applied sciences

Article

A Methodology for Open Information Extraction and Representation from Large Scientific Corpora: The CORD-19 Data Exploration Use Case

Dimitris Papadopoulos [1,2,*], Nikolaos Papadakis [3] and Antonis Litke [3]

[1] Technical University of Crete, Department of Production Engineering and Management, 73100 Chania, Greece
[2] Hellenic Army Academy, 16672 Vari, Greece
[3] Infili Technologies P.C., 15772 Athens, Greece; npapadakis@infili.com (N.P.); alitke@infili.com (A.L.)
* Correspondence: dpapadopoulos6@isc.tuc.gr; Tel.: +30-210-8904000

Received: 26 July 2020; Accepted: 8 August 2020; Published: 13 August 2020

Featured Application: Open Information Extraction on the COVID-19 Open Research Dataset (CORD-19).

Abstract: The usefulness of automated information extraction tools in generating structured knowledge from unstructured and semi-structured machine-readable documents is limited by challenges related to the variety and intricacy of the targeted entities, the complex linguistic features of heterogeneous corpora, and the computational availability for readily scaling to large amounts of text. In this paper, we argue that the redundancy and ambiguity of subject–predicate–object (SPO) triples in open information extraction systems has to be treated as an equally important step in order to ensure the quality and preciseness of generated triples. To this end, we propose a pipeline approach for information extraction from large corpora, encompassing a series of natural language processing tasks. Our methodology consists of four steps: i. in-place coreference resolution, ii. extractive text summarization, iii. parallel triple extraction, and iv. entity enrichment and graph representation. We manifest our methodology on a large medical dataset (CORD-19), relying on state-of-the-art tools to fulfil the aforementioned steps and extract triples that are subsequently mapped to a comprehensive ontology of biomedical concepts. We evaluate the effectiveness of our information extraction method by comparing it in terms of precision, recall, and F1-score with state-of-the-art OIE engines and demonstrate its capabilities on a set of data exploration tasks.

Keywords: information extraction; triple extraction; bioinformatics; data mining

1. Introduction

Open information extraction (OIE) systems aim at distilling structured representations of information from natural language text, usually in the form of triples or n-ary propositions. Contrary to ontology-based information extraction (OBIE) systems which rely on pre-defined ontology schemas, OIE systems follow a relation-independent extraction paradigm tailored to massive and heterogeneous corpora. Therefore, they can play a key role in many NLP (natural language processing) applications like natural understanding and knowledge base construction by extracting phrases that indicate novel semantic relationships between entities. Although there are many approaches for extracting triples in the form of {subject, predicate, object} from unstructured text, there is no standardized way of efficiently generating, mapping, and representing these triples in a manner that facilitates end-user applications. These limitations are primarily prevalent in larger corpora, where the high number of duplicate and/or low-quality triples as a result of topic irrelevant sentences or complex syntactic phenomena (e.g., coreference) further hinders robust triple extraction, ultimately discouraging their extensive deployment.

The goals of this paper are twofold: first, to present a methodology for efficiently extracting information from large corpora, covering all phases from natural language text pre-processing to triple extraction, intuitive visualization, and querying; and second, to concretize this methodology on a set of downstream tasks relying on state-of-the-art tools and pretrained deep learning models to demonstrate its effectiveness in a real-world scenario involving the CORD-19 dataset, which represents the most extensive machine-readable coronavirus-related collection of literature available for data mining to date.

2. Related Work

There is an abundance of proposed strategies for transforming raw text to a structured representation in order to populate a knowledge graph [1–4]. However, especially in the case of OIE approaches and due to concerns on scaling, the use of syntactic or semantic relation extraction techniques has been relatively sparse, with the exception of a few recent examples aiming at domain-specific knowledge extraction [5–9]. Most domain-specific information extraction approaches are focused primarily on evaluating the efficiency of different triple extraction tools on raw data, not taking useful pre-processing and post-processing strategies into account, thus resulting in a large number of potentially uninformative triples [10–12]. There exist a few systems that go beyond triple extraction by implementing a more thorough preprocessing strategy, including coreference resolution or discourse analysis to improve the quality of the extracted triples; however, these do not address the scalability issues that arise from processing large corpora [13,14]. By treating each sentence in the corpus equally, we run the risk of overflowing the graph database with unrelated information compared to the documents' true scope, seriously impeding data exploration tasks. On the other hand, by complying only with a strictly defined ontology schema, we are likely to lose all information that is not covered by the existing ontology properties [15]. Finally, the available bibliography lacks a clear triple representation strategy that would resolve duplication issues and would equip the user with a set of data enrichment processes for visualizing latent information such as the temporal dimension (continuity) of the extracted triples, connections to existing ontologies (entity linking), sentence polarity, and hidden interconnections between different corpora based on similar extracted entities. Our methodology encompasses a number of pre-processing (coreference resolution, text summarization) and post-processing (entity enrichment, graph representation) tasks coupled with a core parallel triple extraction process, combining different approaches to enhance the contextual connectivity of the extracted information.

2.1. Advances in Coreference Resolution

Coreference resolution is the task of finding and grouping all expressions (mentions) that refer to the same entity in a text. Two noun phrases are said to be co-referring to each other if both of them unambiguously resolve to a unique referent. In many cases, the term "coreference" is used interchangeably with the term "anaphora", denoting the non-symmetric syntactic phenomenon of a noun phrase being the anaphoric antecedent of a another noun phrase (i.e., only the former is required for the interpretation of the other) [16]. A key challenge of coreference resolution is that entity information may be spread across multiple mentions over the corpus, thus requiring information to be aggregated from all mentions [17]. Over the last decades, several approaches in tackling coreference problems have emerged, spanning from early, rule-based, and linguistically-motivated approaches [18,19] which are based on the syntactic constraints of the language, to modern deep learning techniques that rely on pairwise scoring of entity mentions [20–22]. The latest research in the field leverages finetuning of existing state-of-the-art language models (e.g., span-based pretraining of BERT models) which are repurposed for the task of coreference resolution [17,23].

2.2. Advances in Text Summarization

Automatic text summarization is the process of computationally shortening a set of data to create a subset that represents the most important information within the original content [24]. Summarization is considered one of the most increasingly demanded tasks in natural language processing, as a means to unlock the abundance of wealth hidden underneath the vastness of textual data. At present, there exist two main methods for text summarization: abstractive and extractive [25]. In abstractive summarization, the summary is generated by novel sentences paraphrasing existing words, while in extractive summarization, the content is composed by unmodified sentences from the original text.

Recent work in abstractive summarization involves the use of sequence-to-sequence frameworks based on attentional recurrent neural network (RNN) encoder-decoder [26–28] or transformer [29,30] architectures to generate concise summaries from input documents. With regard to extractive summarization, proposed approaches employ lexical features [31,32], statistical methods such as TF–IDF [33,34], or unsupervised learning techniques [35] to extract keywords and phrases from large corpora, with the most recent ones also involving transformer architectures [36]. While implementations following the abstractive approach are more closely emulating human summarization, even those based on ANNs (and considered state-of-the-art) are relying on large training corpora, have limited generalization on the document level, and usually suffer from semantic and grammatical errors [37]. Extractive summarization, on the other hand, has reached its maturity stage and—although most extractive summaries may lack in readability and cohesion—they generally succeed at capturing the key points of the digested text [38,39]. At the same time, since they just highlight portions of the original content, there is no danger of generating sentences with irrelevant/wrong interpretations, which is especially important in sensitive domains like biology or medicine.

2.3. Advances in Open Information Extraction

Open information extraction (OIE) systems aim at converting the unstructured information expressed in natural language into a more structured representation, in the form of relational triples consisting of a set of arguments (subject, object) and their semantic relation (predicate), e.g., <subject, predicate, object> [40]. Unlike closed information extraction approaches that are limited to a narrow set of predefined target relations, OIE systems are able to extract any kind of relation, providing increased scalability and usability over heterogeneous corpora [41]. In order to extract OIE triples, most approaches try to identify linguistic extraction patterns, which may be either hand-crafted or automatically learned from annotated data. Rule-based approaches that rely on hand-crafted extraction rules focus on syntactic constraints expressed as part-of-speech (POS)-based regular expressions [42,43]. Self-supervised learning approaches usually leverage annotated data sources (e.g., Wikipedia infoboxes) to train classifiers [44,45] or bootstrap a large training set over which they learn a set of extraction POS pattern templates [46]. Some recent OIE systems are clause-based, using linguistic knowledge about the grammatical and syntactic properties of the language to identify clause constituents, thus restructuring larger complex sentences to many simple ones [47,48]. The emergence of annotated corpora for OIE evaluation paved the way for supervised neural-based models that further pushed the state-of-the-art in this domain [49,50]. Latest approaches are extending the use of deep BIO taggers used for semantic role labeling (where B is assigned to the beginning of named entities, I is assigned to the interior, and O is assigned to other) by leveraging the word embeddings of the processed sentences in deep neural networks (e.g., bi-LSTMs) to produce probability distributions over possible BIO tags [51,52].

2.4. Advances in Entity Linking, Enrichment and Representation

Over the years, various entity enrichment approaches have been introduced, aiming at augmenting the usefulness of the extracted triples, including entity linking and polarity detection processes before representing them through a graph database. Linking the identified entity mentions in text to an ontology or dictionary is considered an essential step in creating informative triples, with various knowledge bases (KBs) pertaining to general knowledge being employed for this purpose based on morphological similarity, such as DBpedia [53], Freebase [54], and Yago [55]. There are cases, however, where many of the entities in these general-interest KBs are irrelevant for certain applications, therefore domain-specific ontologies for semantic information brokering, based on inter-ontology relationships such as synonyms, hyponyms, and hypernyms of the extracted entities are used [56]. In order to further increase the links between morphologically dissimilar extracted entities and KB-related objects, neural-based methods are also implemented, exploiting word embeddings to represent semantic spaces [57,58], also allowing for domain-agnostic entity resolution [59]. With regard to sentiment analysis (neutral vs. emotionally loaded) and polarity (positive vs. negative) detection of a text [60], lexicon-based [61], ML-based [62], and neural-based [63] classifiers are commonly used to identify the polarity of a relation within a sentence. In the field of graph representation, the two main implementations of graph models include resource description framework (RDF) triple stores [64] and labeled property graphs (LPG) [65]; both provide ways to explore and graphically depict connected data.

3. Materials and Methods

Our proposed methodology introduces a processing pipeline that takes as input a large natural language corpus consisting of many documents (e.g., scientific articles) and provides a structured representation of the extracted information in the form of open information triples as output, allowing for interactive data exploration. The system comprises the following components:

1. an in-place neural coreference resolution component,
2. an extractive text summarizer that isolates the key points of the ingested text,
3. a parallel triple extraction component as our core information extraction method, and
4. a toolkit of entity enrichment and representation techniques built around a graph engine.

More information about the data used to demonstrate our methodology and the technical specifications of each component are given at Sections 3.1 and 3.2, respectively. An overview of our pipeline is depicted in Figure 1.

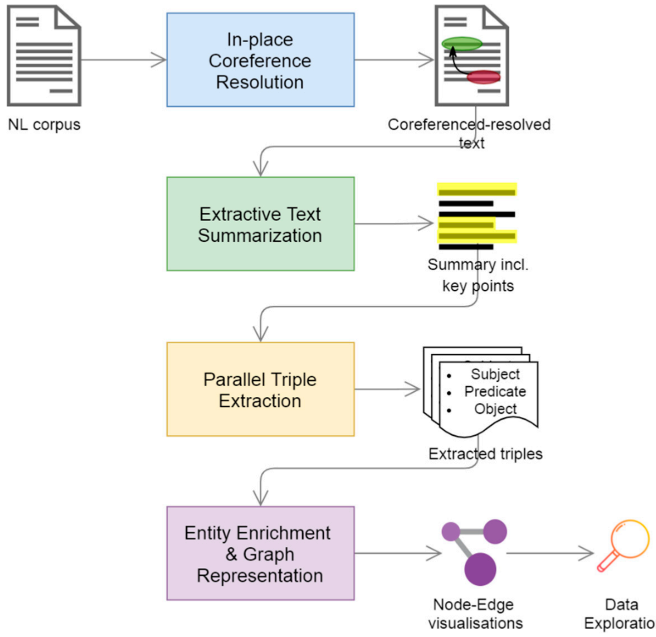
NL corpus
In-place Coreference Resolution
Coreferenced-resolved text
Extractive Text Summarization
Summary incl. key points
Parallel Triple Extraction
Subject
Predicate
Object
Extracted triples
Entity Enrichment & Graph Representation
Node-Edge visualisations
Data Exploration

Figure 1. Overview of our four-step information extraction pipeline.

3.1. Data

For the purposes of our work and in response to the recent COVID-19 pandemic, we leveraged the COVID-19 Open Research Dataset (CORD-19) [66], provided by the Allen Institute for AI. The dataset consists of different subsets for commercial and non-commercial usage, collectively including over 40,000 full-text articles about the coronavirus family of viruses, to be used by the global research community. We specifically focused on the "Commercial use subset" that by the time of access (8 April 2020) contained 9365 articles from PubMed Central (PMC), a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine. Each paper is represented as a single JSON object, whose schema contains a unique paper ID, the paper's title and authors list, an abstract, the main body text, and its corresponding bibliographic entries. The dataset is constantly updated, aiming at facilitating the development of text mining and information retrieval systems over its rich collection of metadata and structured full text papers, in support of the fight against the COVID-19 disease.

3.2. Implementation

The following subsections (Sections 3.2.1–3.2.4) are devoted to each of the four steps of our proposed methodology. Each subsection follows the same pattern; a brief description of the process followed by a detailed analysis of the technical implementation. Finally, a number of examples based on the CORD-19 dataset are given to illustrate how this pipeline may be operationalized for other real-world scenarios.

3.2.1. In-Place Coreference Resolution

Given that our information retrieval task requires the extraction of dependency relations from sentences, i.e., sets of the form {subject, predicate, object}, and that in many cases the entity is replaced with its coreferential pronoun (e.g., "Mary is a nice person, I like hanging out with *her*" rather than "Mary is a nice person, I like hanging out with *Mary*"), we consider in-place coreference resolution as a crucial pre-processing step on the each article's body text, to improve the quality of the extracted triples. Therefore, in the scope of creating our information extraction pipeline, we leveraged the pretrained neural coreference resolution tool from AllenNLP [67], which implements a variant of Lee et al. end-to-end coreference resolution model [68] using Span-BERT embeddings [23]. The model had been trained on the OntoNotes 5.0 dataset (the largest coreference annotated corpus) [69], achieving F1-score of 78.87% on the test set.

Each article of the CORD-19 dataset was pre-processed by the in-place coreference resolution component, where all noun phrases (mentions) referring to the same entity were substituted with that entity. The pretrained neural model from AllenNLP provided good results even on complex situations containing challenging pronoun disambiguation problems, thus facilitating the creation of more informative triples. Indicative examples of the performed coreference substitutions on article extracts are provided in Table 1.

Table 1. In-place coreference resolution on CORD-19 dataset extracts. The patterns of coreference are annotated with subscripts. The anaphors (orange) are replaced by the antecedent to which they refer to (green).

Article ID	Extract
42e321eedb a25d380ae4 4d16cdf0bb deab83d665	Two articles in the top ten cited articles discussed the emergence of New Delhi **metallo-β-lactamase (NDM) gene**$_{\{1\}}$ responsible for carbapenem resistance. ~~**This gene**~~$_{\{1\}}$ **metallo-β-lactamase (NDM) gene**$_{\{1\}}$ belongs to carbapenemase gene family and **bacteria carrying metallo-β-lactamase (NDM) gene**$_{\{2\}}$ are referred to as superbugs because ~~**they**~~$_{\{2\}}$ **bacteria carrying metallo-β-lactamase (NDM) gene**$_{\{2\}}$ are resistant to most antibiotics.
85eb641e06 b0d6b1a0b2 02275add0c 5d27e53d71	**Official health linkages**$_{\{1\}}$ have served to promote good will in some otherwise difficult relationships, as has been the case with Indonesia. ~~**They**~~$_{\{1\}}$ **Official health linkages**$_{\{1\}}$ have also helped to promote a positive international image for Australia.
4c84dbfd01 f7b2009ebe d54376da8 afcbcf1ec64	However, one should also note that the experiment is based on labelling and quantifying proteins about 4 h post-infection. **This relatively early time point**$_{\{1\}}$ allows one to minimize potentially confounding influences of virion particle assembly and production on cytoplasmic levels of viral proteins, but ~~**it**~~$_{\{1\}}$**This relatively early time point**$_{\{1\}}$ also represents a point before the majority of viral proteins have been made.
2c5d1ebec4 04ad8061eb 81e94effbe5 2a6dbe809	**Purified ALV-A virus particles**$_{\{1\}}$ were incubated with PMB for 30 min at 378C, and infectivity was measured on human 293 cells expressing **the ALV receptor Tva (293-Tva)**$_{\{2\}}$. The effect of PMB treatment of ~~**these particles**~~$_{\{1\}}$ **Purified ALV-A virus particles**$_{\{1\}}$ was comparable to native MLV particles. These findings suggest that ~~**Tva**~~$_{\{2\}}$ **the ALV receptor Tva (293-Tva)**$_{\{2\}}$ binding creates or exposes a functionally important cysteine thiolate target for PMB in ALV-A Env.
d9eb8ffffee8 147c850b00f 613a1978c18 505580	**FCoVs and CCoVs**$_{\{1\}}$ are common pathogens and readily evolve. It is necessary to pursue epidemiological surveillance of ~~**these viruses**~~$_{\{2\}}$ **FCoVs and CCoVs**$_{\{1\}}$, so as to detect the emergence of new variants, which may have increased pathogenicity and/or a new host range, as early as possible.

3.2.2. Extractive Text Summarization

We consider text summarization as the second step of our information extraction pipeline for two reasons: i. extracting triples directly from large corpora (such as the CORD-19 dataset) is both costly and time-inefficient due to the increased computational resources required by information extraction engines, and ii. in most cases, only a small fraction of the extracted triples is useful and/or relevant to the discussed topic. Relying on the articles' abstracts—whenever applicable—is a reasonable alternative; however, we run the risk of processing text that contains short descriptions of the article's purpose, along with broad, high-level specifications of little contextual value. Hence, we argue that an extractive summarization method is the optimal way to reduce a document's text length by omitting peripheral or inappropriate information, while highlighting its key features for triple extraction. To this end, we relied on the HuggingFace extractive summarizer library based on Miller's transformer-based approach [36]. For our implementation, and since we were dealing mainly with biomedical articles, we substituted the vanilla BERT model with AllenAI's SciBERT pretrained model for scientific text, which significantly outperforms BERT-base on most NLP-related tasks including named entity recognition, sequence labeling and text classification [70].

We iteratively passed the CORD-19 articles (previously submitted to in-place coreference resolution) through the summarizer, where all sentences were embedded into the multi-dimensional space using SciBERT embeddings. Subsequently, k-means clustering was used on the sentence representations to identify those sentences that were closest to the cluster's centroids for summary selection. The ratio of sentences comprising each summary to the original text was provided as parameter (0.2) to the model. The result was an extractive summary for each one of the 9365 ingested articles, with an average count of 843 words per summary, which resulted in a significantly more condensed corpus, compared to an average count of 4482 words per original body text. An example of the text summarization process is provided in Table A1 of the Appendix A, while detailed information on the word reduction per article is depicted in Figure 2 below. The generated summaries, along with respective metadata (paper_id, title) comprised the corpus for the next phase of our pipeline, parallel triple extraction.

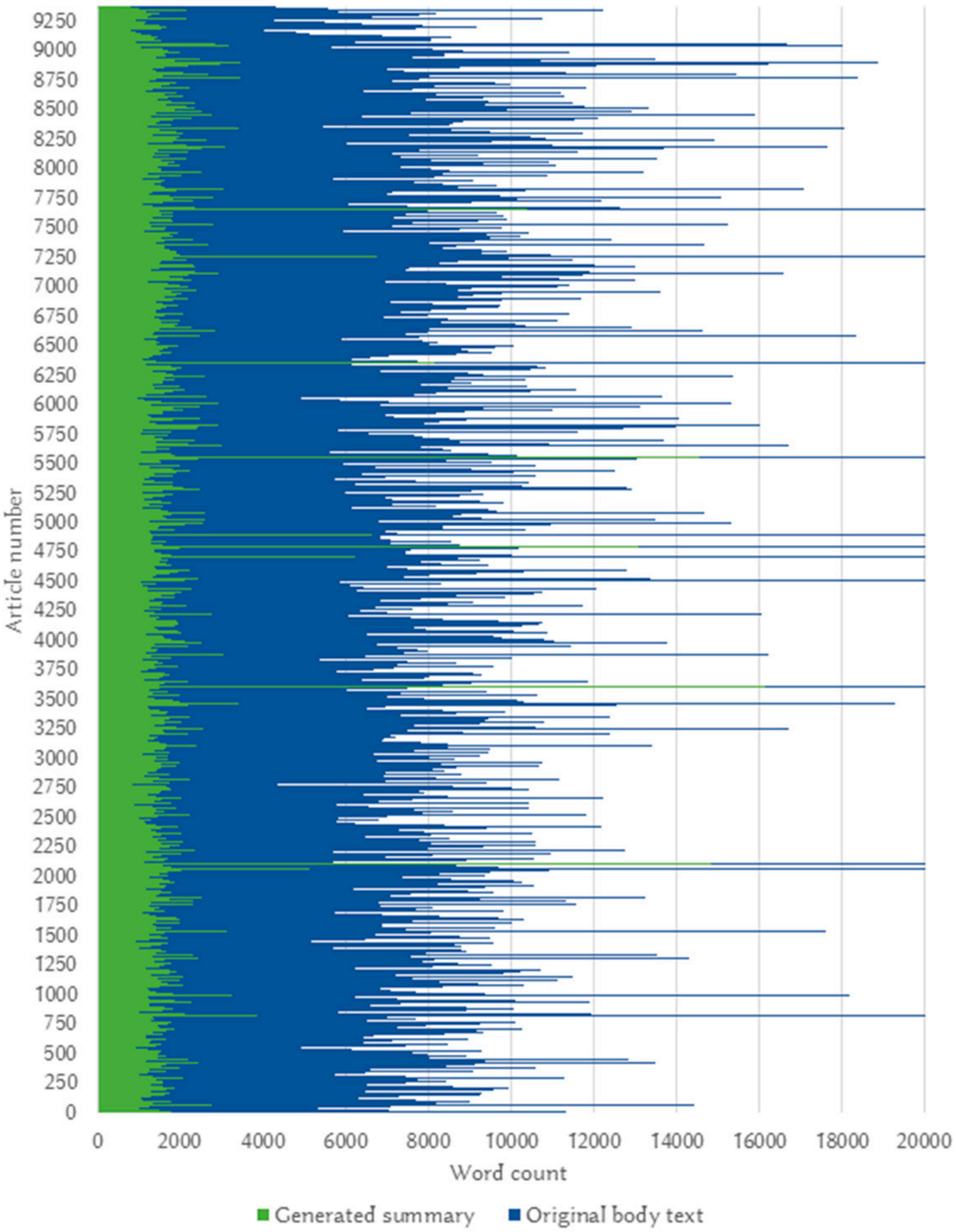

Figure 2. Word reduction per paper using extractive summarization. The original body text of each article (blue) is replaced by its summary (green). This preprocessing step aims at selecting the most representative sentences of the given documents, despite their original length.

3.2.3. Parallel Triple Extraction

Our approach aims at distilling knowledge from texts that can be used directly for end-user applications such as structured searches (e.g., "*Find all triples containing SARS-CoV-2 as a Subject along with all related papers*"). Therefore, triple extraction is the third and core step of our knowledge extraction pipeline, in an effort to discover relations between the entities of the CORD-19 corpus. Instead of restricting ourselves to a subset of relations in a pre-defined ontology (closed information extraction), we leveraged an open information extraction (OIE) approach, where triple formulation is defined as the task of generating a structured, machine-readable representation of the information in text, with the length of each triple element varying from a single word to a short text phrase. Moreover, instead of limiting ourselves to a single engine, we combined the three most popular OIE systems, namely the Open IE system from the University of Washington (UW) and Indian Institute of Technology, Delhi (IIT Delhi, India) [71], ClausIE from Max Planck Institute (MPI) [72], and the neural OIE model from AllenNLP [73]. A brief explanation of the intuition behind each system is provided below:

1. Open IE 5.1 from UW and IIT Delhi is a successor to the Ollie learning-based information extraction system [74]. The latest version is based on the combination of four different rule-based and learning-based OIE tools; namely CALMIE (specializing in triple extraction from conjunctive sentences) [75], RelNoun (for noun relations) [76], BONIE (for numerical sentences) [77], and SRLIE (based on semantic role labeling) [78].
2. ClausIE from MPI follows a clause-based approach, first identifying the clause type of each sentence and then applying specific proposition extraction based on the corresponding grammatical function of the clause's constituents. It also considers nested clauses as independent sentences. Because ClausIE detects useful pieces of information expressed in a sentence before representing them in terms of one or more extractions, it is especially useful in splitting complex sentences into many individual triples [41].
3. AllenNLP OIE system formulates the triple extraction problem as a sequence BIO tagging problem and applies a bi-LSTM transducer to produce OIE tuples, which are grouped by each sentence's predicate [72]. Given that it relies on supervised learning and contextualized word embeddings to produce independent probability distributions over possible BIO tags for each word, it has the potential of discovering richer and more complex relations. On the downside, it is not guaranteed that the neural sequence tagger will produce exactly two arguments for a given predicate (i.e., a subject and an object), thus complicating the triple extraction process.

We implemented a parallel triple extraction sequence based on the aforementioned OIE systems, where the extractive summary of every CORD-19 article from the previous step was passed through each one of the three engines. Given that Open IE 5.1 is based on handcrafted extraction heuristics and automatically constructed triple extractors, that ClausIE follows a rule-based approach which exploits linguistic knowledge about the grammar (clause types) of the English language and that AllenNLP OIE system is dependent on the context's vector representation to detect parts of speech, we relied on the complementarity between the different approaches to ensure maximum recall. The only shortcoming of this method is that it inevitably leads to a higher duplication rate compared to using a single engine; however, this issue was effectively tackled in the next and final stage of our pipeline. Example cases with extracted triples showcasing the complementarity of the different OIE engines are shown in Table 2.

Table 2. Parallel triple extraction using different OIE engines. The left column shows the processed sentence and the Source ID of the corresponding CORD-19 article. The middle column shows the derived triples, while the right column denotes the engine(s) that discovered each triple (O: Open IE, C: ClausIE, A: AllenNLP OIE).

Sentence	Extracted Triples (S/P/O)	Engine
"RA and PBD blocked the attachment of IAV and 3C-like protease (3CLP) of severe acute respiratory syndrome-associated coronaviruses plays a pivotal role in viral replication and is a promising drug target", Source: c85ca5217f9051f839115 69eed1eb52cf992f7dd	RA and PBD/blocked/the attachment of IAV	O,C
	3C-like protease of severe acute respiratory syndrome-associated coronaviruses/is/3CLP	O
	3C-like protease of severe acute respiratory syndrome-associated coronaviruses/plays/a pivotal role in viral replication	O
	3C-like protease of severe acute respiratory syndrome-associated coronaviruses/is/a promising drug target	O
	PBD/blocked/the attachment of 3C-like protease (3CLP) of severe acute respiratory syndrome-associated coronaviruses	C
	PBD blocked the attachment of IAV/is/a promising drug target	C
	PBD/blocked/the attachment of IAV	C
	RA blocked the attachment of 3C-like protease (3CLP) of severe acute respiratory syndrome-associated coronaviruses/is/a promising drug target	C
	RA/blocked/the attachment of 3C-like protease (3CLP) of severe acute respiratory syndrome-associated coronaviruses	C
	the attachment of IAV/plays/a pivotal role in viral replication	C
	RA and PBD/blocked/the attachment of IAV and 3C-like protease (3CLP) of severe acute respiratory syndrome-associated coronaviruses	A
"CRP is an acute phase protein that has been linked to the presence and severity of bacterial infection in numerous studies during the past 2 decades [9,34,35]", Source: b8e9c45dda9cb8c9c4321 a55704ab2a66fb34f7d	CRP/is/an acute phase protein	O
	an acute phase protein/has been linked/to the presence and severity of bacterial infection in numerous studies during the past 2 decades	O,C
	an acute phase protein/has been linked/to the presence and severity of bacterial infection in numerous studies	O
	CRP/is/an acute phase protein that has been linked to the severity of bacterial infection in numerous studies during the past 2 decades	C,A

Table 2. *Cont.*

Sentence	Extracted Triples (S/P/O)	Engine
"Involvement of polyamines, possibly due to loss of epigenetic control of X-linked polyamine genes/is suspected in SjS since the appearance of acrolein conjugated proteins is related to the intensity of SjS and acrolein is an oxidation product of polyamines (88)", Source: 8495f7c65f4a6cbce0e0 d53c0900f10f6740826e	Involvement of polyamines possibly due to loss of epigenetic control of X-linked polyamine genes/is suspected/in SjS since the appearance of acrolein conjugated proteins is related to the intensity of SjS	O
	Involvement of polyamines possibly due to loss of epigenetic control of X-linked polyamine genes/is suspected/in SjS	O
	the appearance of acrolein conjugated proteins/is related/to the intensity of SjS	O
	acrolein/is/an oxidation product of polyamines	O
	acrolein/is/an oxidation product	O
	Involvement of polyamines/possibly due to loss of epigenetic control of X-linked polyamine genes/is related/to the intensity of SjS	C
	Involvement of polyamines, possibly due to loss of epigenetic control of X-linked polyamine/is/suspected	A
"The new genome sequence was obtained by first mapping reads to a reference SARS-CoV-2 genome using BWA-MEM 0.7.5a-r405 with default parameters to generate the consensus sequence.", Source: b5d303cbcfe6be92d733e c593118b388db77452e	The new genome sequence/obtained/by first mapping reads to a reference SARS-CoV-2 genome	O,C,A
	a reference SARS-CoV-2 genome/be using/BWA-MEM 0.7.5a-r405 with default parameters to generate the consensus sequence/	O
	BWA-MEM 0.7.5a-r405/to generate/the consensus sequence	C
	a reference SARS-CoV-2 genome/using/BWA-MEM 0.7.5a-r405 to generate the consensus sequence	C
	The new genome sequence/was/obtained	A

3.2.4. Entity Enrichment & Graph Representation

Our methodology concludes with a series of post-processing activities, including linking the extracted entities to an existing ontology, performing polarity detection on the phrases related to each triple as well as cleaning the duplicate triples that were extracted via the parallel execution of the aforementioned OIE engines, before representing them through a graph data modeling process. More details about the technical implementation of each subtask are given below:

1. Entity linking: We leveraged the EntityLinker component from SciSpacy [79], a Python package containing models for processing biomedical, scientific, or clinical text. The component was used to perform a string overlap-based search (char-3grams) on named entities, comparing them with the concepts of the UMLS (Unified Medical Language System) knowledge base, using an approximate nearest neighbors search. The UMLS knowledge base contains over four million concepts along with additional information (e.g., definitions, hierarchies, concept–concept relations) from many health and biomedical vocabularies and standards (including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT), enabling interoperability between computer systems [80]. For each triple subject or object with one or more entities linked to UMLS concepts, the coded concept name, concept description, and confidence score of the linking process was added along with the existing triple information. In order to address the ambiguity of biomedical terminology, we exploited the parametrization capabilities of the SciSpacy EntityLinker component, mainly by using the *resolve_abbreviations* parameter to resolve any abbreviations identified in the corpus before performing the linking and by tuning the threshold that a mention candidate must reach to be linked to a specific UMLS concept. Of course, this barely scratches the surface of biomedical terms disambiguation which remains a challenging task [81]. Overall, this entity enrichment process not only increases the contextual value of the extracted triples, but also facilitates the research for specific ontologies by mapping the existing entities with their normalized lexical variants (aliases).
2. Polarity detection: By implementing polarity detection at sentence-level, we were able to classify triples that were inherently bearing a positive or a negative value. We relied on AllenNLP's RoBERTa-based binary sentiment analyzer [63], which achieves 95.11% accuracy on the Stanford Sentiment Treebank test set.
3. Triple cleaning: This process is aimed at reducing the redundant triples that resulted from the parallel triple extraction process, as described in Section 3.2.3, while also reducing the number of non-informative triples with little contextual value. To this end, we considered only "fully-linked" triples, i.e., triples whose both subject and object was linked to at least one UMLS concept. For those remaining triples, we additionally implemented a deduplication process to keep only the unique ones, based on the mentions of concepts of each sentence's subjects and objects. In this manner, only one triple containing the same UMLS concepts was stored for each sentence.
4. Continuity representation: In order to enhance the readability of the extracted information, we implemented a script that connected the extracted triples with each other, based on the order of appearance in the original text. This way, the user is able to unravel the scope of the targeted content by interacting with its structured representation.
5. Graph representation: It is the final step of our proposed pipeline, aiming at the practical interaction with the ingested data (e.g., visualizations and queries) that will facilitate data exploration tasks. Due to the rich internal structure that characterizes labeled property graphs, allowing each node or relationship to store more than one property (thus reducing the graph's size), we used the Neo4J graph database based on labeled property graphs for storing, representing, and enriching the extracted relationships [82]. Neo4j is a Java-based graph database management system, described by its developers as an ACID-compliant transactional database with native graph storage and processing. It provides a web interface allowing relationships to be queried

via CYPHER, a declarative graph query language that allows for expressive and efficient data querying in property graphs.

The final structure of the enriched triple extraction output following the above processes is shown in Table 3 for a single triple. The graph representation consists of three distinct types of nodes which are linked via three types of edges:

- **Corpus** nodes (purple): they signify the ingested text (e.g., scientific articles), are represented by the *article_title* and the unique *article_id* properties, and are connected to one or more triples via the *contains triples* one-directional edge,
- **Subject** nodes (green): they denote the subject of the extracted triple and are connected to one or more corpora (as mentioned above), as well as to one or more **Objects**, via *predicate* one-directional edges. The **Subject** node has the following properties: *value*: the natural language text of the subject; *subj_entity_name:* the name(s) of the entities linked to UMLS concepts; *subj_entity_coded*: the coded ID(s) of the entities linked to UMLS concepts; *subj_entity_description:* a short description of the linked entities; *subj_entity_confidence:* the reported confidence of the entity linking process; *article_id*: the unique ID of the extracted article; *article_title*: the title of the extracted article; *sentence_text*: the text of the processed sentence; *sent_num*: the serial number of the sentence in the whole corpus; *triple_num*: the serial number of the triple extracted from the given sentence; and *engine*: the OIE engine that from which the triple was extracted from.
- **Object** nodes (red): they denote the object of the extracted triple and are connected to their respective **Subject** through the *predicate* edge (as mentioned above), as well as with the "Subject" of the following triple in the corpus (if there exists one) via the *followed by* one-directional edge. Their properties are identical to those of the **Subject** nodes.

An example figure of the graph representation that concentrates the aforementioned nodes and edges of one CORD-19 article is in Figure 3.

Table 3. Sample output of the information extraction pipeline. The triple (subject–predicate–object) extracted by the sentence denoted in the *sentence_text* field is linked with its corresponding UMLS entities (*sub_entity_coded*, *obj_entity_coded* fields).

Field	Value
title	NLRP3 Inflammasome-A Key Player in Antiviral Responses
id	372caa549b07492de5fd7064e9cadc62bef0f478
subject	chronic intrahepatic inflammation and liver injury
predicate	be mediated
object	by the NLRP3 inflammasome
subj_entity_name	[‘chronic’, ‘liver’]
subj_entity_coded	[‘C0333383’, ‘C0160390’]
subj_entity_confidence	[0.7110092043876648, 1.0]
subj_entity_description	[‘CUI: C0333383, Name: Acute and chronic inflammation Definition: The coexistence of a chronic inflammatory process with a superimposed polymorphonuclear neutrophilic infiltration. TUI(s): T046\nAliases: (total: 8): Active chronic inflammation, Chronic Active Inflammation, Chronic active inflammation, Active Chronic Inflammation, Subacute inflammatory cell infiltrate, Subacute inflammatory cell infiltration, Subacute inflammation, Acute and chronic inflammation (morphologic abnormality)’, ‘CUI: C0160390, Name: Injury of liver Definition: Damage to liver structure or function due to trauma or toxicity. TUI(s): T037 Aliases (abbreviated, total: 12): of liver injury, Injury to liver, injuries liver, injury liver, liver injury, Injury to Liver, Hepatic injury, injury hepatic, hepatic injury, Hepatic trauma’]
obj_entity_name	[‘NLRP3’]
obj_entity_coded	[‘C1424250’]
obj_entity_confidence	[1.0]
obj_entity_description	[‘CUI: C1424250, Name: NLRP3 gene Definition: This gene plays a role in both inflammation and apoptosis. TUI(s): T028 Aliases (abbreviated, total: 29): NLRP3 gene, NLRP3 Gene, AVP, AII, MWS, CRYOPYRIN, Cryopyrin, FCAS, CIAS1 GENE, CIAS1 gene’]
sentiment	positive
sent_num	265719
sentence_text	HCV infection promotes chronic intrahepatic inflammation and liver injury mediated by the NLRP3 inflammasome (21).
triple_num	1
engine	1

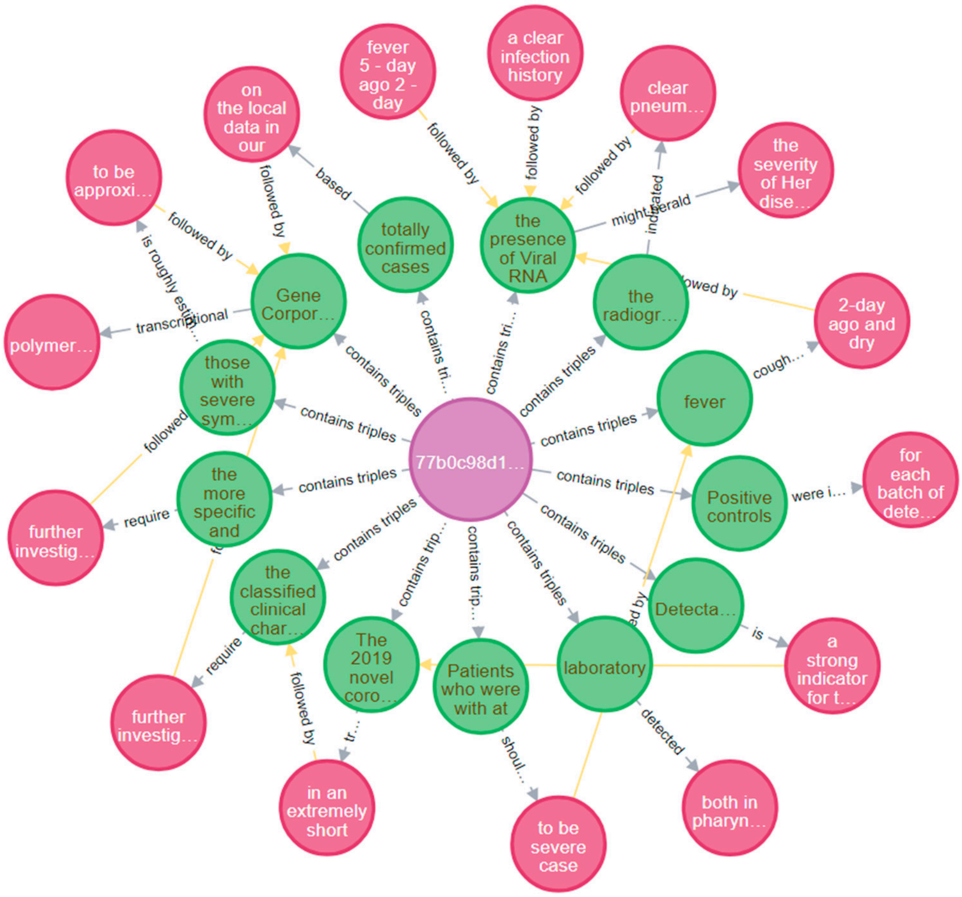

Figure 3. Graph representation example of the triples (*Subjects*—green and *Objects*—red) extracted from one CORD-19 article (purple) with ID 77b0c98d1a2ca46b219ad090074814c387c80d8f using Neo4J (some triples are omitted for better visibility).

4. Results

After submitting the CORD-19 articles to our information extraction pipeline, we acquired 411,189 triples that contain subjects and objects linked to at least one UMLS entity. A compressed file containing the extracted triples is available online (https://github.com/lighteternal/CORD-19-OIE-triple-extraction). In the following subsections, we evaluate the validity of our information extraction process and present a number of indicative data exploration tasks on the CORD-19 dataset, which are enabled by the structured representation of the extracted information.

4.1. Evaluation of the Information Extraction Process

The evaluation of OIE systems on domain-specific corpora is generally a tricky process, mainly due to the lack of gold extractions (i.e., valid, manually annotated triples) for the specific domain. The common approach to tackle this lack of gold standards and metrics is to annotate a small subset of the extracted triples for correctness, thus yielding a precision measure as the ratio of valid extracted triples over the total number of extracted triples [83]. This approach, however, does not measure the extent to which actual valid triples are being overlooked (sensitivity), since it requires the total population of potentially valid triples. In order to acquire indicative evaluation metrics for our methodology, we manually generated all possible triples from a subset of 50 sentences (one sentence usually generates more than one triple) and calculated both the precision (as defined above) and the recall as the proportion of valid triples extracted by our pipeline to the total number of valid triples (automatically extracted and hand-crafted). We also calculated the F1-score as the harmonic mean of precision and recall. It should be noted that during the manual triple generation process, we only considered triples that could be potentially linked to UMLS entities (e.g., triples from sentences such as "*We discovered a mutation of the virus*" would not count as valid, because the subject "*we*" does not correspond to any UMLS entity). Evaluation results are provided in Table 4.

Table 4. Evaluation metrics of the information extraction pipeline, on a subset of 50 randomly selected sentences of the CORD-19 corpus. Precision is the ratio of valid triples from our approach among the total generated triples, while recall is the ratio of valid triples from our approach among the total valid ones (generated and hand-crafted). F1-score is the harmonic mean of precision and recall.

Metric	Value
Precision	0.78
Recall	0.76
F1-score	0.77

Although a direct comparison with other approaches (e.g., end-to-end methods, OIE engines) is infeasible as it would require their implementation on the same dataset, it is easily understandable that the complementarity of our methodology leads to better performance, at least compared to standalone OIE engines. This can also be inferred by the experimental results of several OIE systems on different datasets, with precision, recall, and F-measure barely surpassing the 0.7 threshold for a large number of extractions [47,84].

4.2. Data Exploration Tasks

We present a number of indicative data exploration tasks on the CORD-19 dataset to demonstrate the capabilities of our information extraction pipeline.

1. The first task focuses on visualizing triples from the CORD-19 bibliography whose subject refers to the SARS-CoV-2 virus, the strain of coronavirus that causes the COVID-19 disease. The domain expert can either query the database for subjects containing the name of the virus or he/she can use the UMLS coded ID of the SARS-CoV-2 (C5203676), to acquire results containing all the corresponding aliases of the entity. These results are available in both tabular and graphical form (Table 5, Figure 4):
2. The second task attempts to discover useful relationships regarding IL-6, a pleiotropic proinflammatory cytokine that is found in increased levels in COVID-19 patients and similar viruses. By performing a targeted search on subjects containing the "virus" entity and objects containing the "IL-6" entity, we get the results shown in Table 6 (in natural language form) and Figure 5 (as UMLS coded entities). This time, we are also interested in the title of the scientific article from where the triple was extracted.
3. The final data exploration task allows us to focus on one of the articles and exploit the continuity representation functionality (*followed by* edges) to traverse through the generated triples. The result is akin to a graphical summary of the processed article. The generated chain consists of alternating subject/object nodes, depicting the sequence of their appearance in the original text (Figure 6).

Table 5. Sample of extracted triples related to the SARS-CoV-2 entity (some properties of the related nodes have been omitted for better visibility).

Subject	Predicate	Object	UMLS in Subj.	UMLS in Obj.
the SARS-CoV-2	might be imported	to the seafood market in a short time	('C5203676')	('C0206208', 'C4526594')
the mortality rate due to 2019-nCoV is comparatively lesser than the earlier outbreaks of SARS and MERS-CoVs, as well as this virus	presents	relatively mild manifestations	('C0026565', 'C5203676', 'C0012652', 'C1175743')	('C1513302')
the initial identification of 2019-NCoV from 7 patients	diagnosed	with unidentified viral pneumonia	('C0020792', 'C5203676', 'C0030705')	('C0032310')
2019-nCoV cases	be detected	outside China	('C5203676', 'C0868928')	('C0008115')
the receptor binding domain of SARS-CoV-2	was	capable of binding ACE2 in the context of the SARS-CoV spike protein	('C0597358', 'C5203676')	('C1167622', 'C1422064', 'C1175743')
In-depth understanding the underlying pathogenic mechanisms of SARS-CoV-2	will reveal	more targets	('C0205125', 'C0450254', 'C0441712', 'C5203676')	('C0085104')
SARS-CoV-2 orf8 and orf10 proteins	are	other methods to transmit SARS-CoV-2	('C5203676', 'C1710521')	('C0449851', 'C0332289', 'C5203676')
As of February 20 2020 the 2019 novel coronavirus now named SARS-CoV-2 causing the disease COVID-19 has caused over 75,000	has spread	to 25 other countries World Health Organization	('C0205314', 'C0206419', 'C5203676', 'C0012634', 'C5203670')	('C0043237')
infections due to SARS-CoV-2	have spread	to over 26 countries	('C3714514', 'C5203676')	('C0454664',)
2019-nCoV S-RBD or modified S-RBD of other coronavirus	may be applied	for developing 2019-nCoV vaccines	('C5203676')	('C0042210')
DetecTable 2019-nCoV viral RNA in blood	is	a strong indicator for the further clinical severity	('C3830527', 'C5203676', 'C0035736', 'C0005767')	('C0021212', 'C2981439')
SARS-CoV-2	is causing	the ongoing COVID-19 outbreak	('C5203676')	('C5203670')
SARS-CoV-2	may use	integrins	('C5203676')	('C0021701',)
SARS-CoV-2	appears	to have been transmitted during the incubation period of patients	('C5203676')	('C1521797', 'C2752975', 'C1561960', 'C0030705')
SARS-CoV-2	becomes	highly transmissible	('C5203676')	('C0162534')
SARS-CoV-2	is	an emerging new coronavirus	('C5203676')	('C0206419')
SARS-CoV-2	might exhibit	seasonality	('C5203676')	('C0683922')
SARS-CoV-2	was not detected	suggesting that more evidences are needed before concluding the conjunctival route as the transmission pathway of SARS-CoV-2	('C5203676')	('C3887511', 'C2959706', 'C0040722', 'C5203676')
SARS-CoV-2	is circulating	in China	('C5203676')	('C0008115')
SARS-CoV-2	be induced	pneumonia resulting in severe respiratory distress and death	('C5203676')	('C0032285', 'C1547227', 'C0476273', 'C0011065')
a novel corona virus first 2019-nCov then SARS-CoV-2	was identified	as the cause of a cluster of pneumonia cases	('C0205314', 'C0206750', 'C5203676')	('C0015127', 'C1704332', 'C0032285')

Table 5. *Cont.*

Subject	Predicate	Object	UMLS in Subj.	UMLS in Obj.
The accurate transmission rate of SARS-CoV-2	is	unknown since various factors impact its transmission	(‘C0443131′, ‘C0040722′, ‘C5203676′)	(‘C1257900′, ‘C0726639′, ‘C0040722′)
The observed morphology of SARS-CoV-2	is	consistent with the typical characteristics of the Coronaviridae family	(‘C1441672′, ‘C0332437′, ‘C5203676′)	(‘C0332290′, ‘C1521970′, ‘C0010076′)
the receptor-binding domain of SARS-CoV-2	demonstrates	a similar structure to that of SARS-CoV	(‘C1422496′, ‘C5203676′)	(‘C0678594′, ‘C1175743′)
the sequence divergence between SARS-CoV-2 and RaTG13	is	great to assign parental relationship	(‘C0004793′, ‘C0443204′, ‘C5203676′, ‘C4284300′)	(‘C0030551′, ‘C1706279′)
the mutation rate of SARS-CoV-2	is	suggestive to humans	(‘C1705285′, ‘C1521828′, ‘C5203676′)	(‘C0086418′)
the question of whether it is a potential transmission route of SARS-CoV-2	deserves	further evaluation	(‘C3245505′, ‘C0040722′, ‘C5203676′)	(‘C1261322′)
the patterns and modes of the interaction between SARS-CoV-2 and host antiviral defense	would be	similar share many features	(‘C0449774′, ‘C4054480′, ‘C0037420′, ‘C5203676′, ‘C1167395′, ‘C1880266′)	(‘C1521970′)
the immune response against SARS-CoV-2	is decoupled	from viral replication	(‘C0301872′, ‘C5203676′)	(‘C0042774′)
the detection of SARS-CoV-2 with a high viral load in the sputum of convalescent patients	arouse	concern about prolonged shedding of the virus after recovery	(‘C1511790′, ‘C5203676′, ‘C0376705′, ‘C0038056′, ‘C0740326′, ‘C0030705′)	(‘C0439590′, ‘C0162633′, ‘C0042776′, ‘C0237820′)
a new coronavirus (2019-nCoV), identified through genomic sequencing	was	the culprit of the pneumonia	(‘C0206419′, ‘C5203676′, ‘C1294197′)	(‘C2068043′, ‘C0032285′)
SARS-CoV-2 examination of the viral genome	was	critical for identifying the pathogen	(‘C5203676′, ‘C0582103′, ‘C0042720′)	(‘C2826883′, ‘C0450254′)

Table 6. Sample of extracted triples showing relations between the entities “IL-6” and “virus” (some properties of the related nodes have been omitted for better visibility).

Article	Subject	Predicate	Object	UMLS in Subj.	UMLS in Obj.
Distinct Regulation of Host Responses by ERK and JNK MAP Kinases in Swine Macrophages Infected with Pandemic (H1N1) 2009 Influenza Virus	Highly pathogenic H5N1 viral infection in human macrophages	induced	higher expression of IL-6	(‘C0450254′, ‘C0016627′, ‘C0042769′, ‘C0086418′)	(‘C0205250′, ‘C2911684′, ‘C0021760′)
The Expression of IL-6, TNF-, and MCP-1 in Respiratory Viral Infection in Acute Exacerbations of Chronic Obstructive Pulmonary Disease	patients with viral infections	showed	higher levels of TNF IL-6 and MCP-1	(‘C0030705′, ‘C0042769′)	(‘C0205250′, ‘C0441889′, ‘C1522668′, ‘C0021760′, ‘C0128897′)

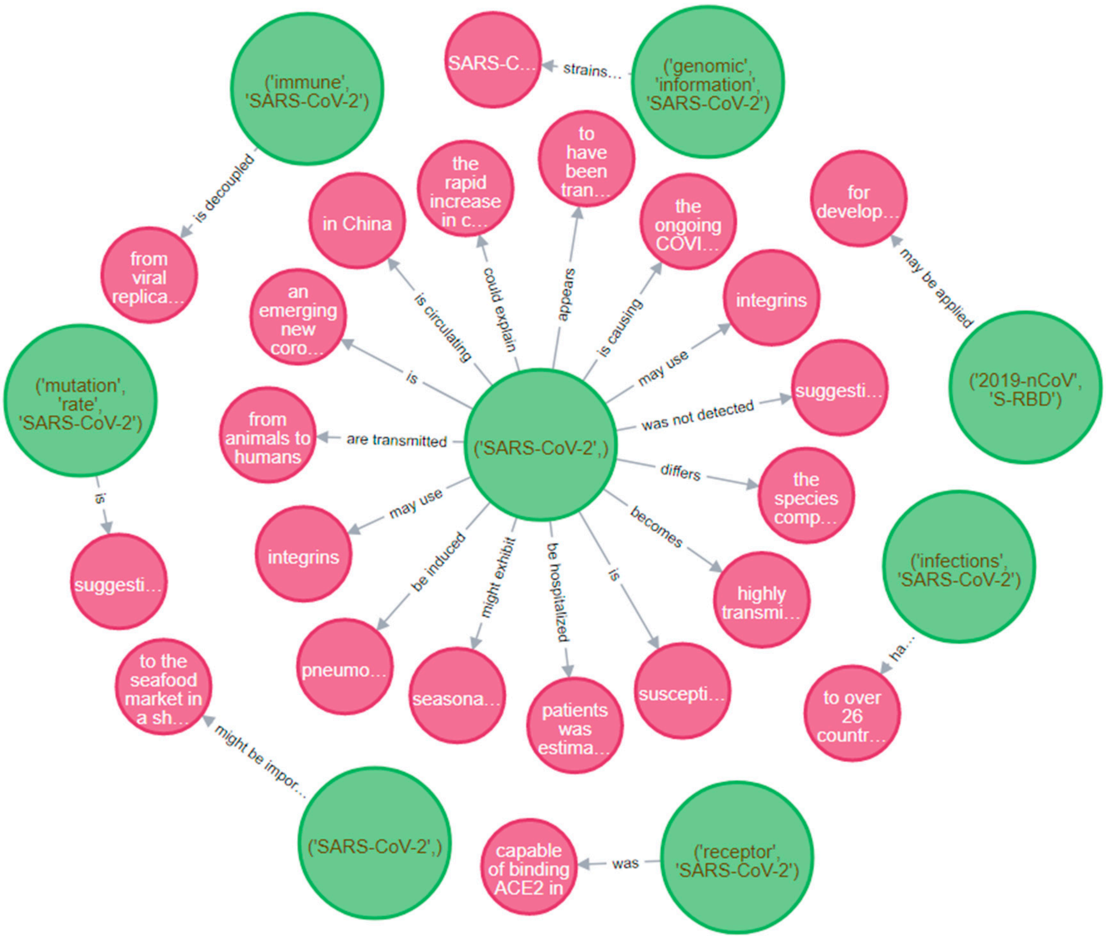

Figure 4. Sample of the Neo4J graph visualization of the triples related to the SARS-CoV-2 entity (some triples and properties have been omitted for better visibility).

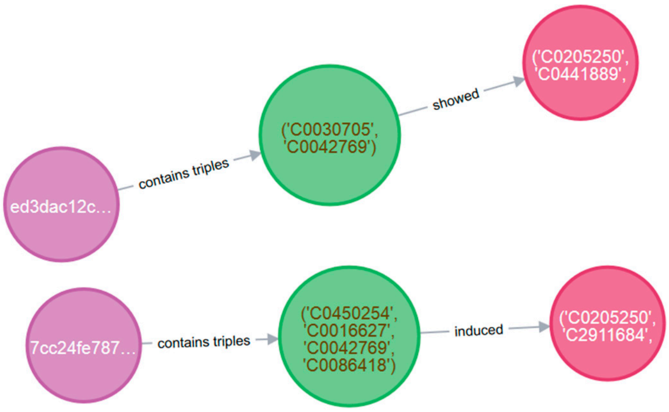

Figure 5. Sample of the Neo4J graph visualization of the triples showing relations between the coded UMLS IDs of “virus” (green) and “IL-6” (red) found in two CORD-19 articles (purple). Other UMLS entities (e.g., “H5N1”, “MCP-1”) are also present in the *Subject* and *Object* nodes.

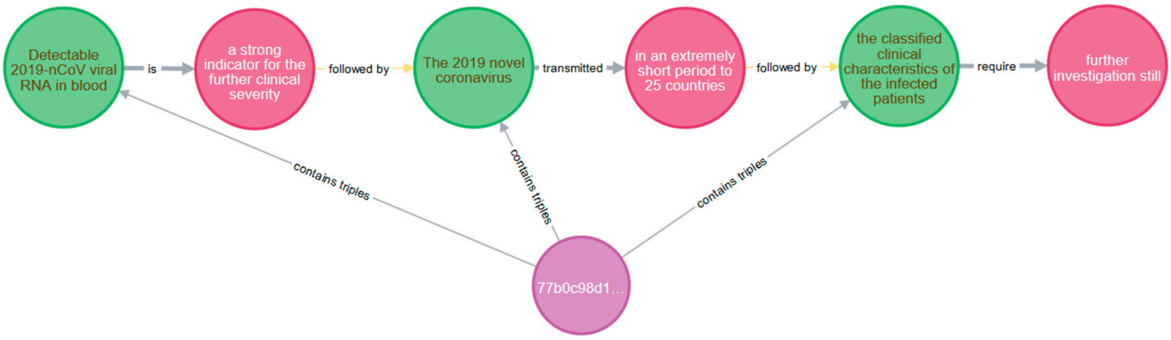
Detectable 2019-nCoV viral RNA in blood
is
a strong indicator for the further clinical severity
followed by
The 2019 novel coronavirus
transmitted
in an extremely short period to 25 countries
followed by
the classified clinical characteristics of the infected patients
require
further investigation still
contains triples
contains triples
contains triples
77b0c98d1…

Figure 6. A sample sequence of extracted triples from an *article's* (purple) sentences as alternating *subject* (green) and *object* (red) nodes, linked with *predicate* (grey) and *followed_by* (yellow) edges.

5. Discussion

This section aims at illustrating the impact of our approach to domain-specific information extraction tasks, through an end-to-end example (Table 7) which showcases the added value of each individual component that comprises our pipeline compared to a standalone OIE process. For reasons of better readability, we have included only a small extract of the processed article body.

As seen in the upper part of Table 7, the coreference resolution component correctly substitutes an entity mention ("the bats") to its more semantically informative antecedent ("horseshoe bats"). Subsequently, the extractive summarization component identifies a set of salient points and groups them to a much shorter, concise corpus, which is subjected to the parallel triple extraction process. The extracted triples are finally passed through an entity enrichment and cleaning process, that results in storing only those linked to one or more UMLS concepts. Other provided information including polarity, UMLS coded IDs, matching confidence scores, concept descriptions as well as the Neo4J graph visualization of the extracted triples are omitted in this example.

For the sake of comparison, a small subset of the triples derived from the application of a standalone, state-of-the-art OIE engine (ClausIE) on the raw article body is provided at the lower part of Table 7. It is apparent from the triple samples that the lack of pre-processing (coreference resolution, summarization) and post-processing (entity resolution, cleaning) methods, usually leads to a large number of uninformative triples, despite the fact of some being syntactically correct. For example, the triple {"This finding", "implies", "a possible recombination event"} although being valid, it unpacks little contextual value with regard to the aims of the article. On the contrary, our approach seems to increase the expressiveness and informativeness of the derived triples, ensuring that they remain relevant to the context of the article.

It should be noted that the inclusion of the aforementioned pre-processing and post-processing operations increases the computational cost of our pipeline compared to adopting a standalone OIE approach; however, this is partially countered by the fact that the core triple extraction process is applied to a significantly shorter and more concise corpus. Furthermore, it is evident that, although the results seem encouraging, none of the pipeline components are guaranteed to produce optimal results in every complex situation. More specifically, the extractive summarization component may fail to capture all the key points of the article resulting in loss of information, the in-place coreference resolution component may fail to find the correct antecedent of a mention, the parallel triple extraction process may miss some triples involving compound syntactic phenomena, and the entity enrichment tool may perform a wrong linking to a knowledge base concept. This is to be expected in real-world machine learning applications, where extractive summarization can be conceived as a feature selection method, in-place coreference resolution can be considered as a feature transformation technique, and entity linking can be seen as a data filtering method, each of them contributing to the usefulness of the overall approach. As discussed in Section 4, the added value of our methodology stems from the effective combination of different NLP tasks, benefiting from their distinct characteristics in an attempt to provide a robust outcome.

Table 7. End-to-end example of information extraction on a CORD-19 article body. The highlighted text composes the extractive summary of the article. The coreference resolution component replaces the anaphors (orange) by their antecedent (green). The parallel information extraction component provides triples, which are linked with existing UMLS concepts via the entity enrichment component. The triples extracted by our pipeline are compared to those extracted using a standalone OIE engine directly on the article body.

CORD-19 Article ID: 85783a36e7e787302307f42460839435d665f4e7
Article Title: SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat
Article body: [. . .] Subsequently, coronaviruses with high similarity to the human SARS-CoV or civet SARS-CoV-like virus were isolated from horseshoe bats$_{\{1\}}$, concluding ~~thebats~~$_{\{1\}}$horseshoe bats$_{\{1\}}$ as the potential natural reservoir of SARS-CoV whereas masked palm civets are the intermediate host [53–56]. It is thus reasonable to suspect that bat is the natural host of SARS-CoV-2 considering its similarity with SARS-CoV. The phylogenetic analysis of SARS-CoV-2 against a collection of coronavirus sequences from various sources found that SARS-CoV-2 belonged to the Betacoronavirus genera and was closer to SARS-like coronavirus in bat [19]. By analyzing genome sequence of SARS-CoV-2, it was found that SARS-CoV-2 felled within the subgenus Sarbecovirus of the genus Betacoronavirus and was closely related to two bat-derived SARS-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21, but were relatively distant from SARS-CoV [15,18,57–59]. Meanwhile, Zhou and colleagues showed that SARS-CoV-2 had 96.2% overall genome sequence identity throughout the genome to BatCoV RaTG13, a bat coronavirus detected in Rhinolophus affinis from Yunnan province [14]. Furthermore, the phylogenetic analysis of full-length genome, the receptor binding protein spike (S) gene, and RNA-dependent RNA polymerase (RdRp) gene respectively all demonstrated that RaTG13 was the closest relative of the SARS-CoV-2 [14]. However, despite SARS-CoV-2 showed high similarity to coronavirus from bat, SARS-CoV-2 changed topological position within the subgenus Sarbecovirus when different gene was used for phylogenetic analysis: SARS-CoV-2 was closer to bat-SL-CoVZC45 in the S gene phylogeny but felled in a basal position within the subgenus Sarbecovirus in the ORF1b tree [57]. This finding implies a possible recombination event in this group of viruses. Of note, the receptor-binding domain of SARS-CoV-2 demonstrates a similar structure to that of SARS-CoV by homology modelling but a few variations in the key residues exist at amino acid level [15,19]. Despite current evidences are pointing to the evolutional origin of SARS-CoV-2 from bat virus [15,57], an intermediate host between bats and human might exist . Lu et al. raised four reasons for such speculation [15]: First, most bat species in Wuhan are hibernating in late December; Second, no bats in Huanan Seafood market were sold or found; Third, the sequence identity between SARS-CoV-2 and bat-SL-CoVZC45 or bat-SL-CoVZXC21, the closest relatives in their analyses, is lower than 90%; Fourth, there is an intermediate host for other humaninfecting coronaviruses that origin from bat. For example, masked palm civet and dromedary camels are the intermediate hosts for SARS-CoV [49] and MERS-CoV respectively [60]. A study of the relative synonymous codon usage (RSCU) found that SARS-CoV-2, bat-SL-CoVZC45, and snakes had similar synonymous codon usage bias, and speculated that snake might be the intermediate host [61]. However, no SARS-CoV-2 has been isolated from snake yet. Pangolin was later found to be a potential intermediate host for SARS-CoV-2 . The analysis of samples from Malytan pangolins obtained during anti-smuggling operations from Guangdong and Guangxi Customs of China respectively found novel coronaviruses representing two sub-lineages related to SARS-CoV-2 [62]. The similarity of SARS-CoV-2 to these identified coronaviruses from pangolins is approximately 85.5% to 92.4% in genomes, lower than that to the bat coronavirus RaTG13 (96.2%) [14,62]. However, the receptor-binding domain of S protein from one sub-lineage of the pangolin coronaviruses shows 97.4% similarity in amino acid sequences to that of SARS-CoV-2, even higher than that to RaTG13 (89.2%) [62]. Interestingly, the pangolin coronavirus and SARS-CoV-2 share identical amino acids at the five critical residues of RBD of S protein, while RaTG13 only possesses one [62]. The discovery of coronavirus close to SARS-CoV-2 from pangolin suggests that pangolin is a potential intermediate host. However, the roles of bat and pangolin as respective natural reservoir and intermediate host still need further investigation.As an emerging virus, there is no effective drug or vaccine approved for the treatment of SARS-CoV-2 infection yet. Currently, supportive care is provided to the patients, including oxygen therapy, antibiotic treatment, and antifungal treatment, extra-corporeal membrane oxygenation (ECMO) etc. [21,22]. To search for an antiviral drug effective in treating SARS-CoV-2 infection, Wang and colleagues evaluated seven drugs, namely, ribavirin, penciclovir, nitazoxanide, nafamostat, chloroquine, remdesivir (GS-5734) and favipiravir (T-750) against the infection of SARS-CoV-2 on Vero E6 cells in vitro [63]. Among these seven drugs, chloroquine and remdesivir demonstrated the most powerful antiviral activities with low cytotoxicity. The effective concentration (EC 50) for chloroquine and remdesivir were 0.77µM and 1.13µM respectively. Chloroquine functions at both viral entry and post-entry stages of the SARS-CoV-2 infection in Vero E6 cells whereas remdesivir does at post-entry stage only. Chloroquine is a drug used for an autoimmune disease and malarial infection with potential broad-spectrum antiviral activities [64,65]. An EC90 (6.90 µM) against the SARS-CoV-2 in Vero E6 cells is clinically achievable in vivo according to a previous clinical trial [66]. Remdesivir is a drug currently under the development for Ebola virus infection and is effective to a broad range of viruses including SARS-CoV and MERS-CoV [67,68]. Functioning as an adenosine analogue targeting RdRp, Remdesivir can result in premature termination during the virus transcription [69,70]. The EC90 of remdesivir against SARS-CoV-2 in Vero E6 cells is 1.76 µM, which is achievable in vivo based on a trial in nonhuman primate experiment [63,69] [...]

Table 7. *Cont.*

Extracted triples using our pipeline:				
subject	**predicate**	**object**	**subj_entity_name**	**obj_entity_name**
coronaviruses with high similarity to the human SARS - CoV or civet SARS - CoV - like virus	were isolated	from horseshoe bats	(’coronaviruses’, ’high’, ’human’, ’SARS-CoV’, ’civet’, ’SARS-CoV-like’)	(’horseshoe’)
horseshoe bats as the potential natural reservoir of SARS-CoV	is	the natural host of SARS-CoV-2	(’horseshoe’, ’natural’, ’SARS-CoV’)	(’natural’, ’SARS-CoV-2’)
masked palm civets	are	the intermediate host	(’palm’)	(’intermediate’, ’host’)
the receptor-binding domain of SARS-CoV-2	demonstrates	a similar structure to that of SARS-CoV	(’receptor-binding’, ’SARS-CoV-2’)	(’structure’, ’SARS-CoV’)
a few variations in the key residues	exist	at amino acid level	(’variations’, ’residues’)	(’amino’)
Pangolin	to be	a potential intermediate host for SARS-CoV-2	(’pangolin’)	(’potential’, ’intermediate’, ’host’, ’SARS-CoV-2’)
The EC90 of remdesivir against SARS - CoV-2 in Vero E6 cells	is	1.76 BμM, which is achievable in vivo	(’EC90’, ’remdesivir’, ’SARS’, ’CoV-2’)	(’in’)

Sample of OIE triples (ClausIE) directly on article body without using the pipeline (115 triples in total):		
subject	**predicate**	**object**
It	is	reasonable to suspect that bat is the natural host of SARS-CoV-2 considering its similarity with SARS-CoV Subsequently
its	has	similarity
SARS-CoV-2	belonged	to the *Betacoronavirus* genera
SARS-CoV-2	was	closer to SARS-like coronavirus
Zhou and colleagues	showed	that SARS-CoV-2 had 96.2% overall genome sequence identity throughout the genome to BatCoV RaTG13
SARS-CoV-2	was felled	in a basal position within the subgenus *Sarbecovirus* in the ORF1b tree
This finding	implies	a possible recombination event
SARS-CoV and MERS-CoV respectively	is	49
that snake might be the intermediate host	is	61
Guangxi Customs of China	found	novel coronaviruses representing two sub-lineages related to SARS-CoV-2 respectively
RaTG13	is	89.20%
Interestingly the pangolin coronavirus and SARS-CoV-2	share	identical amino acids

Table 7. *Cont.*

Sample of OIE triples (ClausIE) directly on article body without using the pipeline (115 triples in total):		
subject	**predicate**	**object**
RaTG13	possesses	one
The effective concentration for chloroquine and remdesivir	is	EC 50
Chloroquine	functions	at both viral entry and post-entry stages of the SARS-CoV-2 infection in Vero E6 cells whereas remdesivir does at post-entry stage only
potential broad-spectrum antiviral activities	is	64 65
An EC90 against the SARS-CoV-2 in Vero E6 cells	is	achievable clinically
Remdesivir	is	effective
1.76 Γ,BμM	is	achievable

6. Conclusions

In this paper, a pipeline for efficient open information extraction, entity enrichment, and graph representation from unstructured text was presented. We analyzed the rationale and functioning of each preprocessing component comprising our methodology—namely, in-place coreference resolution that was applied on the raw CORD-19 corpus to replace pronouns with their original references and extractive summarization which was subsequently used to reflect the diverse topics of each article while keeping redundancy to a minimum. We integrated a parallel triple extraction process based on different OIE engines, relying on the complementarity of different information retrieval approaches (clause-based, learning-based, embeddings-based, etc.) to counter the loss of structural and semantic information. Finally, we combined a number of post-processing and information enrichment tasks, including entity linking with the Unified Medical Language System, polarity detection, triple deduplication, and continuity representation to enhance the usefulness and readability of the extracted triples before visualizing them in a graph database.

We implemented our approach on a large research dataset (CORD-19) consisting of thousands of scientific articles to illustrate its efficiency and we acquired more than 400,000 valid triples linked with UMLS entities and other relevant metadata, ensuring that the generated information remains relevant to the context of the ingested corpus and free of syntactic variations that would lead to triples of low contextual value. Our information extraction pipeline was evaluated in terms of precision, recall, and F1-score and, while a direct comparison would require the evaluation of similar systems on the same dataset, it has shown promising results compared to standalone OIE engines and other end-to-end frameworks. Finally, we demonstrated its capabilities through a series of indicative data exploration tasks for retrieving different types of information by querying or visually interacting with the graph database. In terms of design, we adapted a modular approach, basing each component of our pipeline on state-of-the-art tools and pretrained deep-learning models, hoping that it contributes to its flexibility and future-proofness.

In the future, we plan to further expand our methodology by introducing additional pre-processing and post-processing features, including mapping the extracted triples to formal semantic schemas such as the Comparative Toxicogenomics Database (CTD) COVID-19 database [77], in order to render it compliant with the existing ontology-guided storage systems. Moreover, we plan to exploit the modularity of our approach to experiment with other state-of-the-art tools and compare it on benchmark datasets and shared tasks [85]. Finally, with regard to the validity of the pipeline's output, we plan on involving human experts from the biomedical domain to assess the informativeness of the extracted triples and the usefulness of the produced graphs.

Author Contributions: Conceptualization, D.P., N.P., and A.L.; formal analysis, D.P.; methodology, D.P. and N.P.; software, D.P.; supervision, N.P.; validation, D.P.; writing—original draft, D.P.; writing—review and editing, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: The research work of D.P. was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 50, 2nd call). The research work of N.P. and A.L. was supported by the "INODE—Intelligent Open Data Exploration" project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 863410.

Acknowledgments: We would like to thank the anonymous reviewers for their comments and suggestions on an earlier version of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Extractive text summarization example: The highlighted text composes the summary of the processed article.

CORD-19 Article ID: d99dbae98cc9705d9b5674bb6eb66560b4434305
The current epidemic of a new coronavirus disease (COVID-19), caused by a novel coronavirus (2019-nCoV), recently officially named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has reopened the issue of the role and importance of coronaviruses in human pathology (1) (2) (3) (4) (5). This epidemic definitively confirms that this heretofore relatively harmless family of viruses, Coronaviridae, includes major pathogens of epidemic potential. The COVID-19 epidemic has clearly demonstrated the power of infectious diseases, which have been responsible for many devastating epidemics throughout history. The epidemiological potential of emerging infectious diseases, especially zoonoses, is affected by numerous environmental, epidemiological, social, and economic factors (6, 7). Emerging zoonoses pose both epidemiological and clinical challenges to health care professionals. Since the 1960s, coronaviruses have caused a wide variety of human and animal diseases. In humans, they cause up to a third of all community-acquired upper respiratory tract infections, such as the common cold, pharyngitis, and otitis media. However, more severe forms of bronchiolitis, exacerbations of asthma, and pneumonia in children and adults have also been described, sometimes with fatal outcomes in infants, the elderly, and the immunocompromised. Some coronaviruses are associated with gastrointestinal disease in children. Sporadic infections of the central nervous system have also been reported, although the role of coronaviruses in infections outside the respiratory tract has not been completely clarified (8). Most coronaviruses are adapted to their hosts, whether animal or human, although cases of possible animal-to-hu-man transmission and adaptation have been described in the past two decades, causing two epidemics. The first such outbreak originated in Guangdong, a southern province of the People's Republic of China, in mid-November of 2002. The disease was named severe acute respiratory syndrome (SARS). The cause was shown to be a novel coronavirus (SARS-CoV), an animal virus that had crossed the species barrier and infected humans. The most likely reservoir was bats, with evidence that the virus was transmitted to a human through an intermediate host, probably a palm civet or raccoon dog (8, 9). In less than a year, SARS-CoV infected 8098 people in 26 countries, of whom 774 died (10, 11). Approximately 25% of the patients developed organ failure, most often acute respiratory distress syndrome (ARDS), requiring admission to an intensive care unit (ICU), while the case fatality rate (CFR) was 9.6%. However, in elderly patients (>60 years), the CFR was over 40%. Poor outcomes were seen in patients with certain comorbidities (diabetes mellitus and hepatitis B virus infection), patients with atypical symptoms, and those with elevated lactic acid dehydrogenase (LDH) values on admission. Interestingly, the course of the disease was biphasic in 80% of the cases, especially those with severe clinical profiles, suggesting that immunological mechanisms, rather than only the direct action of SARS-CoV, are responsible for some of the complications and fatal outcomes (8, 9). Approximately 20% of the reported cases during this epidemic were health care workers. Therefore, in addition to persons exposed to animal sources and infected family members, health care workers were among the most heavily exposed and vulnerable individuals (9, 10). During 2004, three minor outbreaks were described among laboratory personnel engaged in coronavirus research. Although several secondary cases, owing to close personal contact with infected patients, were described, there was no further spread of the epidemic. It is not clear how the SARS-CoV eventually disappeared and if it still circulates in nature among animal reservoirs. Despite ongoing surveillance, there have been no reports of SARS in humans worldwide since mid-2004 (11). In the summer of 2012, another epidemic caused by a novel coronavirus broke out in the Middle East. The disease, often complicated with respiratory and renal failure, was called Middle East respiratory syndrome (MERS), while the novel coronavirus causing it was called Middle East respiratory syndrome coronavirus (MERS-CoV). Although a coronavirus, it is not related to the coronaviruses previously described as human pathogens. However, it is closely related to a coronavirus isolated from dromedary camels and bats, which are considered the primary reservoirs, albeit not the only ones (8, 12). From 2012 to the end of January 2020, over 2500 laboratory-confirmed MERS cases, including 866 associated deaths, were reported worldwide in 27 countries (13). The largest number of such cases has been reported among the elderly, diabetics, and patients with chronic diseases of the heart, lungs, and kidneys. Over 80% of the patients required admission to the ICU, most often due to the development of ARDS, respiratory insufficiency requiring mechanical ventilation, acute kidney injury, or shock. The CFR is around 35%, and even 75% in patients >60 years of age. However, MERS-CoV, unlike its predecessor SARS-CoV, did not disappear, but still circulates among animal and human populations, occasionally causing outbreaks, either in connection with exposure to camels or infected persons (12). Overall, 19.1% of all MERS cases have been among health care workers, and more than half of all laboratory-confirmed secondary cases were transmitted from human to human in health care settings, at least in part due to shortcomings in infection prevention and control (12, 13). Post-exposure prophylaxis with ribavirin and lopinavir/ritonavir decreased the MERS-CoV risk in health care workers by 40% (14). THe eMeRGence oF covId-19 cAused by sARs-cov-2In mid-December of 2019,

Table A1. *Cont.*

CORD-19 Article ID: d99dbae98cc9705d9b5674bb6eb66560b4434305
a pneumonia outbreak erupted once again in China, in the city of Wuhan, the province of Hube (1). The outbreak spread during the next two months throughout the country, with currently over 80 000 cases and more than 2400 fatal outcomes (CFR 2.5%), according to official reports. Exported cases have been reported in 30 countries throughout the world, with over 2400 registered cases, of which 276 are in Europe. On February 25, the first case of COVID-19 was confirmed in Zagreb, Croatia, and was linked to the current outbreak in the Lombardy and Veneto regions of northern Italy (15). The case definition was first established on January 10 and modified over time, taking into account both the virus epidemiology and clinical presentation. The clinical criteria were expanded on February 4 to include any lower acute respiratory diseases, and the epidemiological criterion was extended to the whole of China, with the possibility of expansion to some surrounding countries (16, 17). At the early stage of the outbreak, patients' full-length genome sequences were identified, showing that the virus shares 79.5% sequence identity with SARS-CoV. Furthermore, 96% of its whole genome is identical to bat coronavirus. It was also shown that this virus uses the same cell entry receptor, ACE2, as SARS-CoV (18). The full clinical spectrum of COVID-19 ranges from asymptomatic cases, mild cases that do not require hospitalization, to severe cases that require hospitalization and ICU treatment, and those with fatal outcomes. Most cases were classified as mild (81%), 14% as severe, and 5% as critical (ie, respiratory failure, septic shock, and/or multiple organ dysfunction or failure). The overall CFR was 2.3%, while the rate in patients with comorbidities was considerably higher −10.5% for cardiovascular disease, 7.3% for diabetes, 6.3% for chronic respiratory diseases, 6.0% for hypertension, and 5.6% for cancer. The CFR in critical patients was as high as 49.0% (4).It is still not clear which factors contribute to the risk of transmitting the infection, especially by persons who are in the incubation stage or asymptomatic, as well as which factors contribute to the severity of the disease and fatal outcome. Evidence from various types of additional studies is needed to control the epidemic (19). However, it is certain that the binding of the virus to the ACE 2 receptor can induce certain immunoreactions, and the receptor diversity between humans and animal species designated as SARS-CoV-2 reservoirs further increases the complexity of COVID-19 immunopathogenicity (20). Recently, a diagnostic RT-PCR assay for the detection of SARS-CoV-19 has been developed using synthetic nucleic acid technology, despite the lack of virus isolates and clinical samples, owing to its close relation to SARS. Additional diagnostic tests are in the pipeline, some of which are likely to become commercially available soon (21). Currently, randomized controlled trials have not shown any specific antiviral treatment to be effective for COVID-19. Therefore, treatment is based on symptomatic and supportive care, with intensive care measures for the most severe cases (22). However, many forms of specific treatment are being tried, with various results, such as with remdesivir, lopinavir/ritonavir, chloroquine phosphate, convalescent plasma from patients who have recovered from COVID-19, and others (23) (24) (25) (26). No vaccine is currently available, but researchers and vaccine manufacturers have been attempting to develop the best option for COVID-19 prevention. So far, the basic target molecule for the production of a vaccine, as well as therapeutic antibodies, is the CoV spike (S) glycoprotein (27, 28). The spread of the epidemic can only be contained and SARS-CoV-2 transmission in hospitals by strict compliance with infection prevention and control measures (contact, droplet, and airborne precautions) (22, 29). During the current epidemic, health care workers have been at an increased risk of contracting the disease and consequent fatal outcome owing to direct exposure to patients. Early reports from the beginning of the epidemic indicated that a large proportion of the patients had contracted the infection in a health care facility (as high as 41%), and that health care workers constituted a large proportion of these cases (as high as 29%). However, the largest study to date on more than 72 000 patients from China has shown that health care workers make up 3.8% of the patients. In this study, although the overall CFR was 2.3%, among health care workers it was only 0.3%. In China, the number of severe or critical cases among health care workers has declined overall, from 45.0% in early January to 8.7% in early February (4). This poses numerous psychological and ethical questions about health care workers' role in the spread, eventual arrest, and possible consequences of epidemics. For example, during the 2014-2016 Ebola virus disease epidemic in Africa, health care workers risked their lives in order to perform life-saving invasive procedures (intravenous indwelling, hemodialysis, reanimation procedures, mechanical ventilation), and suffered high stress and fatigue levels, which may have prevented them from practicing optimal safety measures, sometimes with dire consequences (30). This third coronavirus epidemic, caused by the highly pathogenic SARS-CoV-2, underscores the need for the ongoing surveillance of infectious disease trends throughout the world. The examples of pandemic influenza, avian influenza, but also the three epidemics caused by the novel coronaviruses, indicate that respiratory infections are a major threat to humanity. Although Ebola virus disease and avian influenza are far more contagious and influenza currently has a greater epidemic potential, each of the three novel coronaviruses require urgent epidemiologic surveillance. Many infectious diseases, such as diphtheria, measles, and whooping cough, have been largely or completely eradicated or controlled through the use of vaccines. It is hoped that developments in vaccinology and antiviral treatment, as well as new preventive measures, will ultimately vanquish this and other potential threats from infectious diseases in the future.

References

1. Augenstein, I.; Padó, S.; Rudolph, S. *LODifier: Generating Linked Data from Unstructured Text BT—The Semantic Web: Research and Applications*; Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 210–224.
2. Clancy, R.; Ilyas, I.F.; Lin, J. Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond. In Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), Hong Kong, China, 3–7 November 2019; pp. 39–46.
3. Kríž, V.; Hladká, B.; Nečaský, M.; Knap, T. *Data Extraction Using NLP Techniques and Its Transformation to Linked Data BT—Human-Inspired Computing and Its Applications*; Gelbukh, A., Espinoza, F.C., Galicia-Haro, S.N., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 113–124.
4. Pertsas, V.; Constantopoulos, P.; Androutsopoulos, I. *Ontology Driven Extraction of Research Processes BT—The Semantic Web—ISWC*; Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.-A., Simperl, E., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 162–178.
5. Exner, P.; Nugues, P. Entity Extraction: From Unstructured Text to DBpedia RDF triples. In Proceedings of the WoLE@ISWC, Boston, MA, USA, 11–15 November 2012.
6. Holzinger, A.; Kieseberg, P.; Weippl, E.; Tjoa, A.M. Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Swizterland, 2018.
7. Xiao, Z.; de Silva, T.N.; Mao, K. Evolving Knowledge Extraction from Online Resources. *Int. J. Comput. Syst. Eng.* **2017**, *11*, 746–752. [CrossRef]
8. Makrynioti, N.; Grivas, A.; Sardianos, C.; Tsirakis, N.; Varlamis, I.; Vassalos, V.; Poulopoulos, V.; Tsantilas, P. PaloPro: A platform for knowledge extraction from big social data and the news. *Int. J. Big Data Intell.* **2017**. [CrossRef]
9. Wu, H.; Lei, Q.; Zhang, X.; Luo, Z. Creating A Large-Scale Financial News Corpus for Relation Extraction. In Proceedings of the 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 28–31 May 2020; pp. 259–263.
10. Amith, M.; Song, H.Y.; Zhang, Y.; Xu, H.; Tao, C. Lightweight predicate extraction for patient-level cancer information and ontology development. *BMC Med. Inform. Decis. Mak.* **2017**. [CrossRef] [PubMed]
11. Wang, X.; Li, Q.; Ding, X.; Zhang, G.; Weng, L.; Ding, M. A New Method for Complex Triplet Extraction of Biomedical Texts. In *International Conference on Knowledge Science, Engineering and Management*; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2019.
12. Haihong, E.; Xiao, S.; Song, M. A text-generated method to joint extraction of entities and relations. *Appl. Sci.* **2019**, *9*, 3795. [CrossRef]
13. Kertkeidkachorn, N.; Ichise, R. T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text. In Proceedings of the AAAI Workshops, San Francisco, CA, USA, 4–5 February 2017.
14. Freitas, A.; Carvalho, D.S.; Da Silva, J.C.P.; O'Riain, S.; Curry, E. A semantic best-effort approach for extracting structured discourse graphs from wikipedia. In Proceedings of the CEUR Workshop Proceedings, Boston, MA, USA, 11–15 November 2012.
15. Blomqvist, E.; Hose, K.; Paulheim, H.; Lawrynowicz, A.; Ciravegna, F.; Hartig, O. The Semantic Web: ESWC 2017 Satellite Events. In Proceedings of the ESWC 2017 Satellite Events, Portorož, Slovenia, 28 May–1 June 2017.
16. Elango, P. *Coreference Resolution: A Survey*; Technical Report; UW-Madison: Madison, WI, USA, 2006.
17. Kantor, B.; Globerson, A. Coreference resolution with entity equalization. In Proceedings of the ACL 2019—57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
18. Hobbs, J.R. Resolving pronoun references. *Lingua* **1978**. [CrossRef]
19. Grosz, B.; Weinstein, S.; Joshi, A. Centering: A Framework for Modeling the Local Coherence of Discourse. *Comput. Linguist.* **1995**, *21*, 203–225.

20. Wiseman, S.; Rush, A.M.; Shieber, S.M.; Weston, J. Learning anaphoricity and antecedent ranking features for coreference resolution. In Proceedings of the ACL-IJCNLP 2015—53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Beijing, China, 26–31 July 2015.
21. Clark, K.; Manning, C.D. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2256–2262.
22. Lee, K.; He, L.; Lewis, M.; Zettlemoyer, L. End-to-end Neural Coreference Resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 188–197.
23. Joshi, M.; Chen, D.; Liu, Y.; Weld, D.S.; Zettlemoyer, L.; Levy, O. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 64–77. [CrossRef]
24. Andhale, N.; Bewoor, L.A. An overview of Text Summarization techniques. In Proceedings of the 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 12–13 August 2016; pp. 1–7.
25. Allahyari, M.; Pouriyeh, S.A.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K.J. Text Summarization Techniques: A Brief Survey. *arXiv* **2017**, arXiv:abs/1707.02268. [CrossRef]
26. Nallapati, R.; Zhou, B.; dos Santos, C.; Gulçehre, Ç.; Xiang, B. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Proceedings of the CoNLL 2016—20th SIGNLL Conference on Computational Natural Language Learning, Proceedings, Berlin, Germany, 11–12 August 2016.
27. Nallapati, R.; Xiang, B.; Zhou, B.; Question, W.; Algorithms, A.; Heights, Y. Sequence-To-Sequence Rnns For Text Summarization. In Proceedings of the International Conference on Learning Representations, ICLR 2016 - Workshop Track, San Juan, Puerto Rico, 2–4 May 2016.
28. Kouris, P.; Alexandridis, G.; Stafylopatis, A. Abstractive Text Summarization Based on Deep Learning and Semantic Content Generalization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July 2019; pp. 5082–5092.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
30. Hoang, A.; Bosselut, A.; Çelikyilmaz, A.; Choi, Y. Efficient Adaptation of Pretrained Transformers for Abstractive Summarization. *arXiv* **2019**, arXiv:1906.00138.
31. Filippova, K.; Altun, Y. Overcoming the lack of parallel data in sentence compression. In Proceedings of the EMNLP 2013—2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013.
32. Colmenares, C.A.; Litvak, M.; Mantrach, A.; Silvestri, F. HEADS: Headline generation as sequence prediction using an abstract: Feature-rich space. In Proceedings of the NAACL HLT 2015—2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015.
33. Shimada, A.; Okubo, F.; Yin, C.; Ogata, H. Automatic Summarization of Lecture Slides for Enhanced Student Preview-Technical Report and User Study. *IEEE Trans. Learn. Technol.* **2018**. [CrossRef]
34. Hassel, M. Exploitation of Named Entities in Automatic Text Summarization for Swedish. In Proceedings of the NODALIDA'03–14th Nordic Conferenceon Computational Linguistics, Reykjavik, Iceland, 30–31 May 2003.
35. Pal, A.R.; Saha, D. An approach to automatic text summarization using WordNet. In Proceedings of the 2014 IEEE International Advance Computing Conference (IACC), Gurgaon, India, 21–22 February 2014; pp. 1169–1173.
36. Miller, D. Leveraging BERT for Extractive Text Summarization on Lectures. *arXiv* **2019**, arXiv:1906.04165.
37. Wang, Q.; Liu, P.; Zhu, Z.; Yin, H.; Zhang, Q.; Zhang, L. A Text Abstraction Summary Model Based on BERT Word Embedding and Reinforcement Learning. *Appl. Sci.* **2019**, *9*, 4701. [CrossRef]
38. Zheng, H.-T.; Guo, J.-M.; Jiang, Y.; Xia, S.-T. Query-Focused Multi-document Summarization Based on Concept Importance. In *Advances in Knowledge Discovery and Data Mining*; Bailey, J., Khan, L., Washio, T., Dobbie, G., Huang, J.Z., Wang, R., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 443–453.

39. Gupta, S.; Gupta, S.K. Abstractive summarization: An overview of the state of the art. *Expert Syst. Appl.* **2019**, *121*, 49–65. [CrossRef]
40. Jurafsky, D.; Martin, J.H. Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing. In *Language*; Prentice Hall: Upper Saddle River, NJ, USA, 2007; ISBN 0131873210.
41. Niklaus, C.; Cetto, M.; Freitas, A.; Handschuh, S. A Survey on Open Information Extraction. In Proceedings of the 27th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2018; pp. 3866–3878.
42. Fader, A.; Soderland, S.; Etzioni, O. Identifying relations for Open Information Extraction. In Proceedings of the EMNLP 2011—Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011.
43. Mesquita, F.; Schmidek, J.; Barbosa, D. Effectiveness and efficiency of open relation extraction. In Proceedings of the EMNLP 2013–2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013.
44. Banko, M.; Cafarella, M.J.; Soderland, S.; Broadhead, M.; Etzioni, O. Open information extraction from the web. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Hyderabad, India, 6–12 January 2007.
45. Wu, F.; Weld, D.S. Open Information Extraction Using Wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp. 118–127.
46. Weld, D.S.; Hoffmann, R.; Wu, F. Using Wikipedia to bootstrap open information extraction. *Acm Sigmod Rec* **2009**, *37*, 62–68. [CrossRef]
47. Del Corro, L.; Gemulla, R. ClausIE: Clause-Based Open Information Extraction. In Proceedings of the 22nd International Conference on World Wide Web; Association for Computing Machinery: New York, NY, USA, 2013; pp. 355–366.
48. Angeli, G.; Premkumar, M.J.; Manning, C.D. Leveraging linguistic structure for open domain information extraction. In Proceedings of the ACL-IJCNLP 2015—53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Beijing, China, 26–31 July 2015.
49. Yang, Z.; Salakhutdinov, R.; Cohen, W.W. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. *arXiv* **2017**, arXiv:1703.06345.
50. Liu, Y.; Zhang, T.; Liang, Z.; Ji, H.; McGuinness, D.L. Seq2RDF: An End-to-end Application for Deriving Triples from Natural Language Text. *arXiv* **2018**, arXiv:abs/1807.0.
51. He, L.; Lee, K.; Lewis, M.; Zettlemoyer, L. Deep semantic role labeling: What works and what's next. In Proceedings of the ACL 2017—55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017.
52. Stanovsky, G.; Michael, J.; Zettlemoyer, L.; Dagan, I. Supervised open information extraction. In Proceedings of the NAACL HLT 2018—2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018.
53. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A nucleus for a Web of open data. In *The Semantic Web*; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2007.
54. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 9–12 June 2008.
55. Suchanek, F.M.; Kasneci, G.; Weikum, G. Yago: A core of semantic knowledge. In Proceedings of the 16th International World Wide Web Conference, Banff, AB, Canada, 8–12 May 2007.
56. Mena, E.; Kashyap, V.; Illarramendi, A.; Sheth, A. Domain Specific Ontologies for Semantic Information Brokering on the Global Information Infrastructure. In Proceedings of the Formal Ontology in Information Systems: Proceedings of FOIS'98, Trento, Italy, 6–8 June 1998.
57. Karadeniz, I.; Özgür, A. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC Bioinform.* **2019**. [CrossRef]
58. Cho, H.; Choi, W.; Lee, H. A method for named entity normalization in biomedical articles: Application to diseases and plants. *BMC Bioinform.* **2017**. [CrossRef]

59. Papadakis, G.; Tsekouras, L.; Thanos, E.; Giannakopoulos, G.; Palpanas, T.; Koubarakis, M. Domain- and Structure-Agnostic End-to-End Entity Resolution with JedAI. *SIGMOD Rec.* **2020**, *48*, 30–36. [CrossRef]
60. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends. Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
61. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**. [CrossRef]
62. Sentiment Classification using Machine Learning Techniques. *Int. J. Sci. Res.* **2016**. [CrossRef]
63. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:abs/1907.1.
64. McBride, B. *The Resource Description Framework (RDF) and Its Vocabulary Description Language RDFS BT—Handbook on Ontologies*; Staab, S., Studer, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 51–65. ISBN 978-3-540-24750-0.
65. Angles, R. The Property Graph Database Model. In Proceedings of the AMW, Cali, Colombia, 21–25 May 2018.
66. Kohlmeier, S.; Lo, K.; Wang, L.L.; Yang, J.J. *COVID-19 Open Research Dataset (CORD-19)*. 2020. Available online: https://pages.semanticscholar.org/coronavirus-research (accessed on 18 April 2020).
67. Allen Institute for AI Coreference Resolution Demo. Available online: https://demo.allennlp.org/coreference-resolution (accessed on 18 April 2020).
68. Weischedel, R.; Palmer, M.; Marcus, M.; Hovy, E.; Pradhan, S.; Ramshaw, L.; Xue, N.; Taylor, A.; Kaufman, J.; Franchini, M.; et al. OntoNotes Release 5.0 LDC2013T19. *Linguist. Data Consort.* **2013**. [CrossRef]
69. Beltagy, I.; Lo, K.; Cohan, A. {S}ci{BERT}: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3615–3620.
70. University of Washington; Indian Institute of Technology Open IE 5.1. Available online: https://github.com/dair-iitd/OpenIE-standalone (accessed on 10 June 2020).
71. Max Planck Institute for Informatics ClausIE: Clause-Based Open Information Extraction. Available online: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/software/clausie/ (accessed on 18 April 2020).
72. Allen Institute for AI Open Information Extraction. Available online: https://demo.allennlp.org/open-information-extraction (accessed on 18 April 2020).
73. Schmitz, M.; Bart, R.; Soderland, S.; Etzioni, O. Open language learning for information extraction. In Proceedings of the EMNLP-CoNLL 2012—2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 12–14 July 2012.
74. Saha, S. Mausam Open Information Extraction from Conjunctive Sentences. In Proceedings of the 27th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2018.
75. Pal, H.-M. Demonyms and Compound Relational Nouns in Nominal Open IE. In Proceedings of the 5th ACL Workshop on Automated Knowledge Base Construction, San Diego, CA, USA, 12–17 June 2018.
76. Saha, S.; Pal, H. Mausam Bootstrapping for numerical open IE. In Proceedings of the ACL 2017—55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017.
77. Christensen, J.; Soderland, S.; Etzioni, O. An Analysis of Open Information Extraction Based on Semantic Role Labeling Categories and Subject Descriptors. In Proceedings of the Sixth International Conference on Knowledge Capture, Banff, AB, Canada, 25–29 July 2011; pp. 113–120.
78. Neumann, M.; King, D.; Beltagy, I.; Ammar, W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *arXiv* **2019**, arXiv:1902.07669.
79. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**. [CrossRef]
80. Kormilitzin, A.; Vaci, N.; Liu, Q.; Nevado-Holgado, A. Med7: A transferable clinical natural language processing model for electronic health records. *arXiv* **2020**, arXiv:2003.01271.
81. Neo4j Neo4j—The Leader in Graph Databases. Available online: https://neo4j.com/ (accessed on 18 April 2020).
82. Gotti, F.; Langlais, P. Weakly Supervised, Data-Driven Acquisition of Rules for Open Information Extraction. In *Canadian Conference on Artificial Intelligence*; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2019.

83. Yuan, S.; Yu, B. An Evaluation of Information Extraction Tools for Identifying Health Claims in News Headlines. In Proceedings of the Workshop Events and Stories in the News, Santa Fe, NM, USA, 21–25 August 2018.
84. Davis, A.P.; Grondin, C.J.; Johnson, R.J.; Sciaky, D.; McMorran, R.; Wiegers, J.; Wiegers, T.C.; Mattingly, C.J. The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Res.* **2018**, *47*, D948–D954. [CrossRef]
85. Lever, J.; Jones, S.J. VERSE: Event and Relation Extraction in the BioNLP 2016 Shared Task. In Proceedings of the 4th BioNLP Shared Task Workshop, Berlin, Germany, 13 August 2016.