Data Models for Annotating Biomedical Scholarly Publications: the Case of CORD-19

Houcemeddine Turki*
turkiabdelwaheb@hotmail.fr
Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax
Sfax, Tunisia

Mohamed Ali Hadj Taieb
mohamedali.hajtaieb@fss.usf.tn
Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax
Sfax, Tunisia

Alejandro Piad-Morffis
apiad@matcom.uh.cu
Faculty of Math and Computer Science, University of Havana
La Habana 10200, Cuba

Mohamed Ben Aouicha
mohamed.benaouicha@fss.usf.tn
Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax
Sfax, Tunisia

René Fabrice Bile
bilerene@gmail.com
National Polytechnic School of Maroua, University of Maroua
Maroua, Far-North, Cameroon

ABSTRACT

Semantic text annotations have been a key factor for supporting computer applications ranging from knowledge graph construction to biomedical question answering. In this systematic review, we provide an analysis of the data models that have been applied to semantic annotation projects for the scholarly publications available in the CORD-19 dataset, an open database of the full texts of scholarly publications about COVID-19. Based on Google Scholar and the screening of specific research venues, we retrieve seventeen publications on the topic mostly from the United States of America. Subsequently, we outline and explain the inline semantic annotation models currently applied on the full texts of biomedical scholarly publications. Then, we discuss the data models currently used with reference to semantic annotation projects on the CORD-19 dataset to provide interesting directions for the development of semantic annotation models and projects.

CCS CONCEPTS

• **Information systems → Document representation**; • **Applied computing → Annotation**; • **Computing methodologies → Knowledge representation and reasoning**.

KEYWORDS

Semantic relations, Semantic annotations, Named entity annotation, Semantic relation annotation, Annotation models, CORD-19

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524675

ACM Reference Format:
Houcemeddine Turki, Mohamed Ali Hadj Taieb, Alejandro Piad-Morffis, Mohamed Ben Aouicha, and René Fabrice Bile. 2022. Data Models for Annotating Biomedical Scholarly Publications: the Case of CORD-19. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3487553.3524675

1 INTRODUCTION

Currently, scholarly literature is ever-changing and growing daily due to the regular discoveries on new scientific facts and findings, particularly related to the exploration of new fields of interest or the development of novel research methods [52]. In late 2019, a new infectious disease called COVID-19, characterized by acute respiratory symptoms and induced by the SARS-CoV-2 virus, has emerged causing a widespread pandemic by March 2020 [27]. In the context of this evolving outbreak, new scholarly papers appear every day to study the multidisciplinary facets of the disease ranging from molecular and clinical aspects of the infection [13, 27] to microbial safety [23]. The set of these COVID-19-related scholarly publications is considered as big data distinguished by its volume, variety, velocity, and veracity [48] and is consequently hard to process by humans due to the rapidly changing patterns of the COVID-19 information involved in research outputs and to the growing number of scholarly findings and evidence about the medical condition [13]. The high volume of unstructured information available on COVID-19 is hard to process without sophisticated infrastructure [48] and complex computational models, based on machine learning and natural language processing techniques [50]. Alternatively, leveraging semantically structured representations of knowledge, such as Wikidata's COVID-19 Knowledge Graph [58, 61], enables the design of computational methods to explore, analyze, and integrate COVID-19-related information in decision-support systems with ease. Such semantic resources allow both manual explorations by domain experts and automatic processing by computational methods, which makes them useful in a variety of scenarios, from tracking epidemiological evolution [45] to generating public health recommendations [20, 62], and supporting different informative [21] or didactic [22] needs.

However, the development of knowledge graphs is only possible through the extraction of the semantic features of scientific knowledge from textual resources such as scholarly papers [28]. This can be simply done by applying the process of identifying and marking semantic information in raw texts to generate in-line semantic annotations of biomedical texts [28]. In the context of the COVID-19 outbreak, the Allen Institute for Artificial Intelligence and other partner institutions have issued an open dataset of scholarly publications about the infectious pandemic (so-called *CORD-19*[1]) to motivate the development of tools for the semantic annotation of COVID-19 information and consequently for the generation of open COVID-19 knowledge graph [63]. Since the development of the CORD-19 dataset, many efforts have been made to build efficient tools for the semantic annotation of COVID-19 research publications. Developed projects have been hosted on linked data interfaces such as *OpenLink Virtuoso* [37] and *BRAT* [14, 41] to allow the sustainable enrichment of the CORD-19 dataset with annotations by human efforts and machines and the reuse and integration of the generated semantic information using flexible user-friendly interfaces, APIs and federated SPARQL queries to identify and validate COVID-19 knowledge [26, 37].

In this systematic review, we will study the data models currently used for the in-line semantic annotation of the full texts of CORD-19 scholarly publications. We will begin by presenting the methods for the identification of evidences about the text annotation of scholarly publications about the COVID-19 pandemic (Section 2). Then, we will outline how CORD-19 semantic annotation project operates thanks to an in-depth analysis of research papers (Section 3). After that, we will describe the features of the data models used for the CORD-19 semantic annotations (Section 4). Later, we discuss these data models by outlining several practical limitations of the models used by CORD-19 annotation projects (Section 5). Finally, we will conclude the status of the efforts for the semantic annotation of biomedical research papers and identify future directions for this research work (Section 6).

2 PROPOSED APPROACH

Most of the main initiatives about the text annotation of CORD-19 scholarly publications have been shown during the two *NLP COVID-19 Workshops*[2] occurring during *ACL 2020* and *EMNLP 2020* conferences, the two *SciNLP workshops*[3] held during *AKBC 2020* and *AKBC 2021* conferences, and the online meetings on *CORD-19 semantic annotations* hosted by the World Wide Web Consortium[4]. That is why we screen the publications of these research venues to identify primary research publications about the semantic text annotations of the CORD-19 scholarly publications. As well, we search for the other papers related to the topic by applying *"Annotation" AND ("CORD-19" OR "CORD19")* as a query to *Google Scholar*[5]. As of December 10, 2021, 473 publications have been analyzed to identify the relevant evidences about the CORD-19 semantic annotation projects: 17 projects from the *W3C Semantic Annotation Projects*' Showcases, 108 publications from COVID-19 NLP-related research events, and 348 *Google Scholar* query search results. After the screening process, only seventeen publications have been identified as relevant for our systematic review. Seven of them (41.2%) are related to the *W3C Semantic Annotation Projects*' Showcases and six of them (35.3%) were shown in COVID-19 NLP-related research workshops. We screen the research evidences to find their data models for the semantic annotation of CORD-19 research papers as well as the target of the annotation projects.

[1]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
[2]https://www.nlpcovid19workshop.org/
[3]https://scinlp.org/
[4]https://github.com/w3c/hcls/wiki/CORD-19-Semantic-Annotation-Projects
[5]https://scholar.google

Table 1: CORD-19 semantic annotation projects per data model: Named Entity Annotation (*NE*), Concept-Based Relation Annotation (*CR*), Action-Based Relation Annotation (*AR*), and Sentence Annotation (*S*)

Work	Country	Data Models
Hope (2021) [24]	USA-SWE-ISR	NE, AR
Colic (2020) [14]	SUI	NE
Du (2021) [17]	USA	NE, CR
Esteva (2021) [19]	USA	S
Huang (2020) [25]	USA	S
Ilievski (2020) [26]	USA-BRA	NE, CR
Lymperopoulos (2020) [35]	USA	NE
Michel (2020) [37]	FRA	NE, S
Piad-Morffis (2020) [41]	CUB-ESP	NE, AR
Reese (2021) [46]	USA	NE, CR
Tykhonov (2020) [59]	NED-UKR	NE, CR
Wang (2021) [64, 65]	USA	NE, CR, S
Suryanarayanan (2021) [54]	USA-KEN-ISR	S
Logette (2021) [33]	SUI	NE, CR
Basu (2020) [5]	IND	NE, CR
Wolinski (2021) [66]	FRA	NE

3 ANALYSIS OF RETRIEVED PAPERS

We found out that most of the CORD-19 semantic annotation projects were led by the *United States of America* (9 out of 17) as shown in the Table 1. This goes in line with the current status of the computer science research where the USA dominates the field for years from the perspective of productivity and citations [51]. Other countries were featured in ten publications led by Switzerland, Israel and France with two research papers for each. In this section, we analyze the *Methods* part of these publications to identify the data models used for representing annotations (e.g., *named entity annotation*). As well, we retrieve the purposes of developing inline CORD-19 semantic annotation projects (e.g., *NLP* and *ML* tasks) and subsequently deduce the influence of the motivation of developing such annotation projects on the design of data models.

3.1 Models

When investigating the data models in the research publications (Table 1), we found a large interest to apply named entity annotation (14 out of 17, 82.3%) to the CORD-19 dataset. This is explained by the availability of annotation tools, pre-trained language models, and machine learning models that allow such a task with very significant accuracy rates [37]. Beyond this, Table 1 reveals that seven works are interested to link between semantically related named

Table 2: CORD-19 semantic annotation projects per target: Knowledge graph (*KG*), Dashboard (*D*), Question Answering (*QA*), Generator of Hypertext Links (*H*), Research Analysis (*RA*), Search Engine (*SE*), and Text Summarization (*TS*)

Work	KG	D	QA	Others
Hope et al. (2021) [24]	✓			SE
Colic et al. (2020) [14]				SE, TS
Du et al. (2021) [17]				RA
Esteva et al. (2021) [19]			✓	TS, SE
Huang et al. (2020) [25]				RA
Ilievski et al. (2020) [26]	✓			
Lymperopoulos et al. (2020) [35]				H
Michel et al. (2020) [37]	✓	✓	✓	
Piad-Morffis et al. (2020) [41]	✓			
Reese et al. (2021) [46]	✓			
Tykhonov et al. (2020) [59]	✓			
Wang et al. (2021) [64, 65]	✓	✓	✓	
Suryanarayanan et al. (2021) [54]		✓		
Logette et al. (2021) [33]	✓			RA
Basu et al. (2020) [5]	✓			
Wolinski et al. (2021) [66]		✓		

entity annotations in scholarly texts to form concept-based relation annotations using a variety of embedding techniques and machine learning algorithms. The common use of concept-based relation annotations is evident as the complexity of such an annotation is reduced where named entity annotations are available. Despite the dependency of many works on these two data models, there are several works that tried to develop the use of other types of semantic annotation for the CORD-19 scholarly publications (7 out of 17, 41.2%). Five works were interested in annotating sentences instead of being based on named entity to develop semantic annotations. Here, the software requirements are not advanced as annotations will be assigned as labels to whole sentences and there is no need to go inside every sentence to assign their relation types or their topics. As well, two works have investigated the use of Action-Based Relation Annotation as a data model for annotating CORD-19 scholarly publications. In this situation, scientists tried to identify relation types using the text span annotation of the phrasal verbs standing for them in the sentences. The aim of such an annotation is to restrict the use of relation types in the CORD-19 semantic annotation to the generic ones [24, 41].

3.2 Targets

To highlight why four different data models are used for CORD-19 semantic annotations, we extracted the reasons of the considered annotation projects. We found out most of the works (9 out of 17, 52.9%) that use named entity annotation coupled with concept-based or action-based relation annotation aim to the creation of knowledge graphs about COVID-19 from the analyzed scholarly publications as clearly shown in Table 2. Other minor reasons for such a combination can range from driving CORD-19 search engines to the analysis of the COVID-19 research outputs. As for sentence annotation, it is used alongside named entity annotation to drive question answering systems about the pandemic. This usage of sentence annotation is encouraged by the long-term use of natural language texts within the framework of the TREC initiatives for answering questions as natural language texts are human-readable and provide details that are not always represented in fully-structured knowledge graphs [19]. Both named entity-based annotations and sentence-based annotations are used in several research papers (4 out of 17, 23.5%) to feed COVID-19 dashboards visualizing aspects of the COVID-19 pandemic and disease as revealed by scholarly publications. On the one hand, this is explained by the easiness of extracting features from knowledge graphs, particularly when represented in the *Resource Description Framework* (RDF) Format, using a variety of tools including APIs and query languages like SPARQL [58]. On the other hand, the availability of open-source analytics tools, particularly *Python* and *R* packages, that generate quality visualizations from a processed input, allowed the creation of real-time human-friendly graphic representations of structured information, including the CORD-19 semantic annotations [67].

4 DATA MODELS

Semantic annotation projects for the CORD-19 dataset use linked data formats such as RDF, XML, and JSON to represent in-line semantic information [37] and mainly rely on text span annotations to extract semantic features at the level of sentences [25] and named entities [35] as shown in Figure 1. Text span annotations are made and aligned to external resources thanks to human efforts [41] or fine-grained annotation automation tools such as *PubTator*[6], *SciSpacy*[7], *DBpedia Spotlight*[8], *Entity-fishing*[9], *NCBO BioPortal Annotator*[10], and *Annotator+*[11] [37, 64]. Such annotations are later enriched with other similar semantic annotations using deep learning techniques like convolutional neural networks and Long Short-Term Memory and language models such as BERT, ELMo, and GloVe [17, 25, 35]. These annotations can be restricted to identifying concepts or sentences in an excerpt [41, 59] or expanded to annotate the classes of recognized items [25, 37]. Although word

Figure 1: Types of text span annotation models

and graph embeddings can be used to identify the links between annotated named entities for the automatic construction of knowledge graphs [12], particularly in the context of the CORD-19 Research Dataset [26, 46, 64], several projects have chosen to perform in-line annotations of semantic relations to ensure the verifiability of generated statements based on user contributions and robust computer methods [24, 41]. When validated by a panel of clinical

[6] https://www.ncbi.nlm.nih.gov/research/pubtator/
[7] https://allenai.github.io/scispacy/
[8] https://www.dbpedia-spotlight.org/
[9] https://github.com/kermitt2/entity-fishing
[10] https://bioportal.bioontology.org/annotator
[11] https://bioportal.bioontology.org/annotatorplus

specialists, these relation annotations can serve as a benchmark for an explainable and more trustworthy machine learning-based retrieval of biomedical and clinical semantic relations [71].

The data models for annotating biomedical relations include concept-based annotation models [46], action-based annotation models [24, 41], and sentence annotation models [37] as shown in Figure 2. The concept-based relation annotation models link between annotated concepts using relation annotations where the property is a non-taxonomic (biomedical) or taxonomic (generic or temporal) relation type [46]. The action-based relation annotation models depend on the text span annotation of the terms corresponding to the evocated relation type as actions and link them with concepts using a limited number of generic properties [24, 41]. The sentence relation annotation models assign an explanation string or a piece of semantic information [37] to a sentence represented as a text span annotation. In this section, we will provide details of the different data models used for the text span annotation and the semantic relation annotation of biomedical texts, in the context of the pandemic.

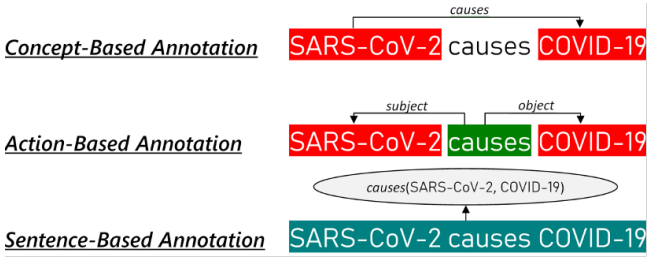

Figure 2: Types of semantic relation annotation models

4.1 Named entity annotation

Named entity annotation is a process that combines named entity recognition and entity linking [14]. The output of a named entity annotation includes the definition of the beginning and end of the annotated text span, assigned labels as well as external identifiers linking the annotation to items in external knowledge resources such as Wikidata and Wikipedia as shown in Figure 3 [37, 64]. Named Entity Recognition (NER) identities concepts in a given sentence and assigns them to their corresponding classes (e.g., *disease, medication*, etc.) or a general class entitled "entity" or "concept" [39]. NER systems can be categorized as (*i*) Knowledge-based NER systems that do not require annotated training data as they rely on

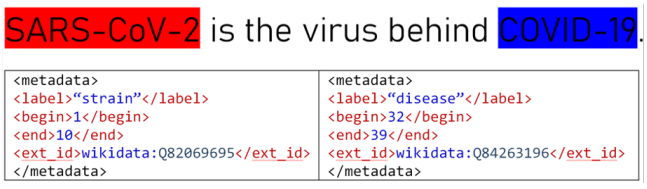

Figure 3: XML Structured Representation of named entity annotations in a sample excerpt about COVID-19

specific resources and domain-specific knowledge. Figure S1 outlines an unsupervised method for NER applied to biomedical context through the use of semantic resources including *UMLS Metathesaurus* and through classification based on "signature" similarity [70]. (*ii*) Unsupervised systems that require training data expressing features including orthography (e.g., *capitalization*), the context of the entity, words contained within named entities, etc. (*iii*) Feature-engineered supervised systems that learn to make predictions by training inputs and their expected outputs. (*iv*) Feature-inferring neural network systems based on neural network architectures for NER, with feature vectors and word embedding models including different granularity levels such as word and character [68–70]. NER is closely related to entity linking (EL), aiming to align the entities mentioned in a text with reference to a knowledge base, e.g., *Unified Medical Language System* (UMLS) in the biomedical domain [49]. Linked entities are useful for correcting the classification of entity types. Exploring approaches for jointly performing NER and EL enhances entity type classification, and entity linking, so that each subtask benefits from the partial output by the subtasks, and alleviates error propagations [14]. This can be achieved through the application of knowledge graph-based semantic similarity measures and word embeddings to assess the semantic features of the annotated concepts and use them later to verify the consistency of their assignment to a given class [31].

NER as an annotation goal can be taken from two sides: **Multiclass classification** means a classification task with more than two classes; e.g., classify a set of images of fruits which may be oranges, apples, or pears. Multiclass classification assumes that each sample is *assigned to one and only one label*: a fruit can be either an apple or a pear but not both at the same time. Also, the specificity degree of the class representing the concept is so important for the end-user application such as *Disease and Symptom*. **Multilabel classification** assigns to each sample a set of target labels. This can be thought of as predicting properties of a data point that are *not mutually exclusive*, such as topics that are relevant for a document. A text might be about any of religion, politics, finance, or education at the same time or none of these. Researchers giving multi-labeled classification NER annotation methods are very limited [39].

The annotation granularity process aims to provide an annotation level. This means that the annotation attempts not only to go into a deeper analysis of the documents but also to consider a fine-granularity [15]. The term "terminal renal insufficiency" will be annotated on different levels. First of all, the complete term "terminal renal insufficiency" will be annotated as a medical condition, which is closest to the UMLS entry. Besides, "renal insufficiency" and "insufficiency" will be also annotated as Medical_Condition to achieve a fine granularity. Furthermore, strings such as "terminal" will be annotated as Medical Specification and "renal" as Body Part. There are different reasons for the detailed annotation level. Firstly, "terminal renal insufficiency" is the most specific term which includes all other information [15]. Often NER systems target the longest and most specific match. However, UMLS might not cover necessarily all variants. A fine granularity might help at a later stage to learn larger constructs (e.g., *adjective + compound noun*) that are not in the dictionary. Multiword expressions (MWE) consist of several words (in the conventionally understood sense) but behave as single words to some extent. MWE discovery methods

treat some specific categories according to the linguistic properties (compound nouns, verb construction, etc.) [15]. Several approaches are proposed including specific pipelines dependent form techniques, language, and resources [See Appendix A in [15]]. The *seed term collection* step is about collecting seed terms for entity classes, upon which signature vectors of the classes will be generated in the third step. In biomedical related applications, for example, classes of entities are defined by users by choosing the corresponding UMLS-specific concepts [39]. Then, the boundary detection step is to detect the boundaries of entities, collecting candidates for entity classification. To remove those noun phrases that are not entities of interest, researchers employ an inverse document frequency (IDF) based technique to filter candidates generated by the NP chunker [39]. Finally, for the last step, entity classification aims to obtain entities of the same class tending to have similar vocabulary and context. For example, in clinical text, the word "insufficiency" is highly likely to be inside an entity of class "Problem", but not "Treatment" or "Test" [39]. The compound nouns like "renal insufficiency" are determined using specific resources like UMLS based on MWE methods. The similarity-based method exploits the distributional semantics. For the signature generation as presented in Figure S2, it is a vector of internal and context words for a certain object.

4.2 Concept-based relation annotation

In several situations, named entity annotations are expanded to link related concepts together for developing a structured annotation of semantic relations presented by the analyzed excerpt. This annotation type is known as concept-based relation annotation [24, 40]. The design of concept-based relation annotation models only represents named entities as text span annotations [46]. Any other type of information including the relations between identified concepts, the references of the annotated relations, and annotation alignment information is represented as structured metadata of the annotations [64]. Here, the relations are labeled by their types (e.g., *direct up-regulation* and *indirect up-regulation* for biological processes) and are assigned to the named entities representing their subjects and objects to appear as arrows linking between concepts in user-friendly interfaces [46]. This implies that the annotation data model should support every type of biomedical relations concerned by the annotation process as a distinct category to efficiently work [46]. Concept-based relation annotation represents the attributes and references as qualifiers in the form of triples (so-called *reification*) where the subject is the original relation, the property is the type of the attribute or reference and the object is a value or an annotated named entity just similarly to the Wikidata knowledge graph [30, 58, 61]. The choice of the information represented by the qualifiers of the annotations is decided by the creators of the data models of the projects according to the context of the work [30]. An example is shown in Figure 4.

In both annotation types, the choice of the categories to represent classes of relations or named entities depends on several variables and constraints. The coverage of the relation types depends on the scope (e.g., *genomic data* or *clinical information*) and purpose (e.g., *knowledge graph creation* or *biomedical text summarization*) of the annotation. Several projects choose to annotate limited types of concepts and relations such as biological processes [24] while

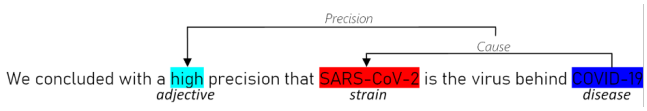

Figure 4: Example semantic annotation using concept-based annotation model

several other initiatives tend to be more exhaustive covering many aspects of biomedical and clinical knowledge [60]. Furthermore, annotation projects can also choose a certain level of abstraction to label the annotations with general categories (i.e., the hypernym of the type of annotation such as concept, direct mechanism, and indirect mechanism) or with extremely detailed categories describing the characteristics of the annotated named entity or relation further than the annotation type (i.e., the hyponym of the type of annotation such as non-drug symptomatic treatment) [24, 60]. Moreover, annotation projects try to avoid interference between chosen categories to ensure a unique representation of semantic annotations [10] and an insignificant ambiguity in the usage and interpretation of semantic annotations [11]. This implies the elimination of inverse properties and of similar or closely related categories that are difficult to differentiate using semantic similarity measures and word embeddings [31] to avoid errors of representations and redundancy [10, 11]. The choice of the categories should inform us of the complexity of the classification of semantic relations in biomedical texts and should be a determining factor in the accuracy of machine learning semantic annotation algorithms [57].

4.3 Sentence annotation

As clearly shown above, sentence annotation models include every sentence of an assessed excerpt in a single text span annotation [25]. Similar to named entity annotations, each sentence annotation is characterized by its start position, end position, and its assigned label [18]. The label can be just a mention that the text span annotation corresponds to a sentence, especially when using simple annotation tools such as *Hypothesis* (https://web.hypothes.is/) [8] and when the sentence annotations are coupled with named entity annotations to allow better semantic link prediction by only considering named entities included in the same sentences [59]. The label can also be a class involved in a set of categories describing a linguistic or functional feature of the sentence [25]. Chosen categories can reflect the intonation of the sentence such as confirmation, uncertainty, denial, warning, judgment, advice or irony [47, 55], a rhetoric relation (e.g., *contrast*, *correction*, *conclusion*, and *support*), or a grammatical feature (e.g., *polarity* and *strength*) the sentence represents [55]. Categories can also serve to identify and characterize the main outcomes of a research paper by representing the context of the sentence such as the author of the sentence and the time and place of its statement [55] or such as the link of the sentence with the process of a research project or with the metadata of a research paper driving Bibliometric-Enhanced Information Retrieval [7, 25]. Categories representing the process of a research project include hypothesis, background, purpose, method, and finding or contribution as shown in Figure S3 [25]. Bibliographic metadata like specified keywords, locations of stated co-authoring institutions, scholarly

references, the source of the publication including the sentence or declared research grants are assigned as labels for sentence annotations where such information or their synonyms or related terms are mentioned in the full text of scholarly publications [7, 37].

Sometimes, sentence annotations can be assigned labels explaining the meaning of the identified phrase. In such a situation, the labeling categories should be defined in a way to respect the constraints described in the "Named entity annotation and Concept-Based Relation annotation" section. Annotated labels can be the type of relation expressed by the sentence or a named entity included in the phrase [37, 59]. The label can also be a natural language raw text such as a question having the annotated sentence as an answer as in the CovidQA dataset [19] or an excerpt in another research publication similar or related to the sentence summarizing or explaining it [29, 32]. It can also be a semantic statement in a knowledge graph or an ontology such as DBPedia and Wikidata as shown in Figure 2 [37]. In that situation, the annotation is called a sentence-based relation annotation and assigns a natural language sentence to machine-readable semantic information to provide a database for biomedical relation extraction from scholarly publications [71]. When combined with other sentence annotations and with named entity annotations, such a relation annotation allows the linguistic analysis of biomedical texts for the automation of natural question answering, biomedical text summarization, and knowledge graph creation and refinement [29, 37, 59]. Examples for every type of sentence annotation are shown in Table S1.

4.4 Action-based relation annotation

Action-based annotation models use Subject-Action-Target triplets as the core semantic element in sentences. The Action identifies a word or phrase that expresses an event, often appearing as a verbal construction, although it can also appear in any other grammatical role [24, 42]. The Subject identifies which entities are performing the given action, while the Target identifies entities that receive the effect of the action. An example of this annotation model is *SAT+R* [42], which is represented in Figure 5.

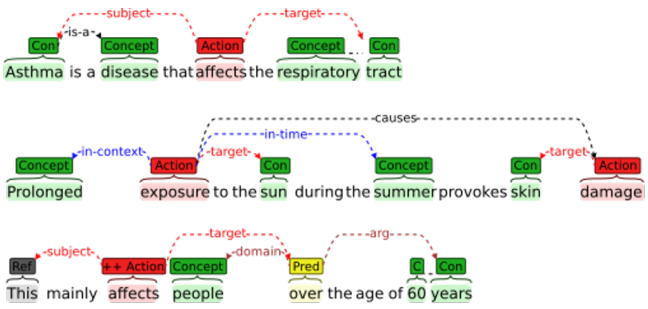

Figure 5: Examples of semantic annotations using *SAT+R*

The *SAT+R* annotation model defines several semantic layers to annotate in a sentence. First, the most relevant Concepts and Actions are identified in the text. An action can be linked to a concept by the subject or the target relation, with the semantic meaning explained above. Furthermore, actions can also be linked to other actions, which allows the annotation of composite concepts. Two additional entities are defined, Reference, which indicates a missing reference, and Predicate, which allows the construction of concepts based on filters or properties.

Predicates define a domain and one or more arguments that identify which is the entity the predicate refers to, and additional semantic elements, respectively. An example of both annotations can be seen in the final sentence in Figure 5. Predicates in *SAT+R* should not be confused with the role of predicates in RDF (actually, that role is closely related to Action in *SAT+R*). The purpose of predicates in *SAT+R* is to construct complex concepts by attaching qualifiers to an atomic concept. As an example, in the third sentence in Figure 5, the concept people is qualified by the predicate over with an argument of 60 years. This creates a composite concept that represents the people whose age is over 60 years, where the meaning of over is to be taken from context. Then, connecting affects with over as its target indicates that who is affected by the subject of the action is not all people, but only those that fulfill the predicate. This model also defines a set of four ontological relations: is-a, part-of, has-property, and same-as, with the usual semantics. These relations can link any of the four entity types. Their purpose is to annotate structural elements, as they could be defined in a taxonomy or ontology. Two additional relations are causes and entails which correspond to the concepts of causality and entailment. Their difference is that causality requires an explicit mechanism by which a concept or action determines the causation of another, while entailment is a logical relation that does not imply causation. Furthermore, three contextual relations are defined, in-time, in-place, and in-context, which allow the definition of events or actions that only occur when conditioned by the existence or occurrence of other concepts or actions.

Finally, the model defines four attributes: emphasized, diminished, uncertain, and negated. The first two allow capturing the grammatical function of augmentatives and diminutives without recurring to the annotation of adverbs or other grammatical elements. The third allows the definition of uncertain events, and the fourth, the annotation of negation. An example of an emphasized concept is present in the last sentence, in the word affects, which is being modified by the adverb mainly. The use of the emphasized attribute on affects represents this notion of emphasis while detaching its semantic meaning from the surface form of whatever adverb or adjective is being used. Similarly, words like sometimes or possibly could hint at the use of the uncertain attribute on a concept or action, again to detach the semantic meaning from the surface textual form.

Dealing with negation in a semantic annotation model such as *SAT+R* is complex because it is not immediately obvious what the negation of a concept means, and what is the scope. The most straightforward case is when an action is negated (e.g., *Asthma does not affect the digestive system*). In this case, the negation particles are not annotated, but rather the action itself is marked with the negated attribute. The scope of the negation affects only the action, and it represents a logical statement equivalent to saying that the triplet <*Asthma,affects,digestive-system*> is false in the domain. However, sometimes it is necessary to negate ontological relations (e.g., *Asthma is not a cancer*). In this case, the negation attribute is attached to the target of the relation, which yields the triplet <Asthma,is-a,not(cancer)>. In this case, the negation of a concept

is interpreted as the complement of the concept in the domain, e.g.,*everything that is not cancer*, which is logically equivalent to negating the is-a relation. Similar reasoning is applied to all the remaining relations, where the negation operator takes a semantic interpretation that stems from the semantic of that relation.

The guiding principle behind the design of *SAT+R* is to annotate the most significant semantic elements in text with the least possible reliance on grammar and syntax. It has been inspired in other general-purpose annotation models such as *AMR* [4] but it has been simplified to improve the performance of both human annotators and automatic systems. Although it was designed for general domain text, it has been used mostly for annotation documents in the health domain, both in English [41] and Spanish language [43].

5 DISCUSSION

When seeing the methods used by semantic annotation projects in the context of the COVID-19 pandemic, we found that a significant part of them is built upon manual annotations or knowledge graphs, particularly Wikidata, as shown in Section 4. Despite the possibility of the automatic retrieval of COVID-19 information from scholarly publications using semantic embeddings and machine learning, the availability of manual semantic annotation projects for CORD-19 datasets is explained by the usefulness of these annotations to provide more accuracy to information retrieval tasks [6]. Driving computer applications by manually curated semantic resources including annotated datasets and biomedical ontologies can open ways to explore the reasons behind the deficiency of automatic annotation algorithms and to have a more trustworthy output that does not conflict with human knowledge [56]. The analysis of the findings returned by the CORD-19 annotation projects (particularly *F-measures*) with the outputs of the Sections 3 and 4 provides interesting insights into the efficiency of semantic annotation models and methods. On the one hand, when comparing human semantic annotations with automatic ones, it is clear that knowledge resources-based systems[12] for the identification and classification of semantic annotations based on word embeddings and neural networks [17, 24, 35] are more efficient than the human annotation efforts and the automatic annotation methods driven by a corpus of manual annotations [41]. This situation is significant for named entity annotations but is more critical for the text span annotations of actions [24, 41]. On the other hand, when comparing the action-based relation annotation projects of the CORD-19 scholarly publications[13], it is clear that the action-based relation annotation model considering subject and object as the only relation types [24] allows a better accuracy of machine learning than the ones using *SAT+R* annotation model [41]. Although the efficiency of the supervised machine learning from action-based semantic annotations seems to significantly vary from a project to another, its accuracy seems to be always inferior to the one of concept-based relation annotation [9]. To study the reasons behind these results, we will discuss all the data models for CORD-19 semantic annotations by applying them to four examples as shown in Table 3.

[12]Knowledge resources involve online encyclopedias, mainly Wikipedia [35] and knowledge graphs, particularly Wikidata and DBpedia [37].

[13]Detailed statistics and human efforts to develop a corpus of manual CORD-19 named entity annotations for Piad-Morffis et al. (2020) [41] are available at https://github.com/knowledge-learning/cord19-ann/tree/master/data/output/packs/submitted.

Table 3: Sample excerpts about the COVID-19 disease.

Identifier	Example
S1	The pathogenesis of COVID-19 is caused by the molecular aspects of SARS-CoV-2 virus
S2	Anemia is rarely a symptom of COVID-19 disease
S3	The development of vaccines by firms will certainly not be a very short journey
S4	The maximal incubation period for COVID-19 is 14 days

The achievement of limited accuracy for named entity annotation, particularly in the context of human annotations, is explained by a lack of an exhaustive definition of the annotation granularity causing the appearance of differences in text span annotation habits between different humans and systems [53]. This matter is common in biomedical natural language processing as most of the subjects and objects of annotated sentences are not just constituted of one term [53]. As shown in the examples S1-S4, noun phrases are commonly used as subjects and objects of sentences [38]. These noun phrases can begin with an article (e.g., *the pathogenesis* [S1], *a symptom* [S2], *14 days* [S4]) and include a preposition (e.g., *pathogenesis of COVID-19* [S1], *symptom of COVID-19 disease* [S2], *development of vaccines by firms* [S3], *maximal incubation period for COVID-19* [S4]), an adjective (e.g., *molecular aspects* [S1], *short journey* [S3], *maximal incubation period* [S4]), an adverb (e.g. *very short journey* [S3]), or a compound noun (e.g., *SARS-CoV-2 virus* [S4], *COVID-19 disease* [S2], *maximal incubation period* [S4]). Here, two major concerns should be highlighted. Articles, prepositions, and conjunctions are generally considered in Natural Language Processing as stop words that should not be taken into consideration in topic modeling and other interesting tasks [1]. However, the consideration of such noun phrase constituents in the text span annotations has not been discussed as it should be and this is what explains the significant disagreements between human experts in annotating named entities in CORD-19 scholarly publications [41].

Furthermore, noun phrases (mainly compound nouns and the ones including prepositions) involve substring terms that exist in reference ontologies and that can be annotated as well. For example, COVID-19 pandemic[14] can be considered as one term and can be split apart and annotated as two terms COVID-19[15] and pandemic[16]. That is why detailed guidelines should be defined to outline the constraints for semantic annotation overlapping and for defining the situations when a word or adjective can be included in a text span annotation. This can enhance the preservation of a unique and accurate normalized representation of semantic annotations for named entities in a sentence and consequently ameliorate the quality of supervised machine learning from manually curated semantic annotations. Such a normalization work should take into consideration the types of relations that will be considered to link between named entity annotations. Effectively, the adoption of

[14]Wikidata:Q81068910

[15]Wikidata:Q84263196

[16]Wikidata:Q12184

several relation types by a given data model can influence the tendencies of users related to the granularity and uniqueness of concept annotation (e.g., *is-a*, *part-of*, and *has-property*) as it can be found by comparing the results of Hope, et al. (2021) [24] to the ones of Piad-Morffis, et al. (2020) [41]. Despite the limitations of named entity annotations, their situation seems to be significantly better than the one of action text span annotations [24, 41]. The identification of verbs in a given sentence is tricky particularly because many factors should be considered when deciding the beginning and the end of a text span annotations for actions.

First, a significant portion of verbs uses adverbs [34] and negation [16] to emphasize the order, certainty, and frequency of facts. Example S3 (Table 3) is a typical sentence illustrating this. On the one hand, the verb in this example is in the negative form (i.e., *will not be*). On the other hand, an adverb (i.e., *certainly*) has been embedded to confirm the statement. Both negation and adverbs provide an important piece of information about the action and should be considered when annotating verbs in textual resources. However, it will be crucial to know if this implies the inclusion of the negation and adverbs in the text span annotation of each action, particularly because this will alter the complexity of the construction of knowledge graphs from the action-based semantic annotations due to a lack of normalization of actions.

Second, the verb conjugation is complicated in itself as it uses modal verbs (e.g., *can*, *should*), suffixes (e.g. *-ed*, *-ing*, and *–s*), and compound tenses (e.g., *will be* [S3]) [16]. It is important to investigate why and how modal verbs should be a part of the action semantic annotations, particularly as they outline important characteristics of the action. For example, the negation and adverbs can be substituted by an attribute of the action annotation (*emphasized*, *diminished*, *uncertain*, and *negated*) as represented in the *SAT+R* models [41]. Similarly, it is interesting to see whether the suffixes of the conjugated verbs should be included in the text span annotation of actions or whether only the verb stem should be included in the annotation (e.g., *caused* → *cause*) to ensure a full normalization of the list of annotated actions and consequently to easily construct knowledge graphs from action-based relation annotations [3].

Third, the use of the passive voice is common in scholarly publications and that is why annotating an action in a passive sentence can be a common challenge for annotators [44]. In such a situation, they are many ways to annotate an action as shown in Figure 6. These ways are motivated by what we have discussed about lemmatization, compound nouns, and prepositions. It will be important to decide the way that is most appropriate for action-based semantic annotation to avoid divergence is the annotation of actions in passive voice and consequently to prevent inconsistencies in the machine learning of action-based relation annotations.

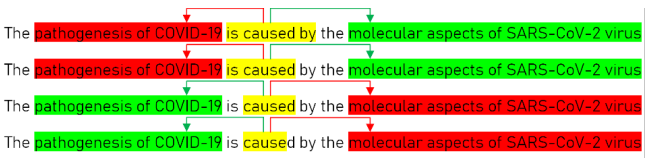

Figure 6: Examples of action-based semantic annotations for S1: Action (Yellow), Subject (Red), and Object (Green).

Although concept-based relation annotations can be practically less difficult from the perspective of data modeling than action-based relation annotations, they suffer from slight limitations that block the achievement of an absolute agreement between manual annotators and consequently a right efficiency of machine learning algorithms for concept-based relation annotation projects [36]. These limitations are mainly related to the choice between a relation type and its reverse when both are supported by the annotation data model (e.g., *medical condition treated* and *drug used for treatment*), to the choice between two close relation types (e.g., *significant drug interaction* and *significant protein interaction*), and the choice of the sense of annotation of a symmetric relation (e.g., *significant drug interaction*) [36]. This deficiency should encourage the development of guidelines to handle reverse and symmetric relation types in concept-based relation annotation projects so that better accuracy rates for such initiatives can be easily achieved.

6 CONCLUSION

In this systematic review, we provided an overview of the data models used to annotate the CORD-19 scholarly publications. We outlined the state-of-the-art of the topic and we explained the data models used to annotate biomedical scholarly publications, mainly during the COVID-19 pandemic. We discussed the advantages of each data model of semantic annotations and we provided an overview of the major matters limiting their practical efficiency, particularly the text span annotation granularity and the assignment of relation types and attributes. Solving such problems can help enhance the accuracy of semantic annotation projects and better the robustness of knowledge-based systems. As a future direction of this paper, we propose to develop our work by including other bibliographic databases such as *Web of Science*, *PubMed*, and *Scopus* and by applying visualization tools such as *Bibliometrix* on the bibliographic metadata of the considered scholarly evidences for a more detailed explanation of the research dynamics behind semantic annotation projects for the CORD-19 dataset [2]. Furthermore, we propose to establish detailed guidelines for annotating textual resources in a standardized way by considering the limitations of semantic annotation data models and then to develop a machine-readable edition of these rules to ameliorate fully automated semantic annotation algorithms.

7 ACKNOWLEDGMENTS

The work of Houcemeddine Turki, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha is supported by the Ministry of Higher Education and Scientific Research in Tunisia (MoHESR) in the framework of Federated Research Project PRFCOV19-D1-P1. The work of Alejandro Piad-Morffis is funded by the University of Alicante, the University of Havana, the Generalitat Valenciana (Conselleria d'Educació, Cultura i Esport), and the Spanish Government through the projects LIVING-LANG (RTI2018-094653-B-C22) and SIIA (PROMETEO/2018/089). We thank David Booth (World Wide Web Consortium, United States of America), Daniel Mietchen (University of Virginia), Franck Michel (Université Côte d'Azur, France), Daniel Schwabe (Pontifical Catholic University of Rio de Janeiro, Brazil), and Guoqian Jiang (Mayo Clinic, United States of America) for providing useful comments for the refinement of this output.

We thank Sisonkebiotik, particularly Chris Fourie (University of the Witwatersrand, South Africa), for supporting this work.

REFERENCES

[1] Bassam Al-Shargabi, Waseem Al-Romimah, and Fekry Olayah. 2011. A comparative study for Arabic text classification algorithms based on stop words elimination. In *Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications*. Association for Computing Machinery, Amman, Jordan, 72–78. https://doi.org/10.1145/1980822.1980833

[2] Massimo Aria and Corrado Cuccurullo. 2017. Bibliometrix : An R-tool for comprehensive science mapping analysis. *Journal of Informetrics* 11, 4 (2017), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

[3] Saeid Balaneshinkordan and Alexander Kotov. 2016. An empirical comparison of term association and knowledge graphs for query expansion. In *European conference on information retrieval*. Springer, Cham, Padua, Italy, 761–767. https://doi.org/10.1007/978-3-319-30671-1_65

[4] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, Sofia, Bulgaria, 178–186. https://aclanthology.org/W13-2322

[5] Sayantan Basu, Sinchani Chakraborty, Atif Hassan, Sana Siddique, and Ashish Anand. 2020. ERLKG: Entity Representation Learning and Knowledge Graph based association analysis of COVID-19 through mining of unstructured biomedical corpora. In *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics, Online, 127–137. https://doi.org/10.18653/v1/2020.sdp-1.15

[6] Asma Ben Abacha and Pierre Zweigenbaum. 2011. Medical Entity Recognition: A Comparaison of Semantic and Statistical Methods. In *Proceedings of BioNLP 2011 Workshop*. Association for Computational Linguistics, Portland, Oregon, USA, 56–64. https://aclanthology.org/W11-0207

[7] Marc Bertin and Iana Atanassova. 2016. Weak Links and Strong Meaning: The Complex Phenomenon of Negational Citations. In *Proceedings of the Third Workshop on Bibliometric-enhanced Information Retrieval co-located with the 38th European Conference on Information Retrieval (ECIR 2016)*. CEUR-WS, Padua, Italy, 14–25. http://ceur-ws.org/Vol-1567/paper2.pdf

[8] Maria Bonn and Jonathan McGlone. 2014. New feature: Article annotation with hypothes.is. *The Journal of Electronic Publishing* 17, 2 (2014). https://doi.org/10.3998/3336451.0017.201

[9] Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics* 9, 1 (2008), 207:1–207:14. https://doi.org/10.1186/1471-2105-9-207

[10] Harry Bunt. 2010. A methodology for designing semantic annotation languages exploring semantic-syntactic ISO-morphisms. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*. CEUR-WS, Hong Kong, 29–46. https://let.uvt.nl/general/people/bunt/docs/bunt-icgl4.pdf

[11] Harry Bunt. 2015. On the Principles of Semantic Annotation. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*. Association for Computational Linguistics, London, UK. https://aclanthology.org/W15-0201

[12] Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research* 63 (2018), 743–788. https://doi.org/10.1613/jair.1.11259

[13] Giovanni Colavizza, Rodrigo Costas, Vincent A. Traag, Nees Jan van Eck, Thed van Leeuwen, and Ludo Waltman. 2021. A scientometric overview of cord-19. *PLoS One* 16, 1 (2021), e0244839. https://doi.org/10.1371/journal.pone.0244839

[14] Nico Colic, Lenz Furrer, and Fabio Rinaldi. 2020. Annotating the Pandemic: Named Entity Recognition and Normalisation in COVID-19 Literature. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.27

[15] Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword Expression Processing: A survey. *Computational Linguistics* 43, 4 (2017), 837–892. https://doi.org/10.1162/coli_a_00302

[16] David Crystal. 1966. Specification and English tenses. *Journal of Linguistics* 2, 1 (1966), 1–34. https://doi.org/10.1017/s0022226700001304

[17] Caifan Du, Johanna Cohoon, Patrice Lopez, and James Howison. 2021. Softcite dataset: A dataset of software mentions in Biomedical and Economic Research Publications. *Journal of the Association for Information Science and Technology* 72, 7 (2021), 870–884. https://doi.org/10.1002/asi.24454

[18] Thomas Emerson and Benson Margulies. 2009. *U.S. Patent No. 7,562,009*. U.S. Patent and Trademark Office, Washington, DC.

[19] Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2021. Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *npj Digital Medicine* 4, 1 (2021), 68. https://doi.org/10.1038/s41746-021-00437-0

[20] Mohamed Frikha, Houcemeddine Turki, Mohamed Ben Ahmed Mhiri, and Faiez Gargouri. 2019. Trust Level Computation based on Time-aware Social Interactions for Recommending Medical Tourism Destinations. *Journal of Information Assurance and Security* 14, 3 (2019), 86–97. http://www.mirlabs.org/jias/secured/Volume14-Issue3/Paper9.pdf

[21] Julien Grosjean, Tayeb Merabti, Nicolas Griffon, Badisse Dahamna, and Stéfan J. Darmoni. 2012. Teaching medicine with a terminology/ontology portal. *Studies in Health Technology and Informatics* 180 (2012), 949–953. https://doi.org/10.3233/978-1-61499-101-4-949

[22] Christian Grévisse, Rubén Manrique, Olga Mariño, and Steffen Rothkugel. 2018. Knowledge graph-based teacher support for learning material authoring. In *Colombian Conference on Computing*. Springer, Cham, Hong Kong, 177–191. https://doi.org/10.1007/978-3-319-98998-3_14

[23] Milad Haghani, Michiel C.J. Bliemer, Floris Goerlandt, and Jie Li. 2020. The scientific literature on Coronaviruses, COVID-19 and its associated safety-related research dimensions: A scientometric analysis and scoping review. *Safety Science* 129 (2020), 104806. https://doi.org/10.1016/j.ssci.2020.104806

[24] Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel Weld, Roy Schwartz, and Hannaneh Hajishirzi. 2021. Extracting a knowledge base of mechanisms from COVID-19 papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4489–4503. https://doi.org/10.18653/v1/2021.naacl-main.355

[25] Ting-Hao Kenneth Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Using a Non-Expert Crowd to Annotate Research Aspects on 10,000+ Abstracts in the COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online. https://aclanthology.org/2020.nlpcovid19-acl.6

[26] Filip Ilievski, Daniel Garijo, Hans Chalupsky, Naren Teja Divvala, Yixiang Yao, Craig Rogers, Rongpeng Li, Jun Liu, Amandeep Singh, Amandeep Schwabe, and Pedro Szekely. 2020. KGTK: A Toolkit for Large Knowledge Graph Manipulation and Analysis. In *International Semantic Web Conference*. Springer, Cham, Athens, Greece, 278–293. https://doi.org/10.1007/978-3-030-62466-8_18

[27] Dima Kagan, Jacob Moran-Gilad, and Michael Fire. 2020. Scientometric trends for coronaviruses and other emerging viral infections. *GigaScience* 9, 8 (2020), giaa085. https://doi.org/10.1093/gigascience/giaa085

[28] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. 2004. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics* 2, 1 (2004), 49–79. https://doi.org/10.1016/j.websem.2004.07.005

[29] Abdullah Kogilavani and Palanisamy Balasubramanie. 2012. Sentence annotation based enhanced semantic summary generation from multiple documents. *American Journal of Applied Sciences* 9, 7 (2012), 1063–1070. https://doi.org/10.3844/ajassp.2012.1063.1070

[30] Valeria Lapina and Volha Petukhova. 2017. Classification of modal meaning in negotiation dialogues. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*. https://aclanthology.org/W17-7406

[31] Juan J. Lastra-Díaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana García-Serrano, Mohamed Ben Aouicha, and Eneko Agirre. 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence* 85 (2019), 645–665. https://doi.org/10.1016/j.engappai.2019.07.010

[32] Chin-Yew Lin and Eduard Hovy. 2002. From Single to Multi-document Summarization. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 457–464. https://doi.org/10.3115/1073083.1073160

[33] Emmanuelle Logette, Charlotte Lorin, Cyrille Favreau, Eugenia Oshurko, Jay S. Coggan, Francesco Casalegno, Mohameth François Sy, Caitlin Monney, Marine Bertschy, Emilie Delattre, and et al. 2021. A machine-generated view of the role of blood glucose levels in the severity of COVID-19. *Frontiers in Public Health* 9 (2021), 695139. https://doi.org/10.3389/fpubh.2021.695139

[34] Chao Lu, Yi Bu, Xianlei Dong, Jie Wang, Ying Ding, Vincent Larivière, Cassidy R. Sugimoto, Logan Paul, and Chengzhi Zhang. 2019. Analyzing linguistic complexity and scientific impact. *Journal of Informetrics* 13, 3 (2019), 817–829. https://doi.org/10.1016/j.joi.2019.07.004

[35] Panagiotis Lymperopoulos, Haoling Qiu, and Bonan Min. 2020. Concept Wikification for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.29

[36] Adam Meyers, Giancarlo Lee, Angus Grieve-Smith, Yifan He, and Harriet Taber. 2014. Annotating Relations in Scientific Articles. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 4601–4608.

http://www.lrec-conf.org/proceedings/lrec2014/pdf/385_Paper.pdf

[37] Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, Olivier Corby, Raphaël Gazzotti, Alain Giboin, Santiago Marro, Tobias Mayer, Mathieu Simon, Serena Villata, and Marco Winckler. 2020. Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research. In *International Semantic Web Conference*. Springer, Cham, Athens, Greece, 294–310. https://doi.org/10.1007/978-3-030-62466-8_19

[38] Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya, and Jun'ichi Tsujii. 2006. Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, 1017–1024. https://doi.org/10.3115/1220175.1220303

[39] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes* 30, 1 (2007), 3–26. https://doi.org/10.1075/li.30.1.03nad

[40] Saša Nešic, Mehdi Jazayeri, Fabio Crestani, and Dragan Gaševic. 2010. Concept-based semantic annotation, indexing and retrieval of office-like document units. In *RIAO '10: Adaptivity, Personalization and Fusion of Heterogeneous Information*. Association for Computing Machinery, Athens, Greece, 134–135. https://doi.org/10.5555/1937055.1937088

[41] Alejandro Piad-Morffis, Suilan Estevez-Velarde, Ernesto Luis Estevanell-Valladares, Yoan Gutiérrez, Andrés Montoyo, Rafael Muñoz, and Yudivián Almeida-Cruz. 2020. Knowledge Discovery in COVID-19 Research Literature. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.22

[42] Alejandro Piad-Morffis, Yoan Guitérrez, Suilan Estevez-Velarde, and Rafael Muñoz. 2019. A General-Purpose Annotation Model for Knowledge Discovery: Case Study in Spanish Clinical Text. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 79–88. https://doi.org/10.18653/v1/W19-1910

[43] Alejandro Piad-Morffis, Yoan Gutiérrez, and Rafael Muñoz. 2019. A corpus to support eHealth Knowledge Discovery Technologies. *Journal of Biomedical Informatics* 94 (2019), 103172. https://doi.org/10.1016/j.jbi.2019.103172

[44] Leong Ping Alvin. 2014. The passive voice in scientific writing. the current norm in science journals. *Journal of Science Communication* 13, 01 (2014), A03. https://doi.org/10.22323/2.13010203

[45] Gergo Pinter, Imre Felde, Amir Mosavi, Pedram Ghamisi, and Richard Gloaguen. 2020. COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. *Mathematics* 8, 6 (2020), 890. https://doi.org/10.3390/math8060890

[46] Justin T. Reese, Deepak Unni, Tiffany J. Callahan, Luca Cappelletti, Vida Ravanmehr, Seth Carbon, Kent A. Shefchek, Benjamin M. Good, James P. Balhoff, Tommaso Fontana, and et al. 2021. KG-COVID-19: A Framework to produce customized knowledge graphs for COVID-19 response. *Patterns* 2, 1 (2021), 100155. https://doi.org/10.1016/j.patter.2020.100155

[47] Spyridon Samothrakis and Maria Fasli. 2015. Emotional sentence annotation helps predict fiction genre. *PLoS One* 10, 11 (2015), e0141922. https://doi.org/10.1371/journal.pone.0141922

[48] Hiba Sebei, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha. 2018. Review of social media analytics process and Big Data Pipeline. *Social Network Analysis and Mining* 8, 1 (2018), 30. https://doi.org/10.1007/s13278-018-0507-0

[49] Wei Shen, Jiawei Han, Jianyong Wang, Xiaojie Yuan, and Zhenglu Yang. 2018. Shine+: A general framework for domain-specific entity linking with heterogeneous information networks. *IEEE Transactions on Knowledge and Data Engineering* 30, 2 (2018), 353–366. https://doi.org/10.1109/tkde.2017.2730862

[50] Junaid Shuja, Eisa Alanazi, Waleed Alasmary, and Abdulaziz Alashaikh. 2020. Covid-19 open source data sets: A comprehensive survey. *Applied Intelligence* 51, 3 (2020), 1296–1325. https://doi.org/10.1007/s10489-020-01862-6

[51] Vivek Kumar Singh, Ashraf Uddin, and David Pinto. 2015. Computer Science Research: The top 100 institutions in India and in the world. *Scientometrics* 104, 2 (2015), 529–553. https://doi.org/10.1007/s11192-015-1612-8

[52] Henry Small. 2006. Tracking and predicting growth areas in science. *Scientometrics* 68, 3 (2006), 595–610. https://doi.org/10.1007/s11192-006-0132-y

[53] Brett R South, Shuying Shen, Makoto Jones, Jennifer Garvin, Matthew H Samore, Wendy W Chapman, and Adi V Gundlapalli. 2009. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *BMC Bioinformatics* 10, S9 (2009), 1–11. https://doi.org/10.1186/1471-2105-10-s9-s12

[54] Parthasarathy Suryanarayanan, Ching-Huei Tsou, Ananya Poddar, Diwakar Mahajan, Bharath Dandala, Piyush Madan, Anshul Agrawal, Charles Wachira, Osebe Mogaka Samuel, Osnat Bar-Shira, and et al. 2021. AI-assisted tracking of worldwide non-pharmaceutical interventions for COVID-19. *Scientific Data* 8, 1 (2021), 94. https://doi.org/10.1038/s41597-021-00878-y

[55] Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and Expression Level Annotation of Opinions in User-Generated Discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, 575–584. https://aclanthology.org/P10-1059

[56] Houcemeddine Turki, Mohamed Hadj Taieb, and Mohamed Ben Aouicha. 2021. Developing intuitive and explainable algorithms through inspiration from Human Physiology and Computational Biology. *Briefings in Bioinformatics* 22, 5 (2021), bbab081. https://doi.org/10.1093/bib/bbab081

[57] Houcemeddine Turki, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha. 2021. How knowledge-driven class generalization affects classical machine learning algorithms for mono-label supervised classification. In *Proceedings of the 21st International Conference on Intelligent Systems Design and Applications*. Springer, Cham, Online.

[58] Houcemeddine Turki, Mohamed Ali Hadj Taieb, Thomas Shafee, Tiago Lubiana, Dariusz Jemielniak, Mohamed Ben Aouicha, Jose Emilio Labra Gayo, Eric A. Youngstrom, Mus'ab Banat, Diptanshu Das, and et al. 2022. Representing COVID-19 information in collaborative knowledge graphs: The case of Wikidata. *Semantic Web* 13, 2 (2022), 233–264. https://doi.org/10.3233/sw-210444

[59] Vyacheslav Tykhonov, Anton Polishko, Artur Kiulian, and Maksym Komar. 2020. CoronaWhy: Building a Distributed, Credible and Scalable Research and Data Infrastructure for Open Science. In *SciNLP workshop at AKBC 2020*. SciNLP, Online. https://doi.org/10.5281/zenodo.3922256

[60] Martin Volk, Bärbel Ripplinger, Špela Vintar, Paul Buitelaar, Diana Raileanu, and Bogdan Sacaleanu. 2002. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics* 67, 1-3 (2002), 97–112. https://doi.org/10.1016/s1386-5056(02)00058-8

[61] Andra Waagmeester, Egon L. Willighagen, Andrew I. Su, Martina Kutmon, Jose Emilio Labra Gayo, Daniel Fernández-Álvarez, Quentin Groom, Peter J. Schaap, Lisa M. Verhagen, Jasper J. Koehorst, and et al. 2021. A protocol for adding knowledge to wikidata: Aligning resources on human Coronaviruses. *BMC Biology* 19, 1 (2021), 12. https://doi.org/10.1186/s12915-020-00940-y

[62] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. Online, 417–426. https://doi.org/10.1145/3269206.3271739

[63] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online. https://aclanthology.org/2020.nlpcovid19-acl.1

[64] Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Ranran Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Yi Fung, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, Jasmine Rah, David Liem, Ahmed ELsayed, Martha Palmer, Clare Voss, Cynthia Schneider, and Boyan Onyshkevych. 2021. 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*. Association for Computational Linguistics, Online, 66–77. https://doi.org/10.18653/v1/2021.naacl-demos.8

[65] Xuan Wang, Xiangchen Song, Bangzheng Li, Kang Zhou, Qi Li, and Jiawei Han. 2020. Fine-Grained Named Entity Recognition with Distant Supervision in COVID-19 Literature. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 491–494. https://doi.org/10.1109/BIBM49941.2020.9313126

[66] Francis Wolinski. 2021. Systematic Extraction of Covid-19 Risk Factors and Vaccine Side Effects. In *SciNLP 2021: 2nd Workshop on Natural Language Processing for Scientific Text*. AKBC, Online. https://github.com/fran6w/vidar-19

[67] Tianzhi Wu, Erqiang Hu, Xijin Ge, and Guangchuang Yu. 2021. nCoV2019: An R package for studying the COVID-19 coronavirus pandemic. *PeerJ* 9 (2021), e11421. https://doi.org/10.7717/peerj.11421

[68] Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2145–2158. https://aclanthology.org/C18-1182

[69] Vikas Yadav, Rebecca Sharp, and Steven Bethard. 2018. Deep Affix Features Improve Neural Named Entity Recognizers. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, New Orleans, Louisiana, 167–172. https://doi.org/10.18653/v1/S18-2021

[70] Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics* 46, 6 (2013), 1088–1098. https://doi.org/10.1016/j.jbi.2013.08.004

[71] Yijia Zhang, Wei Zheng, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier. 2017. Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. *Bioinformatics* 34, 5 (2017), 828–835. https://doi.org/10.1093/bioinformatics/btx659

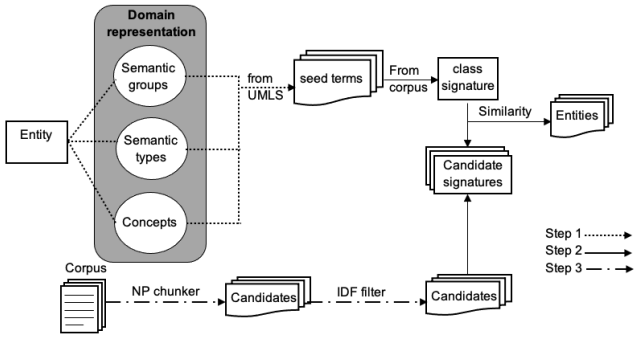

Figure S1: Overall approach to unsupervised biomedical named entity recognition

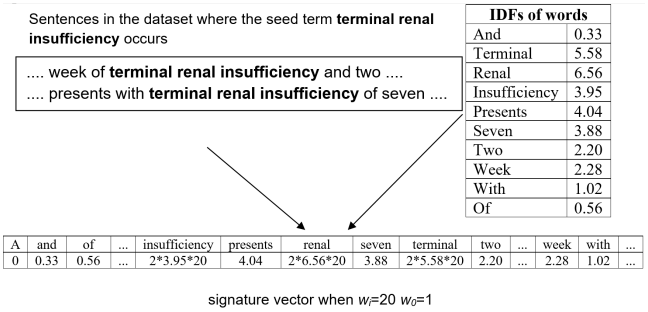

Figure S2: Building a signature vector for the seed term "terminal renal insufficiency" from IDF table and corpus, considering previous and following two words as well as internal words, assuming w_0=1, w_i=20

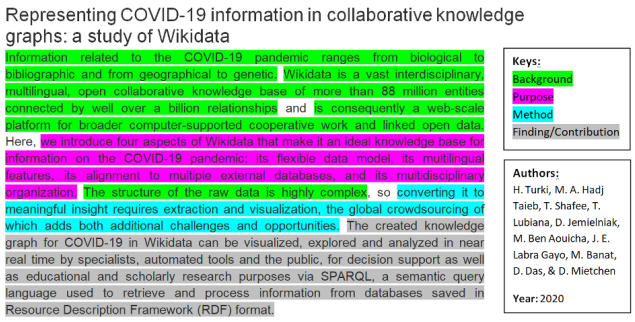

Figure S3: Sentence annotation of an abstract of a sample research publication