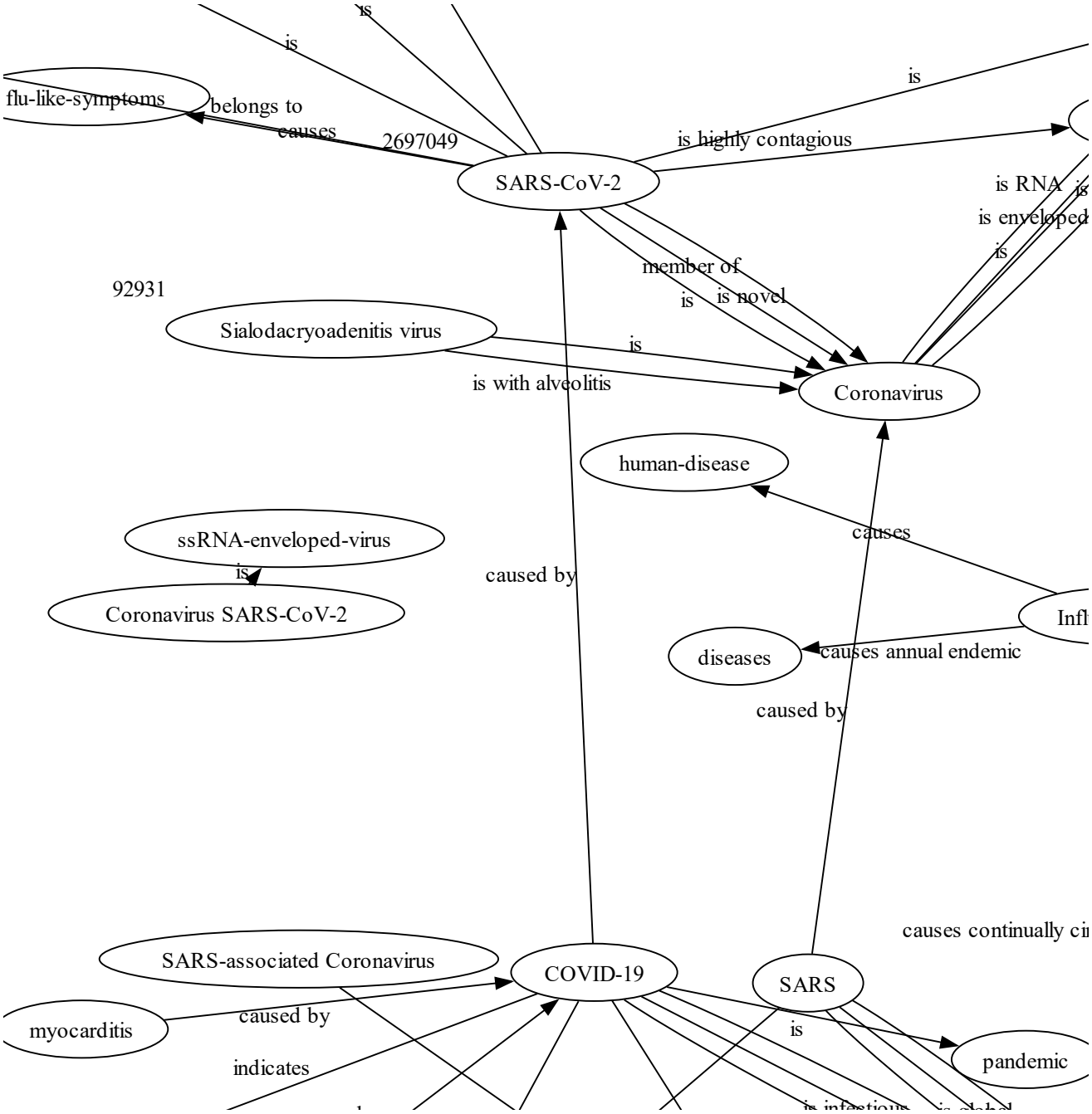

Figure 11. Visualization of multi-graph for the infectious disease ontology panel six.

4. Discussion

In this paper, we have shown how natural logic and Aristotelian logic can be combined to obtain propositions from biomedical sources which could be used in-principle to build a bio-ontology. In particular, we have gleaned and derived new rules from both traditions, which, among other things, allow us to carry out the following:

1. Fill in the true subject of sentences which have pronouns as their subject;
2. Derive a set of propositions from sentences with many clauses separated by commas when combined with "and" or "or";
3. Derive new propositions by substituting one term for another in complex prepositional phrases;
4. Derive a new proposition from an old one by removing adjectives or other restrictive conditions from the predicate of the sentence.

Using this apparatus, we have built a proof-of-principle bio-ontology for infectious diseases using the CORD-19 dataset. The novelty of our approach also includes the very way we implement inference for these rules. Inference based on all of these rules in the Aristotelian logic literature has historically not been carried out in any computational way. In this paper, we give the first computational implementation of these rules using state-of-the-art NLP techniques including dependency parsing. Furthermore, we apply them in a novel way to the finding of relations in text for the building up of a bio-ontology.

The ontology from this manuscript can be used in some ways similar to the COVID ontology CIDO [16]. The ontology can be used first of all to integrate information across multiple resources, as the Taxonomy IDs are cross-linked with genetics and gene expression data in NCBI. Using the integrated information, we can infer novel connections between genetics and disease states, to support COVID-19 and infectious disease data analysis. A second use case is to provide terms and help NLP researchers who are trying to carry out entity analysis (to give them a list of entities) or other kinds of text standardization on medical records, as often there are many variants of terms and the ontology can be an anchoring point. A third use-case is to help NLP researchers mine large volumes of text data related to COVID-19 efficiently, as ontology can be used to enhance NLP [38]. A fourth use case is to help doctors understand and disambiguate between different pathogens that cause the same disease (such as Pneumonia) or understand the different kinds of viruses (for instance large RNA viruses, ssRNA-enveloped viruses).

4.1. Prior Research and Future Directions

Previous work has tried to gather logical propositions from text using dependency analysis in a more grammatical form of logic [4,39], and also to simplify sentences using a form of natural logic [7] (called RelationIE when it is incorporated into the CoreNLP package by Manning). These cover some of the ground of the techniques we use to obtain our rules and proposition tool. The work of Reddy et al. is very relevant to ours, but it uses a different form of grammar-based logic which is rooted in the semantics of Donald Davidson, first-order logic, and lambda calculus. These are all great tools, but ultimately the complex formal propositions produced in their work are given by a long list of priority rules when there are conflicts between different kinds of rule applications. In our work, we do not have to consider the priority of various possibilities since the propositions are in natural language and are all that is required is to transform the sentence. As our rules are rooted in natural language, there is less translation required into formalism, and we need not exert as much control of priority of one rule over another. Though the RelationIE work, like ours, drops adjectives and other restrictive terms, the formal approach used there can sometimes produce unnatural and ungrammatical simplifications. Our approach still has a foundation in the theory of supposition and grammar, so this foundation can help us to avoid ungrammatical simplifications when we automate the applications of the rule (as those are invalid in our rule RP).

Prior work in bio-ontologies has primarily focused on various languages and application domains for bio-ontology. For example, Gene Ontology (GO), by far the most prominent and large bio-ontology, works with genetics data but is built based on user input [40,41]. It is based on a Directed Acyclic Graph (DAG) representation [42], which is flexible for many domains. GO is used primarily to provide a way for users to reconcile and standardize the use of terms for describing the same biological object [43]. GO users can also find biological processes or molecular functions associated with particular genes, and even find over-represented GO categories that assist in establishing the statistical significance of experimental effects [43]. Furthermore, GO has been used as a basic resource to find biological meaning associated with high-throughput genetics studies [44].

Many bio-ontologies are available on the Open Biomedical Ontologies [45,46] (OBO) foundry [47] in a standardized language format. The language of choice for OBO ontologies is the standardized Web Ontology Language (OWL [48]) based on Descriptive Logic [49] (OWL-DL). OWL is very expressive, logically rigorous, and is considered a gold standard

for ontologies in computer science. Note that as all of the following bio-ontologies are on the OBO foundry, no citation will be given for some of them. Bio-ontologies abound in many different areas, including the infectious disease ontology [50], Gene Ontology Molecular Function, Biological Process, and Cellular Component, PRO: Protein Ontology, CHEBI: Chemical Entities of Biological Interest and others. . . a list of sources highlighted by [51]. It is, however, the case that despite the presence of large knowledge bases and their logical rigor, there are some notable drawbacks that have been identified. First of all, the logic model seems to have some unintended consequences in inferring non-biological facts [51], and some issues have been identified in the use of certain models with regards to semantics [42]. Clearly, a great deal of progress has been made in making large databases cross-linking genetics data, gene-expression data, and even infectious diseases. The aim of making useful bio-ontologies has been well realized. Even for COVID-19, there was an ontology made based on existing ontologies [16]. There is also an area of bio-ontologies that seeks to grow the bio-ontology based on natural language processing techniques [52].

4.2. Limitations, Place in the Field, and Future Work

The primary limitation of this work is the number of entities included in the bio-ontology, and the number of relations, in addition to the lack of mechanistic information for the entities in the graph. Future work will expand on this to make it more practically useful for applications.

The limitations of our current approach on a practical front include primarily that the application of the rule RP and oblique syllogisms is not fully automated. The latter could be, reasonably easily. But the former seems less straightforward. However, there is a possibility that we could use the kind of logic that the RelationIE work uses and add more grammatical constraints to realize rule RP. In addition, the proposition tool still produces ungrammatical outputs, which could be rectified by using rule RP. Finally, we did not apply this to a large dataset and derive an ontology. Our present work was to show the usefulness of the rules in obtaining propositions about biomedical subjects, rather than to show the bio-ontology built from them. This would be a future goal for our research, once the inference rules are fully automated. The biggest theoretical limitation of our work as it stands is that though we have stated the new rules of logic, the model theory and logical foundations are not yet realized. Different components of natural logic and Aristotelian logic are combined with grammar. It is feasible that something like this could still be feasibly grounded in terms of model theory and truth conditions, as Terence Parsons' recent work does something like this [15]. A consequence of this, too, would be a mechanism for formal inference in the model, which would be an essential future work for us, especially when applying this to obtaining a more robust bio-ontology.

Existing work in natural logic is very exciting, and it is a great joy to contribute these formulations of inference rules to the literature. The work of Parsons [15] is akin to ours, but does not have some of our logic rules. It has a great deal more in terms of logical foundations though. The existing work in natural logic tries to develop more out of the fragment of first order logic which is monadic, or logic which has generalized quantifiers [9,20]. The tractability of the logical fragments in natural logic [9] is especially important, as the work of Parsons is sufficiently general as to be intractable (as it is equivalent to first-order logic, which is undecidable) [15]. Perhaps we could develop our use of the anaphora rules to exclude first-order-logic-style constructions on grammatical or suppositional grounds, as that is how Parsons obtains equivalence to that logic. More generally, if there is some way to combine the Parsons-style logic with natural logic fragments to realize the Aristotelian logic rules we outline here, we could achieve a natural logic that would be practically useful, computationally tractable, and fully formal.

In the sphere of bio-ontologies, our work is novel in that, in principle, it encodes propositions in a format that can support natural/Aristotelian logic reasoning. Certainly, it is not a large database that was produced by such methods compared to the impressive work carried out by others in this field. However, some of the benefits of this approach

include the lack of unintended consequences of our logical inference approach due to the natural fit between Aristotelian logic and biological reasoning, as opposed to the problems identified with the existential operator in mathematical logic [51]. The user entry of massive amounts of data into an ontology, while very important for validation of results, would have difficulty keeping abreast of current science. One of the ten factors, in fact, which recommend the use of an ontology for biomedical purposes is the fact that it is up to date with the current science [53], which would be easier if it were based on our methodology using NLP techniques and natural logic/Aristotelian logic approaches. The existing work in the field of using NLP for building ontologies is to use biological entity recognition and relation-extraction to build ontologies based on text, as outlined in [52]. Our approach to building ontologies is thus one of a burgeoning field of using NLP for bio-ontology. The key difference is that we are using an approach based on new techniques that are native to the natural language, which would make it easier to support inference in the resulting knowledge base. It is certainly possible to infer expressions in formal logic from text or vice versa [4], but they are more difficult to directly infer in natural language [13].

Specifically in the field of bio-ontologies for COVID, our method for building ontologies is advantageous over prior approaches in two respects: first, its greater scalability than comparable manually built COVID ontologies, and also its accuracy as compared to fully automatic NLP-built COVID ontologies. The scalability of the method presented in this paper is reasonably high as there was little manual post-processing and much of the processing of text data was carried out automatically using NLP techniques. The robustness of the proposed methods is demonstrated by the performance of the method with little to no prior background knowledge built in. The accuracy was 93.6 percent (6.4 percent error). In comparison with manual entry of bio-ontologies, which is by far the dominant approach (as used in [16–18]), our approach is vastly more scalable (though obviously less accurate). Compared to entity-based NLP analysis methods [19], our accuracy is higher, as it takes into account common grammatical structures to generate relations. In particular, there were three different works describing a COVID-related bio-ontology that were based on manual entry of the ontology [16–18]. When compared to CIDO [16,17], we have a similar subject matter with regards to phenotype, genetics information, various attributes, symptoms, etc., but do not cover specific mechanistic information for drug discovery. In comparison to Domingo-Fernandez et al. [18], we cover different things: the diseases and the related viruses cross-linked with genetics information, as opposed to specific protein/mechanistic cause information related to COVID. However, we have a solution based on text mining, which is not manually entered. The COVID ontology given there is vast but is based on manual encoding rather than the automatic (with human correction/supplementation) NLP-based methodology given here. The evaluation of the resulting ontologies would be based on the scientific accuracy of the underlying relationships in the ontology. As the scientific accuracy level for the manually entered ontology is very high (since the ontology was given based on scientific literature alone), our ontology's high accuracy despite automation is important. We found the accuracy of our ontology based on manually scoring the discovered entities and relationships was 93.6 percent (error rate: 6.4 percent). Another approach to a COVID bio-ontology for drug discovery was based on using ontology-informed named entity recognition NLP analysis [19], and is thus very scalable (even more than our approach since we used minimal manual post-processing). However, the entity-based NLP-built ontology has a relatively high rate (6–22 percent) of misses or false alarms for the entities identified in the knowledge graph.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info15110669/s1.

Author Contributions: T.G. provided to the codebase (including the proposition-tool), some results and most of the text describing the proposition tools for the Methods and Materials section. E.C. provided more code for the rest of the proposition handling and ontologies, compiled the principle results, and wrote the manuscript (besides what was indicated as due to T.G.). Software, T.G. and

E.C.; Writing—original draft, E.C.; Writing—review and editing, E.C. All authors have read and agreed to the published version of the manuscript.

Funding: Funding from The Nancy Cain and Jeffrey A. Marcus Science Endowment in Honor of President Donald A. Cowan.

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article and supplementary materials.

Acknowledgments: Lucy Chastain is duly acknowledged for her work in calculating the accuracy of the bio-ontology.

Conflicts of Interest: The authors declare no conflicts of interest. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Bard, J.B.; Rhee, S.Y. Ontologies in biology: Design, applications and future challenges. *Nat. Rev. Genet.* **2004**, *5*, 213–222. [CrossRef] [PubMed]
2. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.w.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 1–9. [CrossRef]
3. Clough, E.; Barrett, T. The gene expression omnibus database. *Stat. Genom. Methods Protoc.* **2016**, *1418*, 93–110.
4. Reddy, S.; Täckström, O.; Collins, M.; Kwiatkowski, T.; Das, D.; Steedman, M.; Lapata, M. Transforming dependency structures to logical forms for semantic parsing. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 127–140. [CrossRef]
5. Strawson, P.F. On referring. *Mind* **1950**, *59*, 320–344. [CrossRef]
6. Russell, B. Mr. Strawson on referring. *Mind* **1957**, *66*, 385–389. [CrossRef]
7. Angeli, G.; Premkumar, M.J.J.; Manning, C.D. Leveraging linguistic structure for open domain information extraction. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Volume 1: Long Papers, pp. 344–354.
8. van Benthem, J. A brief history of natural logic. In *Logic, Navya–Nyaya & Applications, Homage to Bimal Krishna Matilal*; Chakraborti, M.K., Löwe, B., Mitra, M.N., Sarukkai, S., Eds.; College Publications: London, UK, 2008; pp. 21–42.
9. Moss, L.S.; Wollowski, M. Natural Logic in AI and Cognitive Science. In Proceedings of the MAICS, Fort Wayne, IN, USA, 28–29 April 2017; pp. 41–46.
10. Montague, R. *Universal Grammar*; Routledge: London, UK, 1974; Volume 1970, pp. 222–246.
11. Montague, R. English as a Formal Language. In *Logic and Philosophy for Linguists*; Mouton & Co., B.V.: The Hague, The Netherlands, 1974.
12. Montague, R. The proper treatment of quantification in ordinary English. In *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*; Springer: Dordrecht, The Netherlands, 1973; pp. 221–242.
13. Sommers, F.; Englebretsen, G. *An Invitation to Formal Reasoning: The Logic of Terms*; Routledge: London, UK, 2017.
14. Maritain, J. *Formal Logic*; Sheed & Ward: London, UK, 1946.
15. Parsons, T. *Articulating Medieval Logic*; OUP Oxford: Oxford, UK, 2014.
16. He, Y.; Yu, H.; Huffman, A.; Lin, A.Y.; Natale, D.A.; Beverley, J.; Zheng, L.; Perl, Y.; Wang, Z.; Liu, Y.; et al. A comprehensive update on CIDO: The community-based coronavirus infectious disease ontology. *J. Biomed. Semant.* **2022**, *13*, 25. [CrossRef]
17. He, Y.; Yu, H.; Ong, E.; Wang, Y.; Liu, Y.; Huffman, A.; Huang, H.h.; Beverley, J.; Hur, J.; Yang, X.; et al. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Sci. Data* **2020**, *7*, 181. [CrossRef]
18. Domingo-Fernández, D.; Baksi, S.; Schultz, B.; Gadiya, Y.; Karki, R.; Raschka, T.; Ebeling, C.; Hofmann-Apitius, M.; Kodamullil, A.T. COVID-19 Knowledge Graph: A computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics* **2021**, *37*, 1332–1334. [CrossRef]
19. Wang, Q.; Li, M.; Wang, X.; Parulian, N.; Han, G.; Ma, J.; Tu, J.; Lin, Y.; Zhang, R.H.; Liu, W.; et al. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, Online, 6–11 June 2021; pp. 66–77.
20. Barwise, J.; Cooper, R. Generalized quantifiers and natural language. In *Philosophy, Language, and Artificial Intelligence: Resources for Processing Natural Language*; Springer: Berlin/Heidelberg, Germany, 1981; pp. 241–301.
21. Wang, L.L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Eide, D.; Funk, K.; Kinney, R.M.; Liu, Z.; Merrill, W.; et al. CORD-19: The COVID-19 Open Research Dataset. *arXiv* **2020**, arXiv:2004.10706v4.
22. NCBI. *Taxonomy [Internet]*; NCBI: Bethesda, MD, USA, 2004.
23. Honnibal, M.; Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

24. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.Y. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinform.* **2022**, *23*, bbac409. [CrossRef] [PubMed]
25. Hearst, M.A. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics, Nantes, France, 23–28 August 1992.
26. WolframAlpha [Internet]. Available online: https://www.wolframalpha.com/ (accessed on 24 September 2024).
27. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.
28. NCBI. *Nucleotide [Internet]*; NCBI: Bethesda, MD, USA, 2004.
29. Chomsky, N. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*; Cambridge University Press: Cambridge, UK, 2009.
30. Kaufman, L.; Straus, J. *The Blue Book of Grammar and Punctuation: An Easy-to-Use Guide with Clear Rules, Real-World Examples, and Reproducible Quizzes*; John Wiley & Sons: Hoboken, NJ, USA, 2021.
31. Sauteur, P.M.M.; Beeton, M.L.; Pereyre, S.; Bébéar, C.; Gardette, M.; Hénin, N.; Wagner, N.; Fischer, A.; Vitale, A.; Lemaire, B.; et al. Mycoplasma pneumoniae: Delayed re-emergence after COVID-19 pandemic restrictions. *Lancet Microbe* **2024**, 5, e100–e101. [CrossRef] [PubMed]
32. Wikipedia. Mycoplasma Pneumoniae—Wikipedia, The Free Encyclopedia. 2024. Available online: http://en.wikipedia.org/w/index.php?title=Mycoplasma%20pneumoniae&oldid=1191067769 (accessed on 13 February 2024).
33. Wikipedia. Mycoplasma Pneumonia—Wikipedia, The Free Encyclopedia. 2024. Available online: http://en.wikipedia.org/w/index.php?title=Mycoplasma%20pneumonia&oldid=1190118092 (accessed on 13 February 2024).
34. Lobina, D.J.; Demestre, J.; García-Albea, J.E.; Guasch, M. Default meanings: Language's logical connectives between comprehension and reasoning. *Linguist. Philos.* **2023**, *46*, 135–168. [CrossRef]
35. Cock, P.J.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef]
36. Hagberg, A.A.; Schult, D.A.; Swart, P.J. Exploring Network Structure, Dynamics, and Function using NetworkX. In Proceedings of the 7th Python in Science Conference, Pasadena, CA, USA, 19–24 August 2008; Varoquaux, G., Vaught, T., Millman, J., Eds.; SciPy Proceedings: Pasadena , CA, USA, 2008; pp. 11–15.
37. Ellson, J.; Gansner, E.; Koutsofios, L.; North, S.C.; Woodhull, G. Graphviz—Open source graph drawing tools. In Proceedings of the Graph Drawing: 9th International Symposium, GD 2001, Vienna, Austria, 23–26 September 2001; Revised Papers 9; Springer: Berlin/Heidelberg, Germany, 2002; pp. 483–484.
38. Erekhinskaya, T.; Strebkov, D.; Patel, S.; Balakrishna, M.; Tatu, M.; Moldovan, D. Ten ways of leveraging ontologies for natural language processing and its enterprise applications. In Proceedings of the International Workshop on Semantic Big Data, Portland, OR, USA, 14–19 June 2020; pp. 1–6.
39. Reddy, S.; Täckström, O.; Petrov, S.; Steedman, M.; Lapata, M. Universal semantic parsing. *arXiv* **2017**, arXiv:1702.03196.
40. Consortium, G.O. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **2019**, *47*, D330–D338.
41. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
42. Aranguren, M.E.; Bechhofer, S.; Lord, P.; Sattler, U.; Stevens, R. Understanding and using the meaning of statements in a bio-ontology: Recasting the Gene Ontology in OWL. *BMC Bioinform.* **2007**, *8*, 57
43. Rubin, D.L.; Shah, N.H.; Noy, N.F. Biomedical ontologies: A functional perspective. *Briefings Bioinform.* **2008**, *9*, 75–90. [CrossRef]
44. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37*, 1–13. [CrossRef]
45. Mungall, C.J. Obol: Integrating language and meaning in bio-ontologies. *Comp. Funct. Genom.* **2004**, *5*, 509–520. [CrossRef] [PubMed]
46. Smith, B.; Ashburner, M.; Rosse, C.; Bard, J.; Bug, W.; Ceusters, W.; Goldberg, L.J.; Eilbeck, K.; Ireland, A.; Mungall, C.J.; et al. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **2007**, *25*, 1251–1255. [CrossRef] [PubMed]
47. Jackson, R.; Matentzoglu, N.; Overton, J.A.; Vita, R.; Balhoff, J.P.; Buttigieg, P.L.; Carbon, S.; Courtot, M.; Diehl, A.D.; Dooley, D.M.; et al. OBO Foundry in 2021: Operationalizing open data principles to evaluate ontologies. *Database* **2021**, *2021*, baab069. [CrossRef] [PubMed]
48. McGuinness, D.L.; Van Harmelen, F. OWL web ontology language overview. *W3C Recomm.* **2004**, *10*, 2004.
49. Nardi, D.; Brachman, R.J. An introduction to description logics. *Descr. Log. Handb.* **2003**, *1*, 40.
50. Babcock, S.; Beverley, J.; Cowell, L.G.; Smith, B. The infectious disease ontology in the age of COVID-19. *J. Biomed. Semant.* **2021**, *12*, 13. [CrossRef]
51. Boeker, M.; Tudose, I.; Hastings, J.; Schober, D.; Schulz, S. Unintended consequences of existential quantifications in biomedical ontologies. *BMC Bioinform.* **2011**, *12*, 1–10. [CrossRef]

52. Friedman, C.; Borlawsky, T.; Shagina, L.; Xing, H.R.; Lussier, Y.A. Bio-ontology and text: Bridging the modeling gap. *Bioinformatics* **2006**, *22*, 2421–2429. [CrossRef]
53. Malone, J.; Stevens, R.; Jupp, S.; Hancocks, T.; Parkinson, H.; Brooksbank, C. Ten simple rules for selecting a bio-ontology. *PLoS Comput. Biol.* **2016**, *12*, e1004743. [CrossRef]