ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Annotating Scholarly Publications in the Biomedical Field Using Data Models: The Example of CORD-19

Vanapamula Veerabrahmachari[1], Arekatla Madhava Reddy[2], Shaik Guntur Mahabub Subhani[3], Gudipati Mohan Singh Yadav[4]
Assistant Professor[1,2,4], Associate Professor[3]
vveerabrahmachari@gmail.com[1] , amreddy2008@gmail.com[2],
subhanimehandi@gmail.com[3]gudipatimohan20@gmail.com[4]
Department of CSE, A.M. Reddy Memorial College of Engineering and Technology, Petlurivaripalem, Narasaraopet, Andhra Pradesh

Article Info

Received: 29-04-2022 *Revised: 18-05-2022* *Accepted: 28-05-2022*

ABSTRACT

Various computer applications, from knowledge graph building to biological question answering, have benefited greatly from semantic text annotations. As part of this comprehensive evaluation, present an evaluation of the data models used in semantic annotation initiatives for the academic articles included in the CORD-19 dataset, a publicly accessible collection of COVID-19-related article abstracts and complete texts. Using Google Scholar and manual screening, we compile a list of seventeen mostly American-authored papers on the subject. After that, we describe the inline semantic annotation models that are presently used on the whole texts of biomedical scientific papers. Then, we analyze the existing data models in light of the CORD- 19 dataset, suggesting promising new avenues for research and development in the field of semantic annotation models and projects.

1 INTRODUCTION

In the modern day, there is a constant ebb and flow to the body of academic writing due to the steady stream of new scientific facts and conclusions, especially those that pertain to the discovery of new areas of study. creation of cutting-edge investigation strategies [52]. Late in 2019, a new viral illness known as COVID-19 appeared, creating a global pandemic by March 2020 [27]. COVID-19 is characterized by acute respiratory symptoms and is caused by the SARS-CoV-2 virus. Every day, new academic articles arise that investigate different dimensions of this developing pandemic, from the infection's genetic and clinical features [13, 27] to its impact on microbiological safety [23]. Due to the rapidly changing patterns of the COVID-19 information involved in research outputs and to the increasing number of scholarly findings and evidence about the medical condition [13], the set of these scholarly publications is considered as big data distinguished by its volume, variety, velocity, and veracity [48]. Large amounts of unstructured data like those found on COVID-19 are challenging to handle without high-powered hardware and software [48], as well as intricate computational models founded on machine learning. Along with NLP methods [50]. However, computational approaches to explore, evaluate, and integrate COVID-19-related information in decision-support systems may be easily designed by using semantically organized representations of knowledge, such as Wiki data's COVID-19 Knowledge Graph [58, 61]. These semantic resources are helpful in many contexts, such as epidemiological evolution tracking [45], public health recommendation generation [20, 62], and Supporting various informative [21] or didactic [22] needs, because they can be manually explored by domain experts and automatically processed by computational methods. Semantic aspects of scientific knowledge must be extracted from textual resources like academic publications in order for knowledge graphs to be developed [28].

In-line semantic annotations of biological texts may be easily created by employing the procedure of finding and marking semantic content in raw texts [28]. The Allen Institute for Artificial Intelligence and other partner institutions have released an open dataset of scholarly publications about the infectious pandemic (so-called CORD-191) to spur the development of tools for the semantic annotation of COVID-19 information and, ultimately, the generation of an open COVID-19 knowledge graph [63].

Efforts have been undertaken to construct effective tools for the semantic annotation of COVID-19 since the creation of the CORD-19 dataset. Books and articles based on research. In order to sustainably enrich the CORD-19 dataset with annotations by human efforts and machines, several projects have been hosted on linked

data interfaces such as Open Link Virtuoso [37] and BRAT [14, 41] that make it possible to reuse and integrate the generated semantic information through intuitive APIs and federated SPARQL queries that can be used to identify and validate COVID-19 knowledge [26, 37].\ Semantic annotation initiatives in CORD-19 organized by data model (Table 1): Annotations for Named Entities (NE), Conceptual Relations (CR), and Activated Relations (AR) (AR), and Annotated Sentences (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

Studies and 348 search results from Google Scholar. Our systematic review has been limited to the seventeen articles that made it through the screening phase. Among the seven, 41.2% are connected to the Showcases of the W3C Semantic Annotation Projects; six of them (35.39%) were presented at COVID-19 Natural Language Processing (NLP) Workshops. We analyze the available information to identify the data models used in the semantic annotation of CORD-19 research publications and the eventual recipients of these efforts.

3 ANALYSES OF RETRIEVED PAPERS

Based on the data in Table 1, we can see that nine out of seventeen CORD-19 semantic annotation initiatives were conducted by American institutions. This is consistent with the existing state of U.S. computer science researchers have been at the forefront of the subject for many years; both in terms of output and scholarly recognition (see reference [51]). Among the other nations, Switzerland, Israel, and France led the pack with two publications apiece, while eight other countries each received coverage in 10 publications. To determine which data models are being utilized to express annotations (such named entity annotation); we examine the Methods sections of these papers. We also collect the reasons for creating inline CORD-19 semantic annotation projects (such as NLP and ML jobs) and derive how those reasons affect the construction of data models.

Models

Table 1 shows that there is a lot of interest in using named entity annotation (82.3%) on the CORD-19 dataset, which was discovered by looking at the data models used in published research. As will be shown, in large part because of the accessibility of annotation tools, pre-trained language models, and machine learning models that make it possible to carry out the work with very high degrees of accuracy [37]. In addition, as shown in Table 1, seven publications show an interest in creating links between semantically similar names. CORD-19 semantic annotation initiatives, organized by goal, are listed in Table 2.The acronyms KG, D, Q, H, RA, SE, QA, and RA stand for "Knowledge Graph," "Dashboard," "Research Analysis," "Search Engine," and "Text Summarization," respectively (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]		✓		

Employing a wide range of embedding methods and machine learning algorithms, transform entity annotations in academic texts into concept-based connection annotations. Concept-based relational analysis is widely used in Annotations, as the availability of named entity annotations makes it much easier to create such an annotation. Many studies rely on these two data models, although seven out of seventeen (41.2%) of the studies analyzed in the CORD-19 study attempted to advance the use of other forms of semantic annotation for the academic articles that make up the study's subject matter. Instead of using named entities as a foundation for semantic annotation development, five studies sought to annotate sentences directly.

Rather of assigning connection types and themes to individual words, annotations will be applied as labels to whole sentences, reducing the complexity of the necessary software. In addition, two papers have looked at CORD-19 annotation using Action-Based Relation Annotation as a data model. Researchers attempted to determine connection types by annotating text spans containing phrasal verbs that stood in for them in phrases. The purpose of this annotation is to limit the CORD-19 semantic annotation to to those relation types that are considered "generic" [24, 41].

Targets

We abstracted the reasoning of the investigated annotation projects to show why four distinct data models are employed for CORD-19 semantic annotations. Nine out of seventeen works were discovered by us. Knowledge graphs regarding COVID-19 may be constructed from the studied academic papers if annotators utilize named entity annotation in conjunction with concept based or action based connection annotation, as illustrated in Table 2. The combination might also be used for other, less significant purposes, such as powering CORD-19 search engines or analyzing COVID-19 research findings. Sentence annotation, together with named entity annotation, is utilized to power pandemic-related question answering systems.

Long-term use of natural language texts within the context of the TREC programs has fostered this kind of sentence annotation. Natural language writings are human-readable and may include information that aren't necessarily reflected in fully-structured knowledge graphs [19], making them ideal for use in answering queries. A number of studies (four out of seventeen, or 23.5%) employ both named entity-based and sentence-based annotations to feed COVID-19 dashboards that visualize features of the COVID-19 pandemic and illness as disclosed in scientific publications. One explanation for this is that knowledge graphs, especially those in the Resource Description Framework (RDF) Format, are very amenable to feature extraction utilizing a wide range of tools, including as application programming interfaces (APIs) and query languages like SPARQL [58]. However, the accessibility of open-source analytics tools, especially Python and R packages, that generate quality visualizations from a processed input, allowed the development of real-time human-friendly graphical representations of structured information, such as the CORD-19 semantic annotations [67].

4 DATA MODELS

Linked data formats including RDF, XML, and JSON are used by CORD-19 semantic annotation projects to express in-line semantic information [37], with a primary focus on text span annotations. To glean sentence-

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

level [25] and entity-level [35] semantic characteristics, as seen in Fig. 1. By using either human labour [41] or fine-grained annotation automation systems like PubTator6, SciSpacy7, DBpedia Spotlight8, Entity-fishing9, NCBO Bio Portal Annotator10, and Annotator+11 [37, 64], annotators create text span annotations and align them to external resources. Later, comparable semantic annotations are added using deep learning methods like convolution neural networks and Long Short-Term Memory and language models like BERT, Elmo, and GloVe [17, 25, and 35]. These comments may be limited to only pointing out key ideas or words within a passage [41, 59], or they can be broadened to annotate the categories of identified objects [25, 37]. However, the

Figure 1: Graph embeddings and other types of text span annotation models may be utilized to find the relationships between annotated named entities for the purpose of building knowledge automatically. Several initiatives have opted to undertake inline annotations of semantic relations in order to guarantee the verifiability of produced statements based on user contributions and strong computer techniques [24, 41], in particular in the context of the CORD-19 Research Dataset [12, 26, 46, 64]. These relation annotations may be used as a standard for an explainable and more trustworthy machine learning-based retrieval of biological and clinical semantic connections [71], provided they are evaluated by a panel of clinical professionals.

Data models for annotating biological interactions include, as illustrated in [46], concept-based annotation models, action-based annotation models, and phrase annotation models. In Figure 2, we see how the concept-based relation annotation models connect between annotated concepts using relation annotations, where the property is a non-taxonomic (biomedical) or taxonomic (generic or temporal) relation type [46]. Using a small set of generic features, action-based connection annotation models link words with concepts that correspond to the evocative relation type [24, 41]. In the sentence relation annotation models, a string of explanatory text or a piece of semantic information [37] is associated with a sentence that is annotated as a text span. In this part, we will describe the various data models used in the text span annotation and the semantic connection annotation of biomedical texts in the context of the pandemic.

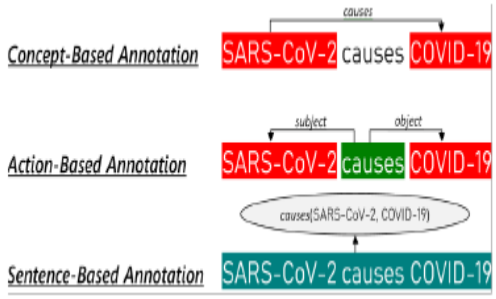

Figure 2: Types of semantic relation annotation models

Named entity annotation

Annotating names with additional information is called named entity annotation [14]. Annotations for named entities provide information about the start and end of a Figure 3 illustrates an annotated text span, labels, and external identifiers that relate the annotation to objects in Wiki data and Wikipedia [37, 64]. Concepts in a phrase are identified using Named Entity Recognition (NER), which then places them in one of many predefined categories (such as "illness," "medicine," "entity," "concept," etc.) [39].

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Figure 3 shows the XML structured representation of named entity annotations in a COVID-19 extract.
In general, NER systems may be broken down into the following types: I Knowledge-based NER
Systems that depend on specialized resources and domain-specific knowledge instead of labelled training data. Using semantic resources like the UMLS Met thesaurus and "signature" similarity classification, Figure S1 describes an unsupervised approach to NER in a biomedical setting [70]. (ii)Unsupervised systems, which do not have human trainers and hence need training data conveying properties such as orthography (e.g. capitalization), entity context, words included inside named entities, etc. (iii) Predictive models built using supervised learning, where inputs and predicted outputs are trained together to provide accurate results. (iv) Neural network designs for NER that use feature-inferring systems, complete with feature vectors and word embedding models that take into account various degrees of granularity (from words to characters) [68–70].
To align the entities referenced in a text with reference to a knowledge base, such the Unified Medical Language System (UMLS) in the biomedical sector, EL is closely connected to NER [49]. Entity links help in reclassifying entities into more accurate categories. By investigating methods for executing NER and EL together, we may improve entity type categorization and entity linking, with the added bonus of reducing error propagation [14]. Knowledge graph-based semantic similarity measures and word embeddings may be used to evaluate the annotated concepts' semantic properties and confirm the accuracy of their classification [31].
There are two perspectives on NER as an annotation target: By "multiclass classification," we imply a classification problem having more than two classes, such as "classifying" a series of fruit photos as "oranges," "grapefruit," "kiwis," etc. You may have your pick of apples and pears. It is assumed in single-class classification, for example, that a sample may be an apple or a pear, but never both. And for an end-user app like Disease and Symptom, the degree of precision of the class expressing the notion is crucial. In multi-label categorization, many desired labels are applied to each sample. This may be seen as anticipating characteristics of a data item that are not exclusive of one another, such as potential document-related themes. Religion, politics, economics, and/or education may all be present in the same work, or they can all be completely unrelated. However, researchers have provided few examples of multi-labelled categorization NER annotation approaches [39]. An annotation granularity is sought for in the process of annotation. This indicates that the annotation makes an effort to undertake a fine-grained study of the texts [15]. Multiple degrees of meaning will be assigned to the phrase "terminal renal insufficiency." As a first step, the medical condition terminal renal insufficiency will be annotated using the full word, which is the one that comes closest to the UMLS entry. Renal insufficiency and sufficiency will both be marked as Medical Condition for even more specificity.
Then, in the boundary detection phase, candidates for entity categorization are gathered by identifying the borders between entities. Getting rid of dummy noun phrases that refer to nonexistent things to filter candidates produced by the NP chunkier [39], researchers use an IDF-based method. Entity categorization, the last phase, seeks to acquire entities of the same class, with the expectation that they will have a common lexicon and context. For instance, the term "insufficiency" is more likely to be included inside an object of type "Problem" than "Treatment" or "Test" [39]. Compound nouns, such as "renal insufficiency," are identified utilizing targeted resources, such as the UMLS, that are grounded on MWE principles. The technique based on similarities takes use of the distributional meaning of data. Signature creation is shown in Figure S2 as a vector of internal and context words for a certain item.
Annotating relationships between concepts Expanded named entity annotations are used to bring similar ideas together to provide a structured annotation in a variety of contexts. Relationships between concepts as given in the studied passage. Concept-based relation annotation (sometimes referred to as "CBRA"; [24, 40]) is the name given to this kind of annotation. Named entities are exclusively represented as text span annotations in the design of concept-based connection annotation models [46]. The recognized relations between concepts, the references of the annotated relations, and the alignment information of the annotations are all provided as structured metadata of the annotations [64]. Here, the named entities representing their subjects and objects are assigned to the relations according to their types (for example, direct up-regulation and indirect up-regulation for biological processes), and the relations are displayed as arrows connecting ideas in user-friendly interfaces [46]. In order for the annotation process to function effectively, this means that the annotation data model must include all possible forms of biological connections as a separate category [46]. Like the Wiki data knowledge graph, concept-based relation annotation uses triples (so-called reification) to represent attributes and references as qualifiers, with the subject being the original relation, the property being the type of the attribute or reference,

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

and the object being a value or an annotated named entity [30, 58, 61]. Data modellers make decisions about which pieces of information to use as qualifiers in their annotations based on the needs of their projects [30]. Figure 4 depicts such a case.

Multiple factors and limitations influence the selection of classes to represent relations or named things in either style of annotation. Dependent on, the material being annotated (for example, genetic data or clinical information) and why it is being annotated (for example, knowledge graphs generation or summary of biomedical writing). While some projects opt to just annotate certain kinds of ideas and relations, such biological processes [24], others prefer to be more comprehensive, including many facets of scientific and clinical information [60]. Figure 4: Example semantic annotation utilizing concept-based annotation model. Furthermore, annotation projects can select a level of abstraction to label the annotations with either broad categories (i.e., the hyponym of the type of annotation, such as concept, direct mechanism, and indirect mechanism) or fine-grained categories (i.e., the hyponym of the type of annotation, such as non-drug symptomatic treatment) describing the characteristics of the annotated named entity or relation beyond the annotation type [24, 6].
To further guarantee an unambiguous representation of semantic annotations, annotation initiatives strive to minimize overlap across target categories. Both the use and interpretation of semantic annotations [10] and [11] are somewhat murky, although not by much. To eliminate representational mistakes and duplication [10, 11], this means getting rid of inverse qualities and closely related categories that are hard to tell apart using semantic similarity measures and word embeddings [31]. An important component in the efficacy of machine learning semantic annotation algorithms [57] is the selection of categories, which should reveal the degree of difficulty in the categorization of semantic linkages in biomedical texts.

Sentence annotation

As was made abundantly evident above, sentence annotation models annotate every sentence in the evaluated excerpt in a single text span [25]. Each sentence annotation is quite similar to named entity annotations in that it begins at a certain point, ends at a specific point, and is given a specific name [18]. When using simple annotation tools like Hypothesis (https://web.hypothes.is/) [8] and when the sentence annotations are coupled with named entity annotations to enable better semantic link prediction by only considering named entities included in the same sentences [59], the label need only mention that the text span annotation corresponds to a sentence. The tag may also represent a group inside a taxonomy that characterizes some grammatical or pragmatic aspect of the text [25].

The Annotation of Relations through Action Subject-Action-Target (SA-T) triplets is the building blocks of the semantic structure of sentences in action-based annotation models. The Action is a grammatical function that may be filled by any word or phrase that describes an event [24, 42]. It is most often seen in verbal constructions, although it is not limited to them. The Subject specifies the agents carrying out the operation, whereas the Target details the intended recipients. Figure 5 depicts the SAT+R [42] annotation model as one example.

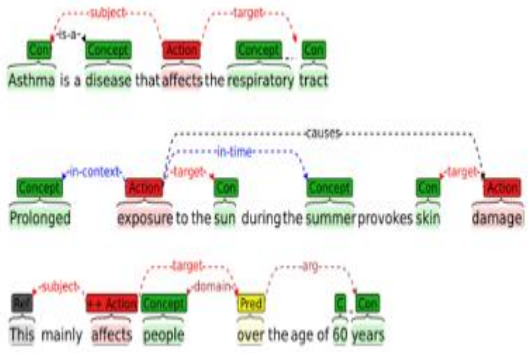

Figure 5: Semantic Annotations Employing the SAT+R Algorithm
With the SAT+R paradigm, you may annotate a phrase on many semantic levels. To begin, the Most Important Ideas and The wording clearly identify the actions. Both the subject and the object relations may be used to associate an action with a notion; their semantic significance has already been discussed. As an added bonus, composite ideas may be annotated by connecting actions together. Reference, which shows a missing reference, and Predicate, which enables the generation of ideas based on filters or characteristics, are two more entities defined. Each predicate has its own domain and set of arguments that specify what the predicate means and what

other semantic aspects are included in the analysis. Figure 5's concluding phrase has both types of annotations. However, the function of predicates in RDF (which is really connected to Action in SAT+R) should not be mistaken with the role of predicates in SAT+R.

In SAT+R, predicates are used to build complicated ideas by tacking on a set of conditions to a simple idea. In the third statement, for instance, Figure 5 presents an argument of 60 years to qualify the noun people by the predicate over. This results in a composite idea that stands in for those who are sixty years of age or older, with "over" being defined contextually. Then, the conjunction affects, with over as its target, specifies that the group of persons impacted by the action's subject is limited to those who also meet the predicate's criteria. Additionally, the model specifies the standard semantics for the four ontological relations of "is-a", "part-of", "has-property", and "same-as". Any two of the four possible entity kinds may be linked using these relations.

Their goal is to provide context for structural components as they may be described in a taxonomy or ontology. Causes and entails are two more relations, and they map onto the ideas of causality and logical inference, respectively. While entailment is a logical link that does not imply causation, causality needs an explicit method whereby one thought or action affects the causation of another. Further, three contextual relations—intimae, in place, and in context—are specified, allowing events or actions to be described that are contingent upon the presence or occurrence of other ideas or activities. The concept concludes by defining four characteristics: highlighted, reduced, unknown, and negated. Both augmentatives and diminutives have grammatical purposes, and the first two allow for their capture without the use of affixes. Repeated use of a certain annotation for adverbs or other grammatical components. The third one permits the annotation of negation, while the fourth one permits the definition of doubtful occurrences. In the previous line, the word affects, which is modified by the adverb principally, serves as an example of a stressed notion. The accentuated attribute on affects symbolizes this idea of emphasis by distancing it from the superficial form of the adverb or adjective it modifies. Words like "sometimes" or "maybe" may also be used to infer the ambiguous attribute being applied to a concept or an action, and thus serves to disentangle the semantic meaning from the text's surface form.

DISCUSSION

We observed that a significant fraction of COVID-19-related semantic annotation efforts rely on human-generated annotations or expert-level understanding to produce their final products. Charts, notably Wiki data as seen in Chapter 4. The availability of manual semantic annotation projects for CORD-19 datasets can be explained by the usefulness of these annotations to provide more accuracy to information retrieval tasks [6], despite the possibility of automatic retrieval of COVID-19 information from scholarly publications using semantic embeddings and machine learning. Using humanly curate semantic resources such as annotated datasets and biological ontology’s to power computer applications might pave the way for more trustworthy output that does not contradict with human knowledge and may provide light on the limitations of automated annotation methods [56].

Combining the results from Sections 3 and 4 with those from the CORD-19 annotation projects (especially the F-measures) gives a comprehensive picture of the annotation effort. The effectiveness of semantic annotation models and methodologies are examined in detail in this article's Section 4. When contrasting human and automatic semantic annotations, it becomes apparent that knowledge resources-based systems12 utilizing word embeddings and neural networks [17, 24, and 35] are more effective than human annotation efforts and automatic annotation methods driven by a corpus of manual annotations [41]. Named entity annotations are affected, while action text span annotations are much more so [24, 41]. However, when comparing the action-based relation annotation projects of the CORD-19 scholarly publications13, it is evident that the action-based relation annotation model considering subject and object as the only relation types [24] allows a higher degree of machine learning accuracy than the ones using the SAT+R annotation model [41].

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

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

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

A lack of a comprehensive specification of the annotation granularity explains why minimal accuracy has been achieved for named entity annotation, especially in the context of human annotations.

Resulting in significant variations in how annotated text spans are handled by people and machines [53]. Since most subjects and objects of annotated sentences are comprised of more than one word, [53] this issue arises often in biological natural language processing. Noun phrases are often the subjects and objects of sentences, as seen by the examples in S1–S4 [38]. The nouns in these phrases may or may not be preceded by an article (for example, the pathogenesis [S1], a symptom [S2], 14 days [S4]), a preposition (for example, the pathogenesis of COVID-19 [S1], a symptom of COVID-19 disease [S2], the development of vaccines by firms [S3], the maximal incubation period for COVID-19 [S4]), an adjective (for example, mo There are two primary considerations at play here. In Natural Language Processing, articles, prepositions, and conjunctions are regarded as stop words that should be ignored in topic modelling and other engaging tasks [1]. The large discrepancies across human experts in annotating named items in CORD-19 academic articles [41] may be traced back to a failure to properly account for such noun phrase elements in the text span annotations.

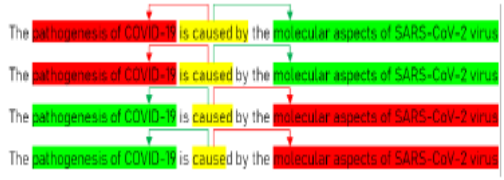

Figure 6: Examples of action-based semantic annotations for S1: Action (Yellow), Subject (Red), and Object (Green).

6 CONCLUSIONS

In this comprehensive evaluation, we described the many types of data models that were used in the annotation of the CORD-19 papers. We described the current knowledge in this area and indicated the data models for annotating medical journals, especially those published in the wake of the 2009 COVID-19 epidemic. We outlined the benefits of each semantic annotation data architecture and examined the main issues limiting their actual effectiveness, such as the granularity of text span annotations and the assignment of relation types and attributes. Fixing such issues would strengthen knowledge-based systems and improve the precision of semantic annotation efforts. In order to better explain the research dynamics behind semantic annotation projects for the CORD-19 dataset [2], we propose expanding our work to include additional bibliographic databases like Web of Science, Pub Med, and Scopus, and applying visualization tools like Bibliometrix on the bibliographic metadata of the considered scholarly evidences. In addition, we suggest establishing comprehensive rules for uniformly annotating textual materials, taking into account the constraints to improve completely automated semantic annotation techniques; we need to provide a machine-readable version of the rules behind these models for semantic annotation data.

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

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

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/bunticgl4.pdf