BioChem

MDPI

Review

The Treasury Chest of Text Mining: Piling Available Resources for Powerful Biomedical Text Mining

Nícia Rosário-Ferreira [1,2,*], Catarina Marques-Pereira [2,3], Manuel Pires [2,4], Daniel Ramalhão [2,4], Nádia Pereira [2], Victor Guimarães [4,5], Vítor Santos Costa [4,5,*] and Irina Sousa Moreira [6,7,*]

1 CQC-Coimbra Chemistry Center, Chemistry Department, Faculty of Science and Technology, University of Coimbra, 3004-535 Coimbra, Portugal
2 CIBB, University of Coimbra, 3000-456 Coimbra, Portugal; catarina.103@gmail.com (C.M.-P.); manuelmoreirapires@hotmail.com (M.P.); dani.ramalhao.97@gmail.com (D.R.); nadia.pereira.nnp@gmail.com (N.P.)
3 IIIs-Institute for Interdisciplinary Research, University of Coimbra, 3000-456 Coimbra, Portugal
4 Department of Sciences, University of Porto, 4169-007 Porto, Portugal; victorguimaraes13@gmail.com
5 INESC-TEC-Centre of Advanced Computing Systems, 4169-007 Porto, Portugal
6 Department of Life Sciences, University of Coimbra, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal
7 CNC-Center for Neuroscience and Cell Biology, CIBB-Center for Innovative Biomedicine and Biotechnology, University of Coimbra, 3004-535 Coimbra, Portugal
* Correspondence: nicia.ferreira@student.uc.pt (N.R.-F.); vsc@dcc.fc.up.pt (V.S.C.); irina.moreira@cnc.uc.pt (I.S.M.)

Abstract: Text mining (TM) is a semi-automatized, multi-step process, able to turn unstructured into structured data. TM relevance has increased upon machine learning (ML) and deep learning (DL) algorithms' application in its various steps. When applied to biomedical literature, text mining is named biomedical text mining and its specificity lies in both the type of analyzed documents and the language and concepts retrieved. The array of documents that can be used ranges from scientific literature to patents or clinical data, and the biomedical concepts often include, despite not being limited to genes, proteins, drugs, and diseases. This review aims to gather the leading tools for biomedical TM, summarily describing and systematizing them. We also surveyed several resources to compile the most valuable ones for each category.

Keywords: text mining; biomedical articles; artificial intelligence; deep learning; machine learning

Citation: Rosário-Ferreira, N.; Marques-Pereira, C.; Pires, M.; Ramalhão, D.; Pereira, N.; Guimarães, V.; Costa, V.S.; Moreira, I.S. The Treasury Chest of Text Mining: Piling Available Resources for Powerful Biomedical Text Mining. *BioChem* **2021**, *1*, 60–80. https://doi.org/10.3390/biochem1020007

Academic Editor: Yehia Mechref

Received: 12 June 2021
Accepted: 14 July 2021
Published: 27 July 2021

1. Introduction

Text mining is already widely used, mainly on social media, for, e.g., Twitter, to explore disease symptoms [1], reactions to public regulations [2], or to study the opioid crisis [3]. Moreover, it has been increasingly applied in various industries, such as the financial sector for decision-making processes [4]. Big text data in this sector, from websites or even social media, has been used on stock price prediction, financial fraud detection, and market forecast [4]. In the health sector, particularly in diabetes, corpus-based terminology from online texts, manuals, or professional papers have been automatically extracted. These were used to developed a specific terminology list of patients whilst browsing phrases to compare professional and common terminology and evaluate statistics of different terminologies in two different languages [5]. In the pharmaceutical sector, automatic terminology extraction from pharmaceutical documents of meaningful information has been applied to classify documents [6].

This review aims at providing a broad perspective of the major developments in biomedical text mining. First, in Section 1, the basis of text mining and some fundamental definitions will be clarified. In this scope, the focus will be mainly on natural language processing (NLP) methods, a field that allies artificial intelligence (AI), linguistic and computer science (CS) methodologies. Thenceforth, in Section 2, the most recent and

promising resources (i.e. as corpora) as well as models to perform text mining will be gathered. The build of dedicated to the biomedical field NLP systems and algorithms (BioNLP) integrated in the biologist workflow, highlights text mining value as a crucial method for biomedical science. Lastly, in Section 3, some questions and future directions regarding the field will be addressed.

1.1. What Is Biomedical Text Mining

The amount of data available from scientific papers, patents, or other sources of information, particularly in the biomedical field, is continuously rising. Often, these data are unstructured and not ready for computational interpretation. Using text mining in the biomedical field has widely increased due to the emergent need to analyze and gain knowledge from large data sources [7]. Text mining provides a set of automated methods that can distill text from heterogeneous sources into actionable data. Text mining applies NLP methods to extract and retrieve information from text just like a human reader would. An NLP model should understand the language, semantics, and vocabulary to predict token features correctly [8]. BioNLP, the application of NLP models in the biomedical field, adds the required knowledge of specific biological contexts [7,9]. Text mining comes upon the first step of automated information retrieval (IR) to retrieve all the information relevant to a specific problem from disperse data resources [7,10].

Biomedical text mining, at its core, comprises three stages: named entity recognition (NER), named entity normalization (NEN), and relation extraction (RE) (Figure 1).

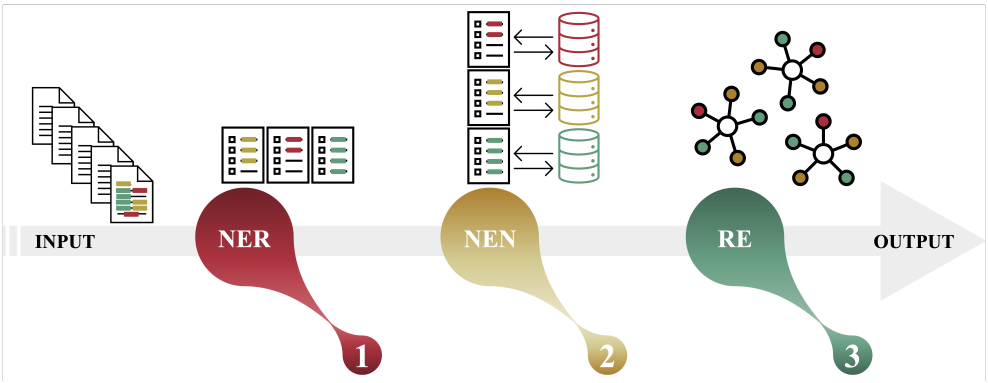

Figure 1. Biomedical text mining pipeline: A comprehensive text mining pipeline generally encompasses these 3 steps: named entity recognition (NER), named entity normalization (NEN) and Relation Extraction (RE). These steps can be interdependent and are hierarchical in this form. Still, each step can be dealt with separately in which case the input needed varies.

NER, also known as 'entity tagging' or 'concept extraction', is fundamental for automatically extracting information from text [11]. Biomedical NER aims to retrieve relevant biomedical entities such as genes, drugs, diseases, species, proteins, mutations, and cell lines existent in natural language documents and tag each word´s location and class [11,12]. Hence, this step identifies concepts and keywords, categorizing them in user-defined classes. NER tools implement the text pre-processing stage, where all data are cleaned and tokenized, a step in which typically words are broken down into words or sub-word tokens to build the vocabulary of such text. Then, all remaining unique tokens are processed through different methods to extract features that represent biomedical classes to be transformed into a suitable representation [13]. Given a text, a NER tool will categorize, for example, the words 'breast cancer', 'fever', 'migraine' or 'hepatites' as diseases and 'human', 'rat', 'patient' or 'person' as species. Usually, these tools categorize the keywords as genes/proteins, diseases, species, cell lines, and mutations. Several articles explain this step more deeply, present more complex examples, either biomedical related or in other areas, and explain the different phases, approaches and methods [13,14].

After accomplishing the NER step, NEN algorithms are invoked to semantics and coherence for all the retrieved tokens, solving the disambiguation. As such, constitute an essential step in the automated construction of a biomedical database describing and relating concepts, which can be organized either as a hierarchy or as a set of relationships. Abbreviation recognition and synonym recognition are advantageous to unify and normalize biomedical terms [12]. Biomedical NEN intends to map entity terms in biomedical text to typical entities in a particular knowledge base, i.e., a database which compiles information about a topical domain in a hierarchical or relationship manner [15]. Furthermore, NEN models can exhibit additional steps such as abbreviation resolution, in which acronyms are reformed to the original long words by using the abbreviation dictionaries [16]. So, after the NER step, NEN will normalize the terms, for example, the tool will recognize the term 'IL6' as the abbreviation of 'Interleukin 6' while the NER step only associates 'IL6' with the category 'gene' or 'protein'. For further information on this step and more detailed examples, the readers may explore other articles, both biomedical and other topics. Various articles describe this step and all its complexity [17–19].

Lastly, RE is a task that aims to identify syntactic and semantic relations between the entities that originated in the previous text mining tasks automatically [20,21]. Basic RE methods encompassed simple systems based on co-occurrence statistics that evolved to more intricate ones using syntactic analysis and machine learning (ML)/deep learning (DL) models [21,22]. The extracted relations are expressed in a machine-understandable format ready for post-text mining analysis [23]. In the biomedical field, relations among entities are pivotal towards understanding complex biological mechanisms by being able to retrieve new relations from previously known ones. The extraction of homo and heterogeneous interactions between chemicals, diseases, genes, proteins, and/or other classes is needed to decipher new knowledge mainly in the fields of, e.g., regulatory pathways, metabolic processes, or adverse drug relations [20,21]. For example, to the sentence 'Individuals with a BRCA1 gene mutation are more likely to develop breast cancer at a younger age', NER can recognize that BReast CAncer gene 1 (BRCA1) is a gene and that 'breast cancer' is a disease and categorize them like that. NEN is responsible to disambiguate this previous step, to categorize all keywords correctly and recognize the term 'BRCA1' as the abbreviation of 'breast cancer susceptibility gene 1'. Lastly, RE is able to associate the BRCA1 gene with the disease 'breast cancer'. The process of relations extraction can be difficult and there are several different approaches in this phase. Some articles explain these approaches more in-depth and use thorough examples [24–26].

1.2. Text Mining Challenges—What Makes Text Mining Complex?

Part of text mining complexity lies in the fact that different sources compile data in different formats, which often requires specific techniques [7,11]. These types of data frequently lack common structural frameworks and can have errors like improper grammar, spelling errors or semantic ambiguities. Text errors increase the complexity of data pre-processing and text mining analysis [7,11,27]. The recognition and mapping of certain terms in the NER and NEN steps can also be troublesome. In fact, Biomedical NER is usually considered more challenging since there are numerous difficulties for automatic identification of biomedical terms due to irregularities in how known entities are entitled [11,12]. Common challenges arise when terms are not a part of the ontology used, as misspellings or ambiguity in the term´s designations can occur. Hence, to deal with this issue, choosing the right corpus and/or ontology is crucial. This is particularly true for genes and proteins where nomenclature is frequently messier since proteins and genes can share the same abbreviation and different ontologies may have different spellings [7,15]. However, this type of heterogeneity and ambiguity can also happen in key classes such as drugs or chemicals [10,15]. Correctly choosing a corpus to train a text mining model and then being able to retrieve relations from text is an intricate task due to the complexity of grammatical construction hindering the machine retrieval of relations from text and, at the

same time, incorporation of data from external sources can foster the advances of the RE step [28].

Lastly, biological knowledge is complex, and the lack of certain specific information can compile conflicting answers. For instance, the same species under different conditions (e.g., age, gender, treatments) may not have the same biological system and what happens in one species may not happen in another. These differences, if not noted, may lead to different answers upon text mining application [29].

1.3. Traditional Versus Machine Learning Driven Text Mining

Traditional text mining approaches consist of finding patterns in the evaluated text based on previously known patterns of interest. Arguably, the simplest way to mine text is by using dictionaries in order to find words of interest in the text, and grammatical rules to find the relation between those words. Both techniques rely on the existence of dictionaries containing the words of interest and grammatical rules for the given language. As such, it struggles to find new patterns different from the already known ones [30].

The advance in ML methods made possible the application of such methods to text mining, allowing a step forward in their performance. ML algorithms are fed with labeled text, where the results are known, to learn a model able to connect the text with the results, in such a way that it can generalize to new unknown text. This advance allows the user to train models, based on already known classified text, to generalize over new text, instead of explicitly defining the rules and words to be looked for in the text. One caveat of this technique is the requirement of a large amount of labeled data, already classified text, which might be unavailable for the task at hand.

Figure 2 shows an example of a sentence annotated for the three tasks: NER, NEN, and RE. The annotation for NER is at the token level and usually follows the IOB format, which stands for inside, outside, and beginning, respectively. Tokens marked as O do not belong to any entity, while the first token of an entity is marked as B, and the remaining tokens of the entity are marked as I. In the case of a multi-class NER, the class name can be attached after the B and I tokens, such as B-Disease, B-Gene, B-Chemical, among others. The labels for the NEN task are at the entity level and may represent a sequence of tokens, for example, "breast cancer", which is represented by two tokens. In both NER and NEN, tokens marked as O are not part of any entity. Finally, the label for RE is at the sentence level and states whether two entities is a sentence are related (or unrelated) to each other. In this example, it states that the "breast cancer" disease is related to the "BRCA1" gene.

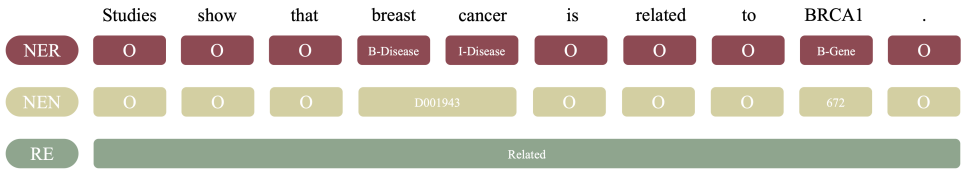

Figure 2. Example of a sentence labeled for Named Entity Recognition (NER), Named Entity Normalization (NEN) and Relation Extraction (RE).

To process text using ML models, the text must first be converted to a numeric form. With the advances of DL and Neural Networks (NN), word2vec [31] is a very popular model that represents words as embeddings, that is, vectors of numbers relating words to other words or concepts. These vectors were computed by using co-occurrences in a sentence, allowing the system to obtain interesting relationships without labeled data. It allows the use of a much larger set of text to initially train the model, and then one can use a smaller amount of labeled text to fine-tune to the final task. This technique improves the performance of the models and allows the user to transfer learning the initial vectors between different tasks [32]. More recently, several have obtained good results with DL mechanisms. These systems combine embeddings with attention mechanisms, that allow the system to focus on elements of the embedding. By using different attention

mechanisms to construct encoder/decoders, Bidirectional Encoder Representations from Transformers (BERT) [33] and more recent systems such as Embedding from Language Models (ELMo) [34] have been too achieved impressive performance. More biological specific, there is a BERT version trained on biomedical text, called Bidirectional Encoder Representations from Transformers for Biomedical text mining (BioBERT), which will be detailed in the next sections [35]. For example, the occurrence of words in the text can be used to create a labeled data set containing approximately correct examples. To further enhance the number of labeled data for ML methods, one can use other simpler methods, such as dictionaries and rule-based methods. This technique is sometimes referred to as weak or distantly supervised approach and was already used in works such as Wright et al. [36].

2. Resources for Text Mining

2.1. Biomedical Corpora

Models of biomedical text mining require specific data to be successfully and effectively trained and evaluated. Therefore, the development and implementation of novel resources to train and evaluate these algorithms have become a fundamental process to obtain better results and unveil new insights within the biomedical field. Due to the rapid growth of biomedical literature, literature knowledge annotation and extraction are becoming increasingly demanding [35,37]. Indeed, one of the biggest challenges of text mining in the biomedical field of research is the construction of appropriate biomedical corpora, the construction of a set of annotated texts and/or sentences that establish terms and/or relationships within a domain [38,39]. General corpora challenges lie on the different word distribution, as well as the domain-dependent type of terms and expressions of the area whose meaning is similar and rarely appears in documents from other domains [35,37]. The annotation process is a crucial step in biomedical corpora construction since when poorly made, severely hinders the accuracy and benefit of biomedical text mining tools [38]. This annotation process can be performed through a manual curation based on a guideline, which, although known as the gold standard corpus due to higher-level outcomes, is time-consuming and implies knowledge in linguistic and semantic fields.

Annotated corpora are crucial for both training and evaluation of text mining tasks since their text is enriched by adding features and patterns [40]. Nevertheless, it is necessary to take into consideration the type of annotations existent in the corpus as well as the task to be performed. If the goal is to use an annotated corpus for NER, then the corpus must have the entities of interest for a given category annotated to be able to identify terms for that category [41]. However, in the biological context more than knowing which entities are involved in a process, it is often more valuable to retrieve relations from the text to understand how these entities interact in order to gain knowledge of the biological system [42]. Thus, to perform RE, annotated corpora with the same type of relationship between entities and their characteristics must be used [43]. In the Biomedical field, most of the frequently used corpora are focused on genes and/or proteins. These may include GENome Information Acquisition (GENIA) ([44]), Colorado Richly Annotated Full Text Corpus (CRAFT) [45], and Critical Assessment of Information Extraction in Biology (BIOCREATIVE) II [46] corpora. Furthermore, depending on the purpose of the study other corpora can be used, thus contributing to a wider range of categories, such as diseases—National Center for Biotechnology Information (NCBI) disease corpus [47]) and chemicals—CHEMicals Disease Named Entity Recognition (CHEMDNER) [48]).

Corpora are often published individually and, therefore, difficult to gather. However, BioCreative https://biocreative.bioinformatics.udel.edu/resources/corpora/ (accessed on 1 July 2021), HUNER https://hu-ner.github.io/corpora.html (accessed on 1 July 2021), and Corposaurus https://corposaurus.github.io/corpora/ (accessed on 1 July 2021) are great centralized sources for corpora. Table 1 provides a list of the most important corpus in the biomedical field are provided.

Table 1. List of available corpora useful specifically for biomedical text mining tasks.

Corpus	Description	Number of Documents	Annotations	References
BC5CDR	BioCreative V Chemical Disease Relation—disease and chemicals annotations for chemicals-disease interactions retrieval	1500 PubMed abstracts	Diseases, chemicals, chemical-disease interactions	[49]
BRONCO	Biomedical entity Relation ONcology COrpus—focused on cancer research and anti-tumor drug screening containing 400 genomic variants and their relation to genes, diseases, drugs, and cell lines from 108 full-text articles	108 full-length articles	Variants, genes, diseases, cell lines, drugs	[50]
CellFinder	A corpus based on stem cells	10 full-length articles	Anatomical parts, cell components, cell lines, cell types, genes/proteins, species, several binary relationships, biological processes	[51]
ChemDNER	CHEMicals Disease Named Entity Recognition—focused on chemical substances and their characteristics	10,000 PubMed abstracts	Chemicals	[48]
ChemDNER patents	CHEMicals Disease Named Entity Recognition patents—focused on detecting mentions in running patent text. Manual annotation led to 2 gold standard corpora Chemical Entity Mention in Patents (CEMP) and Gene and Protein Related Object (GPRO)	21,000 medicinal chemistry-related patents abstracts	Chemicals, gene and gene products	[52]
CoMAGC	Corpus with Multi-faceted Annotations of Gene-Cancer relations—focused on gene-cancer relations (namely regarding prostate, breast, and ovarian cancers)	821 sentences from 408 documents	Change in gene expression, change in cell state, proposition type, and initial gene expression level	[53]
CRAFT	Colorado Richly Annotated Full-Text Corpus	97 full-length articles	Chemicals, cell types, biological processes, cellular and extracellular components and regions, molecular function, chemical reactions, biological taxa, proteins, biomacromolecular entities and sequences, anatomical entities	[54]
DDI corpus	Drug-Drug Interactions—focused on pharmacological substances and their relationships	792 texts from DrugBank database and 233 MEDLINE abstracts	Pharmacological substances; DDIs	[55]
GENIA	GENome Information Acquisition—focused on biological reactions base on transcription factors in human blood cells	2000 MEDLINE abstracts	47 biologically relevant nominal categories	[44]

Table 1. *Cont.*

Corpus	Description	Number of Documents	Annotations	References
GETM	GeneExpression Text Miner corpus—gold standard corpus focused on gene expression events and their anatomical locations	150 MEDLINE abstracts	Genes, anatomical locations	[56]
GNI corpus	GeNomics And Informatics - Originally developed to identify trends from publications from the Genomics and Informatics Journal	499 full texts from Genomics and Informatics Journal	Proteins; DNA; RNA; cell lines; cell types	[57]
GREC	Gene Regulation Event Corpus—developed to train text mining systems to extract biologically meaningful events	240 MEDLINE abstracts	Biological events (13 semantic roles) and biological concepts (10 categories for *E. coli* and 10 categories for Human)	[43]
MedTag	A biomedical corpus that combines 3 corpora: MedPost, ABGene and GENETAG	MedPost: 6700 sentences; ABGene: 4265 sentences; GENETAG: 15,000 sentences	Genes, proteins, clinical medicine semantics	[58]
MLEE	Multi-Level Event Extraction—focused on event extraction	262 PubMed abstracts on angiogenesis	3 entity categories: organism, anatomy (11 subcategories), and molecule (2 subcategories); and 4 event types: anatomical (7 subcategories), molecular (6 subcategories), general (5 subcategories), and planned	[59]
NCBI disease corpus	National Center for Biotechnology Information disease corpus—A corpus to disease recognition	793 PubMed abstracts	Diseases	[47]
new corpus	Unnamed corpus meant to automatize curation of the ChEBI database	200 abstracts and 100 full-text articles	6 entities (metabolites, chemicals, proteins, species, biological activities, spectral data) and 4 relations (isolated from, associated with, binds with, metabolite of)	[60]
NLM-Chem	National Library of Medicine—Chemical—gold standard dataset focused on chemical NER	150 Pubmed full-text articles	Chemicals	[61]
NLM-Gene	National Library of Medicine—Gene—gold standard dataset focused on gene NER	550 Pubmed full-text articles	Gene	[62]
PGR	Phenotype-Gene Relations—a silver standard corpus based on human genes and phenotype, as well as their relations	1712 abstracts	Genes, human phenotypes and phenotype-gene relations	[63]
Variome	A corpus focused on the relationship between the inherited colorectal cancer and human genetic variation	10 articles	11 entities and 13 relations	[64]

2.2. Text Mining Toolkits

Performing text mining tasks from the ground up can be complex. Toolkits are key steps for a simpler implementation of text mining complex tasks, such as text processing

(tokenization, stemming, part-of-speech tagging), NER, NEN, and RE, without losing versatility, necessary for a high-quality approach. These general toolkits can then be adapted for specific contexts, in this case, the biomedical domain. General Architecture for Text Engineering (GATE) was originally released in 1999 and nowadays evolved to encompass a family of tools [65]. GATE tackles text mining and NLP problems and comprises three main components: Gate Document Manager (GDM), Collection of REusable Objects for Language Engineering (CREOLE), and GATE Graphical Interface (GGI). GDM acts as a storage for all the information created by the Language Engineering systems. CREOLE does the actual text processing and analysis. Existing algorithms can be implemented using wrappers around those methods. However, it is also possible to develop approaches not included in the toolkit through CREOLE's API. GGI, as the name suggests, provides a graphical interface for the various tools and resources provided in GATE. This workflow was successfully applied to several fields like life sciences and medicine, such as cancer research, medical records analysis, and drug patent-related research [66]. Unstructured Information Management Applications (UIMA) is yet another software with an architecture based on four main components, namely acquisition, unstructured information analysis on a document level, unstructured information analysis on a collection level, and structured information access. The acquisition is used to retrieve the documents to process and allows the implementation of external applications. Document-level analysis performs a wide range of tasks in each document, such as language translation, grammatical parsing, named-entity detection, document summarization, and document classification. Collection-level analysis can be performed on a whole collection or sub-collection of documents to infer common characteristics between documents such as glossaries of terms, taxonomies, feature vectors, databases of extracted relations, and detected entities. The structured information access component allows browsing knowledge obtained from documents or searching available methods to perform these tasks. Semantic search uses the document-level and collection-level analysis and annotations to return an ordered list of documents. This toolkit also presents a directory service to browse through the different text processing tools and knowledge source adapters, a tool for uniform access to several knowledge sources in this architecture. Recombining Unstructured Information Management (UIM) technology, UIMA is capable to accelerate scientific advances through text mining allowing the development of other tools [67]. ClearTK is a toolkit for statistical NLP, developed for UIMA in 2008 [68]. First, it performs feature extraction on an annotated corpus using a wide range of methods with different complexities. Then, it passes the features onto the training data consumer, used to generate a training data file that can then be further used in an ML model building library. In 2014, ClearTK 2.0 was released as further development and adaptation according to the feedback community of the original toolkit [69]. UIMA was also successfully used for clinical diagnosis [70], as well as a wrapper for an annotator [71]. Both GATE and UIMA approaches are quite complex and need a deep understanding of their architectures, not only for their use but also for the development of applications that can be used within those architectures.

BioC is a simple workflow for NLP and text mining tasks, originally implemented in C++ but extended to Python, Perl, Go and Ruby. BioC toolkit is based in XML, and converts from this format into BioC data classes and vice-versa using two connectors, one input connector and one output connector. Between the input and the output, it is possible to perform several text processing tasks. A good application example for this toolkit was the release of the PubMed Central (PMC) corpus in the XML BioC format, allowing easier text retrieval tasks without losing information present in the document set. This format might also decrease the necessary learning curve for researchers getting started with text mining [72]. As it is a widely used toolkit, new adaptations were released such as tmBioC, which makes the necessary changes to other algorithms, such as DNorm [73], tmVar [74], SR4GN [75], tmChem [76], GenNorm [77], and PubTator [78]. As such, these become compatible with BioC, which further increases the application potential of BioC [79]. BioC has been used in web text mining tools [80,81] as well as in ontologies and annotations [82,83].

Some tools used in the biomedical domain lack an adaptation for the specific scientific context. Due to this, these tools are mainly used for pre-processing purposes, such as tokenization of sentences or words, for instance, Stanford Core NLP and Natural Language ToolKit (NLTK) toolkits. Stanford Core NLP was originally developed for in-house use, nonetheless, it was later released to provide a tool with a simpler architecture that did not require deep knowledge to be used. It was implemented in Java and can be run either from the API or the command line. It provides a wide range of annotators, speech tagging, or NER, among others. This toolkit was designed for English and Chinese languages, but models for other languages can be easily constructed. Like with languages, other annotators can also be added to a pipeline within Stanford Core NLP [84]. This toolkit has been used for processing in biomedical text mining pipelines [85]. NLTK was developed in Python, to solve some of the challenges inherent with the teaching of computational linguistics and NLP. As such, it was developed to be simple to use, work with consistent data structures, be easily expanded through the development of new tools and have detailed documentation. Furthermore, NLTK includes models from the biomedical domain corpus regarding protein-protein interactions as, for, e.g., BioCreative-protein-protein interaction corpus http://www.nltk.org/nltk_data/ (accessed on 1 July 2021). Hence, this toolkit is capable of performing a wide range of NLP operations regarding the biomedical field, such as tokenization, parsing, token tagging, and text classification. In fact, NLTK was the toolkit used for text pre-processing in the GeNomics & Informatics (GNI) corpus which highlights its usefulness in the biomedical field despite its broad scope application [57].

2.3. Text Mining Tools for NER, NEN, and RE

Even though there are available a variety of tools to perform text mining NER, the most widely accepted State-Of-The-Art (SOTA) model is BioBERT, a model based on Google's BERT model, pre-trained on biomedical corpora encompassing PubMed abstracts and PMC full-text articles. For NER, BioBERT uses bidirectional transformers and directly learns WordPiece embeddings during pre-training and fine-tuning, improving its performance [35]. HUNER is a stand-alone NER tool that was trained in 34 corpora for five entity types as chemicals, cell lines, diseases, genes, and species whilst using scientific literature and patents. HUNER was also evaluated on the CRAFT corpus outperforming previous SOTA tools as GNormPlus and tmChem by 5–13 pp on chemicals, species, and genes entities types. HUNER model uses Long Short Term Memory Conditional Random Fields (LSTM-CRF) to learn feature correlations and predicting the entities' tags [86]. To facilitate its use and forgo Docker, HunFlair was released, combining an improved HUNER where a bidirectional LSTM-CRF (biLSTM-CRF) model was used along with the implementation of the Flair NLP framework. This eased its use even for inexperienced users [87]. To tackle the challenges in NER related to spelling errors mainly found in clinical records, Cimind, a NER tool was developed. The authors developed a new dataset that provides multiple versions of the 10th revision of the International Classification of Diseases (ICD) in both English and French, so the system can rely on a dataset to store the double metaphone code for every word in each available language [88]. Cimind allows overcoming spelling errors through a phonetic approach.

Neural Biomedical named Entity Recognition and multi-type Normalization (BERN) is a NER and NEN joint tool, which encompasses the NER model from the SOTA BioBERT, and a high-performance NEN model for each entity type in a single step. Moreover, BERN also applies probability-based decision rules to the entities retrieved from the NER stage to differentiate coinciding ones. Besides its command-line use via GitHub repository, BERN is also available as a RESTful Web service. Other methods were developed to tackle specific problem areas as genetic variants, SETH, and diseases, AuDis. SETH is a tool that performs NER and NEN in natural language text specific to genetic variants, facilitating the identification of genetic variants, attiring these variants to established nomenclatures, and linking them to databases covering multiple mutation types. SETH's modular implementation makes it simple to substitute the gene recognition tool to be

used, so it can adapt to other types of texts [89]. AuDis is a disease NER and NEN tool that applies CRF based model for the NER task optimized with multiple post-processing steps and improved abbreviation resolution and stopwords filtering. The disease mentions were normalized to specific concepts in an existing repository and the authors developed a dictionary-lookup method [90].

Disease-Expression Relation Extraction from Text (DEXTER) [91] is an RE tool that extracts information from the literature regarding gene and microRNA expression in specific diseases-related framework returning expression level, experimental context, and the compared conditions. DEXTER first step of NER is performed using the Stanford CoreNLP toolkit to gather all entities that are genes, microRNAs, and diseases. Then, the sentences retrieved undergo a search for trigger words, words that are specific to the DEXTER field. Next, parsing is used to extract relations and build a triplets entity1-relation-entity2 Standard Dependency Graph (SDG). DEXTER is both available as a stand-alone tool or as part of BioExpress resource [92]. Protein-protein association Extraction with Deep Language (PEDL) is another RE tool to predict protein–protein associations using a Multi-Instance Learning (MIL) framework. PEDL is a two-step approach, which, first, uses a BERT model to extract information and then finds relationship predictions from the transformer layers using CLaSsification (CLS) tokens. PEDL uses distantly supervised data to retrieve all protein pairs and relations from Protein Interaction Database (PID) and directly supervised data from gold standard datasets from the BioNLP-shared tasks. PEDL was able to extend the knowledge present in pathways databases by predicting additional protein-protein associations which is a strong indicator of the usefulness of such approaches [93]. Biomedical Relation (BioRel), a dataset for distantly supervised RE, is a full text mining approach encompassing all text mining steps. BioRel uses Medline as corpus and Unified Medical Language System (ULMS), specifically MetathesauRus RELationships (MRREL), as Knowledge Database (KB) since it includes binary relations. Relationships were assessed through National Drug File—Reference Terminology (NDFRT) and the relation between genes and cancer were retrieved from National Cancer Institute (NCI) to add to the MRREL vocabulary. MetaMap was used to retrieve and normalize entities from the corpus to UMLS. Further feature extraction was performed using the StanfordNLP tool and further distantly supervised annotation labels were created before the final steps of filtering and building the final dataset. Moreover, BioRel was tested and showed itself as a useful resource for training Deep Neural Networks (DNN) models [23].

Despite the methods highlighted in this section, a recap of useful text mining tools is detailed listed in Table 2. Biocreative, National Center for Text Mining (NaCTeM), and NCBI also provide related resources in http://biocreative.sourceforge.net/bionlp_tools_links.html, http://www.nactem.ac.uk/software.php, and https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/ (accessed on 1 July 2021), respectively, where several text mining tools are compiled.

Table 2. List of NLP models for different text mining steps.

Tool/Model	NER	NEN	RE	Class	Architecture	Reference
AuDis	X	X	—	Diseases	CRF	[90]
BERN	X	X	—	Chemical, Disease, Specie, Gene/Protein, and Mutation	NER—Transformers (BioBERT based)	[94]
BioBERT	X	—	X	Fine-tuning is available for desired classes when criteria are met	Transformers	[35]

Table 2. *Cont.*

Tool/Model	NER	NEN	RE	Class	Architecture	Reference
BioRel	X	X	X	Entities: Clinical drugs, pharmacologic substance, organic chemical, disease or syndrome, biologically active substance, molecular function, food, organ or tissue function and neoplastic process; Relations: 124 classes + NA	DNN-based approaches and distance supervised learning	[23]
Cimind	X	—	—	Diseases	Double metaphone phonetic algorithm and weighted distance scale algorithm	[88]
CLRG	X	—	—	Chemical compounds and their specific types	CRF and Artificial Neural Networks (ANN)	[95]
CollaboNet	X	—	—	Chemical, gene/protein, and disease	BiLSTM-CRF	[96]
D3NER	X	—	—	Any type of entity as long as criteria are met	CRF-biLSTM	[97]
DEXTER	X	—	X	Gene/microRNA, disease, and expression information	Standard dependency graph	[91]
Dnorm	X	X	—	Diseases		[73]
HunFlair	X	—	—	Cell line, chemical, disease, gene and species	BiLSTM- CRF	[87]
GNormPlus	X	X	—	Gene, gene family, protein domain	CRF	[98]
NeuroNER	X	—	—	Chemicals, disease, species and gene/protein	LSTM-CRF	[99]
PEDL	X	X	X	Protein-protein association	Multi-instance learning (NER and NEN are BERT-based)	[93]
REflex	X	—	X		CNN	[100]
Saber	X	X	—	Chemical, disorder, organism, gene/gene product	BiLSTM-CRF	[101]
SciSpacy	X	X	—	Depends on the chosen model from the repository	spaCy based	[102]
SETH	X	X	—	Genetic variants		[89]
tmChem	X	X	—	Chemicals, drugs	CRF	[76]
tmVar	X	X	—	Variants at protein and gene level	CRF	[74]
VinAI	X	—	—	Chemicals	BiLSTM-CNN-CRF	[103]
NA	X	—	—	29 entities types from 15 datasets	Multi-Task Learning (MTL)-BC and MTL-LBC	[104]
NA	X	—	—	Chemicals, diseases, species, genes/proteins and cell lines	BiLSTM-CRF	[105]
NA	—	X	—	Diseases	CNN	[15]

2.4. Web-Based Applications

Despite the accuracy and broader application of the previous tools presented, less experienced users tend to prefer web-based applications to perform their analysis. Hence, there are several web servers available to retrieve key concepts from biomedical data.

PubTator Central (PTC) is a freely accessible, daily updated server able to automatically annotate more than 30 M PubMed abstracts and more than 3 M full articles from the PMC-TM subset. PTC is commonly used in biocuration support, gene prioritization, genetic disease analysis, literature-based knowledge discovery, or downstream text mining. This web server is capable of recognizing tokens and classify them into six categories as genes or proteins, genetic variants, diseases, chemicals, and cell lines. Each of these categories was trained using the same taggers as PubTator or re-trained to increase performance whenever a new corpus was available as in the case of variants that used tmVar 2.0, and chemicals that used an improved version of TaggerOne via re-training with the BioCreative V Chemical Disease Relation (BC5CDR) and CHEMDNER corpora. Normalization was conduced with a Convolutional Neural Network (CNN), able to identify the correct bioconcept through the syntax and semantics of the surrounding words. This CNN was trained with human-curated databases attaining accuracy that is human-comparable. Annotated articles and abstracts are freely available using a raw text input through the command line, but also in the PTC web server (https://www.ncbi.nlm.nih.gov/research/pubtator/ [106]) (accessed on 1 July 2021). This web tool has been used in several COrona VIrus Disease 2019 (COVID-19)-related works [107–109].

SciLite is another web server that helps users to find essential concepts in documents and correlate them to available resources and tools (https://www.scilit.net/) (accessed on 1 July 2021). This server is not only capable of identifying genes or proteins, organisms, diseases, GO terms, chemicals, and accession numbers but it is also capable to link these concepts to related databases and provide more information on these concepts without leaving the webserver page. For instance, it can retrieve protein structures corresponding to Protein Data Bank (PDB) accession numbers in documents. SciLite also has an evaluation mechanism in which users can confirm or report annotations to improve text mining algorithms [110]. SciLite has been used to improve access to protein motif articles [111] as well as in annotators [112].

Textpresso is another web server with two main goals: separate a group of full articles into individual sentences and categorize terms in article databases and sentences in order to be easily searched (https://textpressocentral.org/tpc) (accessed on 1 July 2021). Although this web server is dedicated to *Caenorhabditis elegans* literature, it may be extended to other organisms. Words can be classified into several biological concepts such as genes, alleles, cell groups, and phenotypes, or can be correlated as associations, regulations or as biological processes. When combined, these classes form an ontology and the article corpus can be marked with words of these classifications. Users can search for one or several keywords from these classes and the webserver provides sentences of articles with those words to help users to select relevant articles. At the moment, Textpresso has more than 2.5 million full-text articles from several corpora. Overall, Textpresso helps users to identify important articles and focuses on article information related to the user's query [113]. This tool has been used to accelerate the annotation process for the creation of knowledge graphs [30] as well as for annotation and curation in more complex pipelines [114].

Egas https://demo.bmd-software.com/egas/ (accessed on 1 July 2021) is a web-based tool designed to be user-friendly with an extensive interface developed with six main components: project management, project and document navigators, processing tools, account management, concept and relation type filters, and real-time collaboration. Usually, this tool's workflow begins with the selection and import of study documents that can be local files in raw text, A1 and BioC format, or a query for PubMed abstracts as well as PubMed Central full-text documents. If the query is used, a list of documents for selection is presented to the user [115]. After document selection, the next step is to automatically annotate these retrieved documents and for that, Biomedical Concept Annotation System (BeCAS) REST API [116] is used to annotate a range of biomedical entities such as genes, proteins, or drugs. To analyze protein-protein interactions, an ML model was deployed to recognize protein names, using BioThesaurus [117] for normalization and then a rule-based approach for protein-protein interactions recognition [115]. Annotation results are then

displayed in the document viewer and annotated documents can be exported in A1 or BioC formats [115]. This web tool was used, for example, in work for semiautomatic curation of text data related to a rare disease [118].

PolySearch2 is a text mining web tool available in http://polysearch.ca/index (accessed on 1 July 2021) that can effectively relate two entities in a "given X, find all associated Y" type query [119]. PolySearch2 mines several sources encompassing MEDLINE and PMC papers, Wikipedia, US patents, open-access textbooks, and Medline Plus comprising a 43 million articles text collection with further integration of 13 public databases [119]. The entities available for such requests range from human diseases, genes, single nucleotide polymorphisms, proteins, drugs, metabolites, toxins, metabolic pathways, organs, tissues, subcellular organelles, positive health effects, negative health effects, and drug actions to integration to ontology terms from a plethora biological and chemical taxonomies [119]. The dictionaries that enable this type of query include over 1.13 million terms and 2.84 synonyms gathered from various sources [119]. Polysearch2 was used as an evaluation tool in [120]. It was also used for searching proteins related to liver cancer [121].

Finding Associated Concepts with Text Analysis (FACTA)+ available at http://www.nactem.ac.uk/facta/ (accessed on 1 July 2021) was released to expand the original FACTA version [122] and fill the need for a tool that identifies and explores a range of associations [123] adding to the previous webserver. Originally, FACTA allowed users to obtain entities relevant and related to a query [122]. Searches are performed using a MEDLINE query based on input queries that can be a word, an ID or the combination of both, and results presented are within six categories: human genes or proteins, diseases, symptoms, drugs, enzymes, or chemical compounds. Concept identification in documents is accomplished by a dictionary-based approach that involves Universal Protein resource (UniProt) [124], BioThesaurus [117], UMLS [125], Human Metabolome DataBase (HMDB) [126], Kyoto Encyclopedia of Genes and Genomes (KEGG) [127], and DrugBank [128]. In FACTA+, three new features were added from the recognition of biomolecular events to the discovery of associations and an enhanced result's visualization. To address the event's recognition, it is important to detect triggers, words considered as indicative of a relation between two entities, done by FACTA+ via an ML-based approach for NER which uses CRF models. To discover hidden associations, FACTA+ considers that if a central entity is related to two different entities, then these two entities may be related among them as well. Due to the high noise attaining these associations, it is of the utmost importance to correctly rank these possible indirect association favoring their retrieval whilst maintaining them reliably. To incorporate these hidden associations, a new treemap visualization scheme for the directly associated concepts and treemap with linking where co-occurrences are the relation strength measure the indirectly associated concepts [123]. This web tool has been used to extract information about the indirect interactions between post-translational modifications of histone proteins [129].

A wide range of tools is available online to help researchers automatically annotate biomedical literature, resorting to text mining techniques. However, there are two main sub-tasks, RE and scoring functions, that still need to be further improved in the next years.

2.5. Public Databases That Incorporate Text Mining Models

Databases that gather biological information often incorporate information retrieved via a text mining approach to widening their data. This is the case for several databases such as Search Tool for Retrieval of Interacting Genes/Proteins (STRING), Search Tool for Interacting Chemicals (STITCH), microRNA-Target interactions dataBase (miRTarBase), Biological General Repository for Interaction Datasets (BioGRID), and DisGeNet.

STRING database seeks to gather, score and incorporate available protein-protein interaction data and complement this information with computational predictions [130]. STRING ultimately aims to establish a global network with proteins' direct physical interactions and indirect functional interactions. Nowadays, this database includes protein-protein interaction information from 5090 different organisms and more than 24.6 million proteins.

Two proteins are related if they are functionally associated, meaning, they both contribute to a particular biological function. To be considered functionally associated, they may interact physically or share a specific cellular pathway. These associations can be established through genomic, co-expression, text mining, biochemical experiments, or pathway information. Each association has a score and a number of views associated to evaluate the protein-protein interaction information and give a confidence estimation. Text mining associations are obtained through a statistical co-citation analysis from more than 28 million PubMed, Medline, and Online Mendelian Inheritance in Man (OMIM) full articles and abstracts [130].

STITCH aims to integrate protein-protein interaction and protein-chemical interactions into a single database with a global network for each organism [131]. This database displays 430 k different chemicals and the correspondent binding affinities for users to get an analysis of the chemical´s impact on a protein of interest. STITCH database combines experimental data from ChEMBL [132], Psychoactive Drug Screening Program (PDSP) K_i Database [133] and PDB [134], computational predictions and information from manually curated datasets such as DrugBank [135], GPCR-ligand database [136], Matador [137], Therapeutic Targets Database [138], Comparative Toxicogenomics Database [139], and pathway databases like KEGG [140], Reactome [141] and BioCyc [142]. Redundant interactions along these datasets are only counted once to improve the interaction confidence level and the final score is computed based on the strongest described binding affinity. Information from these experimental and manually curated datasets are integrated with text mining predictions, involving co-occurrence text mining and NLP from MEDLINE and Research Portfolio Online Reporting Tools (RePORTER) abstracts and PubMed full articles. STITCH later version filters, for each organism tissue, proteins, and chemicals that are not associated with that particular tissue [131].

MiRTarBase is an online database for interactions between gene targets and microRNAs, non-coding RNAs of 18–25 nucleotides that regulate gene expression [143]. This database has more than 479,000 curated microRNA–target interactions (MTIs) with more than 4000 microRNAs and more than 23,000 target genes from more than 11,000 manually curated articles, identified through a text mining system with a scoring system. Several databases and tools were integrated into the miRTarBase database to improve accuracy such as gene information, microRNA regulators information, disease information, and gene and microRNA expression [143].

BioGRID database aims to curate and store human and model organisms´ protein, genetic and chemical interactions [144]. This database has more than 1.74 million biological interactions from more than 70 species manually curated from more than 55 thousands articles and other databases. BioGRID takes interaction data from experimental expressions and is guided by text mining approaches.

DisGeNet is a knowledge management platform that establishes genes and genomic variants associations with human diseases [145]. This comprehensive database includes over 24,000 diseases and traits, 17,000 genes, and 117,000 genomic variants, as well as over 625,000 gene-disease associations and over 115,000 variant-disease associations. Both gene-disease and variant-disease associations were extracted based on gene or variant list similarity as well as using text mining tools applied to scientific literature via the Literature-derived Human Gene-Disease Network (LHGDN) or BEFREE text mining resources. These tools can identify as well as standardize entities and relationships and review linguistics and semantics to identify relationships between genotype and phenotype [145].

3. Future Perspectives

As the number of available biological and biomedical literature repositories is quickly increasing, the manual curation of every publication is becoming tougher and the prioritization of important experimental publications is becoming harder. This leads to a necessary integration of text mining methods in a daily researcher life as it allows publication scoring regarding articles interaction information and automates articles annotations [144]. The

main caveat related to text mining resources lies in the lack of centralization of such resources. This fact severely hinders tool comparison and even the finding of such tools regardless of the user proficiency. Approaches to gathering resources and creating a centralized database must be a priority. To this end, an attempt to find and aggregate such resources was made by Amália Lourenço's Lab [146]. In this work, over 135 active websites were found and characterized regarding text mining tools. Despite all the results presented, similar work must be done for repositories, universities' websites, and biomedical literature. This highlights the need for, as above mentioned, the creation of a centralized and permanently updated database of such tools. Recently, bio.tools by Elixir website, a centralized option that is user-dependent for the intended input, encompasses many NLP https://bio.tools/t?topicID=%22topic_0218%22 (accessed on 1 July 2021) and text mining https://bio.tools/t?operationID=%22operation_0306%22 (accessed on 1 July 2021) resources.

Text mining tools blooming is also boosted by the improvement of DL algorithms and their ability to provide new insights. However, the RE step is still far from being resolved. To address these difficulties, text mining competitions as BioCreative, Biomedical Natural Language Processing Workshop (BioNLP), or Biomedical Linked Annotation Hackathon (BLAH) take place, most of them, yearly to encourage discussion and push the boundaries of text mining. New methods to improve the incorporation of data from KB, new corpora suitable to an increasing range of subjects and new models towards a higher accuracy RE step are, generally speaking, the first demands.

Systems Biology is a hot area nowadays despite its complexity due to the inherent difficulties to hoard vast insights from related fields such as the ones provided by Omics methodologies. Text mining can effectively connect information from different sources, integrate it, and provide an accurate way to visualize the results often providing even deeper insights and, hence streamlining this broader area.

Author Contributions: Conceptualization, I.S.M.; writing—original draft preparation, N.R.-F., C.M.-P., M.P., D.R., V.G. and N.P.; writing—review and editing, N.R.-F., V.S.C. and I.S.M.; supervision, I.S.M.; project administration, I.S.M. and V.S.C.; funding acquisition, I.S.M. and V.S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by COMPETE 2020—Operational Programme for Competitiveness and Internationalisation and Portuguese national funds via FCT—Fundação para a Ciência e a Tecnologia, under projects POCI-01-0145-FEDER-031356 and UIDB/04539/2020. Authors would also like to acknowledge STRATAGEM—New diagnostic and therapeutic tools against multidrug-resistant tumors, CA17104.

Acknowledgments: N.R.-F, V.G. and C.M.-P. were also supported by FCT through Ph.D. scholarships PD/BD/135179/2017, 2020.05718.BD and 2020.07766.BD (DOCTORATES 4 COVID-19), respectively.

Conflicts of Interest: The authors declare no conflict of interest. The funding agencies had no role in the design of the study, in the collection, analyzes, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Guo, J.W.; Radloff, C.L.; Wawrzynski, S.E.; Cloyes, K.G. Mining twitter to explore the emergence of COVID-19 symptoms. *Public Health Nurs.* **2020**, *37*, 934–940. [CrossRef] [PubMed]
2. Lazard, A.J.; Wilcox, G.B.; Tuttle, H.M.; Glowacki, E.M.; Pikowski, J. Public reactions to e-cigarette regulations on Twitter: A text mining analysis. *Tobacco Control* **2017**, *26*, e112–e116. [CrossRef] [PubMed]
3. Nasralah, T.; El-Gayar, O.; Wang, Y. Social Media Text Mining Framework for Drug Abuse: Development and Validation Study With an Opioid Crisis Case Analysis. *J. Med. Internet Res.* **2020**, *22*, e18350. [CrossRef] [PubMed]
4. Bach, M.P.; Krstić, Ž.; Seljan, S.; Turulja, L. Text Mining for Big Data Analysis in Financial Sector: A Literature Review. *Sustainability* **2019**, *11*, 1277. [CrossRef]
5. Seljan, S.; Baretić, M.; Kučiš, V. Information retrieval and terminology extraction in online resources for patients with diabetes. *Coll. Antropol.* **2014**, *38*, 705–710.

6. Seljan, S.; Stančić, H.; Dunđer, I. Innovation and Intellectual Property Rights. In *Translation Studies and Translation Practice: Proceedings of the 2nd International TRANSLATA Conference 2014*; Fagerberg, J., Mowery, D.C., Nelson, R.R., Eds.; Peter Lang D: Bern, Switzerland, 2017; pp. 141–147. [CrossRef]
7. Fleuren, W.W.; Alkema, W. Application of text mining in the biomedical domain. *Methods* **2015**, *74*, 97–106. [CrossRef]
8. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Deep Learning applications for COVID-19. *J. Big Data* **2021**, *8*. [CrossRef]
9. Gachloo, M.; Wang, Y.; Xia, J. A review of drug knowledge discovery using BioNLP and tensor or matrix decomposition. *Genom. Inform.* **2019**, *17*, e18. [CrossRef]
10. Zheng, S.; Dharssi, S.; Wu, M.; Li, J.; Lu, Z. Text Mining for Drug Discovery. In *Methods in Molecular Biology*; Springer: New York, NY, USA, 2019; pp. 231–252. [CrossRef]
11. Gonzalez, G.H.; Tahsin, T.; Goodale, B.C.; Greene, A.C.; Greene, C.S. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Briefings Bioinform.* **2015**, *17*, 33–42. [CrossRef]
12. Zhu, F.; Patumcharoenpol, P.; Zhang, C.; Yang, Y.; Chan, J.; Meechai, A.; Vongsangnak, W.; Shen, B. Biomedical text mining and its applications in cancer research. *J. Biomed. Inform.* **2013**, *46*, 200–211. [CrossRef]
13. Perera, N.; Dehmer, M.; Emmert-Streib, F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. *Front. Cell Dev. Biol.* **2020**, *8*, 673. [CrossRef] [PubMed]
14. Beheshti, S.M.R.; Venugopal, S.; Ryu, S.H.; Benatallah, B.; Wang, W. Big Data and Cross-Document Coreference Resolution: Current State and Future Opportunities. *arXiv* **2013**, arXiv:1311.3987.
15. Li, H.; Chen, Q.; Tang, B.; Wang, X.; Xu, H.; Wang, B.; Huang, D. CNN-based ranking for biomedical entity normalization. *BMC Bioinform.* **2017**, *18*. [CrossRef]
16. Cho, H.; Choi, W.; Lee, H. A method for named entity normalization in biomedical articles: Application to diseases and plants. *BMC Bioinform.* **2017**, *18*. [CrossRef]
17. Shirakawa, M.; Wang, H.; Song, Y.; Wang, Z.; Nakayama, K.; Hara, T. Entity Disambiguation based on a Probabilistic Taxonomy. Technical Report MSR-TR-2011-25; 2011. Available online: https://www.microsoft.com/en-us/research/publication/entity-disambiguation-based-on-a-probabilistic-taxonomy/ (accessed on 12 June 2021).
18. Gentile, A.L.; Zhang, Z.; Xia, L.; Iria, J. Semantic Relatedness Approach for Named Entity Disambiguation. In *Communications in Computer and Information Science*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 137–148. [CrossRef]
19. Zhu, G.; Iglesias, C.A. Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Syst. Appl.* **2018**, *101*, 8–24. [CrossRef]
20. Yadav, S.; Ramesh, S.; Saha, S.; Ekbal, A. Relation Extraction from Biomedical and Clinical Text: Unified Multitask Learning Framework. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**. [CrossRef]
21. Zhang, Y.; Lu, Z. Exploring semi-supervised variational autoencoders for biomedical relation extraction. *Methods* **2019**, *166*, 112–119. [CrossRef]
22. Muzaffar, A.W.; Azam, F.; Qamar, U. A Relation Extraction Framework for Biomedical Text Using Hybrid Feature Set. *Comput. Math. Methods Med.* **2015**, *2015*, 1–12. [CrossRef] [PubMed]
23. Xing, R.; Luo, J.; Song, T. BioRel: Towards large-scale biomedical relation extraction. *BMC Bioinform.* **2020**, *21*. [CrossRef] [PubMed]
24. Shah, P.; Perez-Iratxeta, C.; Bork, P.; Andrade, M. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinform.* **2003**, *4*, 20. [CrossRef] [PubMed]
25. Dai, H.; Wu, C.Y.; Tzong, R.; Tsai, R.T.H.; Hsu, W.L. From Entity Recognition to Entity Linking: A Survey of Advanced Entity Linking Techniques. In Proceedings of the 26th Annual Conference of the Japanese Society for Artificial Intelligence, Tokyo, Japan, 12–15 June 2012; pp. 1–10.
26. Collovini, S.; Bonamigo, T.; Vieira, R. A review on Relation Extraction with an eye on Portuguese. *J. Braz. Comput. Soc.* **2013**, *19*. [CrossRef]
27. Sun, W.; Cai, Z.; Li, Y.; Liu, F.; Fang, S.; Wang, G. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *J. Healthc. Eng.* **2018**, *2018*, 1–9. [CrossRef] [PubMed]
28. Ghamami, F.; Keyvanpour, M. Why biomedical relation extraction is an open issue? *ICIC Express Lett. Part B Appl.* **2018**. [CrossRef]
29. Saffer, J.D.; Burnett, V.L. Introduction to Biomedical Literature Text Mining: Context and Objectives. In *Methods in Molecular Biology*; Springer: New York, NY, USA, 2014; pp. 1–7. [CrossRef]
30. Nicholson, D.N.; Greene, C.S. Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1414–1428. [CrossRef]
31. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
32. Sachan, D.S.; Xie, P.; Xing, E.P. Effective Use of Bidirectional Language Modeling for Medical Named Entity Recognition. *arXiv* **2017**, arXiv:1711.07908.
33. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
34. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.

35. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**. [CrossRef]
36. Dustin Wright, Y.K. NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction. 2019. Available online: https://openreview.net/forum?id=BJerQWcp6Q (accessed on 12 June 2021). [CrossRef]
37. Ison, J.; Ménager, H.; Brancotte, B.; Jaaniso, E.; Salumets, A.; Raček, T.; Lamprecht, A.L.; Palmblad, M.; Kalaš, M.; Chmura, P.; others. Community curation of bioinformatics software and data resources. *Briefings Bioinform.* **2020**, *21*, 1697–1705. [CrossRef]
38. Sammartino, J.C.; Krallinger, M.; Valencia, A. Annotation Process, Guidelines and Text Corpus of Small Non-Coding RNA Molecules: The MiNCor for MicroRNA Annotations. In Proceedings of the Semantic Mining in Biomedicine (SMBM) 2016 CEUR Workshop Proceedings, Potsdam, Germany, 4–5 August 2016; pp. 56–63.
39. Lamurias, A.; Couto, F.M. Text mining for bioinformatics using biomedical literature. *Encycl. Bioinform. Comput. Biol.* **2019**, *1*, 602–611.
40. Campos, D.; Matos, S.; Oliveira, J.L. Biomedical named entity recognition: A survey of machine-learning tools. *Theory Appl. Adv. Text Min.* **2012**, *11*, 175–195.
41. Li, F.; Zhang, M.; Fu, G.; Ji, D. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinform.* **2017**, *18*, 1–11. [CrossRef]
42. Ananiadou, S.; Pyysalo, S.; Tsujii, J.; Kell, D.B. Event extraction for systems biology by text mining the literature. *Trends Biotechnol.* **2010**, *28*, 381–390. [CrossRef]
43. Thompson, P.; Iqbal, S.A.; McNaught, J.; Ananiadou, S. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinform.* **2009**, *10*, 1–19. [CrossRef]
44. Kim, J.D.; Ohta, T.; Tateisi, Y.; Tsujii, J. GENIA corpus—A semantically annotated corpus for bio-textmining. *Bioinformatics* **2003**, *19*, i180–i182. [CrossRef] [PubMed]
45. Bada, M.; Eckert, M.; Evans, D.; Garcia, K.; Shipley, K.; Sitnikov, D.; Baumgartner, W.A.; Cohen, K.B.; Verspoor, K.; Blake, J.A.; et al. Concept annotation in the CRAFT corpus. *BMC Bioinform.* **2012**, *13*, 1–20. [CrossRef] [PubMed]
46. Smith, L.; Tanabe, L.K.; nee Ando, R.J.; Kuo, C.J.; Chung, I.F.; Hsu, C.N.; Lin, Y.S.; Klinger, R.; Friedrich, C.M.; Ganchev, K.; et al. Overview of BioCreative II gene mention recognition. *Genome Biol.* **2008**, *9*, 1–19. [CrossRef] [PubMed]
47. Doğan, R.I.; Leaman, R.; Lu, Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **2014**, *47*, 1–10. [CrossRef]
48. Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D.M.; et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminformatics* **2015**, *7*, 1–17. [CrossRef]
49. Li, J.; Sun, Y.; Johnson, R.J.; Sciaky, D.; Wei, C.H.; Leaman, R.; Davis, A.P.; Mattingly, C.J.; Wiegers, T.C.; Lu, Z. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database* **2016**, *2016*. [CrossRef] [PubMed]
50. Lee, K.; Lee, S.; Park, S.; Kim, S.; Kim, S.; Choi, K.; Tan, A.C.; Kang, J. BRONCO: Biomedical entity Relation ONcology COrpus for extracting gene-variant-disease-drug relations. *Database* **2016**, *2016*. [CrossRef] [PubMed]
51. Neves, M.; Damaschun, A.; Kurtz, A.; Leser, U. Annotating and evaluating text for stem cell research. In Proceedings of the Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012) at Language Resources and Evaluation (LREC), Manchester, UK, 26 May 2012; pp. 16–23.
52. Krallinger, M.; Rabal, O.; Lourenço, A.; Perez, M.P.; Rodriguez, G.P.; Vazquez, M.; Leitner, F.; Oyarzabal, J.; Valencia, A. Overview of the CHEMDNER patents task. In Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, 2015. pp. 63–75. Available online: https://www.jdb.uzh.ch/id/eprint/37857 (accessed on 12 June 2021).
53. Lee, H.J.; Shim, S.H.; Song, M.R.; Lee, H.; Park, J.C. CoMAGC: A corpus with multi-faceted annotations of gene-cancer relations. *BMC Bioinform.* **2013**, *14*, 323. [CrossRef] [PubMed]
54. Cohen, K.B.; Verspoor, K.; Fort, K.; Funk, C.; Bada, M.; Palmer, M.; Hunter, L.E. The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain. In *Handbook of Linguistic Annotation*; Springer: Dordrecht, The Netherlands, 2017; pp. 1379–1394. [CrossRef]
55. Herrero-Zazo, M.; Segura-Bedmar, I.; Martínez, P.; Declerck, T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *J. Biomed. Inform.* **2013**, *46*, 914–920. [CrossRef] [PubMed]
56. Gerner, M.; Nenadic, G.; Bergman, C.M. An Exploration of Mining Gene Expression Mentions and Their Anatomical Locations from Biomedical Text. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*; Association for Computational Linguistics: Uppsala, Sweden, 2010; pp. 72–80.
57. Oh, S.Y.; Kim, J.H.; Kim, S.J.; Nam, H.J.; Park, H.S. GNI Corpus version 1.0: Annotated full-text corpus of Genomics & Informatics to support biomedical information extraction. *Genom. Inform.* **2018**, *16*, 75.
58. Smith, L.H.; Tanabe, L.; Rindflesch, T.C.; Wilbur, W.J. MedTag: A collection of biomedical annotations. In Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, Stroudsburg, PA, USA, 24 June 2005; pp. 32–37.
59. Pyysalo, S.; Ohta, T.; Miwa, M.; Cho, H.C.; Tsujii, J.; Ananiadou, S. Event extraction across multiple levels of biological organization. *Bioinformatics* **2012**, *28*, i575–i581. [CrossRef] [PubMed]
60. Shardlow, M.; Nguyen, N.; Owen, G.; O'Donovan, C.; Leach, A.; McNaught, J.; Turner, S.; Ananiadou, S. A new corpus to support text mining for the curation of metabolites in the Chebi database. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 280–285.

61. Islamaj, R.; Leaman, R.; Kim, S.; Kwon, D.; Wei, C.H.; Comeau, D.C.; Peng, Y.; Cissel, D.; Coss, C.; Fisher, C.; et al. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci. Data* **2021**, *8*, 1–12. [CrossRef]
62. Islamaj, R.; Wei, C.H.; Cissel, D.; Miliaras, N.; Printseva, O.; Rodionov, O.; Sekiya, K.; Ward, J.; Lu, Z. NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *J. Biomed. Informatics* **2021**, *118*, 103779. [CrossRef]
63. Sousa, D.; Lamúrias, A.; Couto, F.M. A silver standard corpus of human phenotype-gene relations. *arXiv* **2019**, arXiv:1903.10728.
64. Verspoor, K.; Jimeno Yepes, A.; Cavedon, L.; McIntosh, T.; Herten-Crabb, A.; Thomas, Z.; Plazzer, J.P. Annotating the biomedical literature for the human variome. *Database* **2013**, *2013*. [CrossRef]
65. Cunningham, H.; Tablan, V.; Roberts, A.; Bontcheva, K. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Comput. Biol.* **2013**, *9*, e1002854. [CrossRef]
66. Johansson, M.; Roberts, A.; Chen, D.; Li, Y.; Delahaye-Sourdeix, M.; Aswani, N.; Greenwood, M.A.; Benhamou, S.; Lagiou, P.; Holcátová, I.; et al. Using Prior Information from the Medical Literature in GWAS of Oral Cancer Identifies Novel Susceptibility Variant on Chromosome 4—The AdAPT Method. *PLoS ONE* **2012**, *7*, e36888. [CrossRef] [PubMed]
67. Ferrucci, D.; Lally, A. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.* **2004**, *10*, 327–348. [CrossRef]
68. Ogren, P.V.; Wetzler, P.G.; Bethard, S. ClearTK: A UIMA toolkit for statistical natural language processing. In Proceedings of the Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP Workshop at Language Resources and Evaluation Conference (LREC), Marrakech, Morocco, 31 May 2008; Volume 32, pp. 32–38.
69. Bethard, S.; Ogren, P.; Becker, L. ClearTK 2.0: Design patterns for machine learning in UIMA. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*; European Language Resources Association (ELRA): Reykjavik, Iceland, 2014; Volume 2014, pp. 3289–3293.
70. Wang, Y.; Mehrabi, S.; Sohn, S.; Atkinson, E.J.; Amin, S.; Liu, H. Natural language processing of radiology reports for identification of skeletal site-specific fractures. *BMC Med. Inform. Decis. Mak.* **2019**, *19*. [CrossRef]
71. Roeder, C.; Jonquet, C.; Shah, N.H.; Baumgartner, W.A.; Verspoor, K.; Hunter, L. A UIMA wrapper for the NCBO annotator. *Bioinformatics* **2010**, *26*, 1800–1801. [CrossRef]
72. Comeau, D.C.; Dogan, R.I.; Ciccarese, P.; Cohen, K.B.; Krallinger, M.; Leitner, F.; Lu, Z.; Peng, Y.; Rinaldi, F.; Torii, M.; et al. BioC: A minimalist approach to interoperability for biomedical text processing. *Database* **2013**, *2013*, bat064. [CrossRef] [PubMed]
73. Leaman, R.; Islamaj Doğan, R.; Lu, Z. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics* **2013**, *29*, 2909–2917. [CrossRef]
74. Wei, C.H.; Harris, B.R.; Kao, H.Y.; Lu, Z. tmVar: A text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics* **2013**, *29*, 1433–1439. [CrossRef]
75. Wei, C.H.; Kao, H.Y.; Lu, Z. SR4GN: A species recognition software tool for gene normalization. *PLoS ONE* **2012**, *7*, e38460. [CrossRef]
76. Leaman, R.; Wei, C.H.; Lu, Z. tmChem: A high performance approach for chemical named entity recognition and normalization. *J. Cheminformatics* **2015**, *7*, 1–10. [CrossRef]
77. Wei, C.H.; Kao, H.Y. Cross-species gene normalization by species inference. *BMC Bioinform.* **2011**, *12*, 1–11. [CrossRef] [PubMed]
78. Wei, C.H.; Kao, H.Y.; Lu, Z. PubTator: A web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* **2013**, *41*, W518–W522. [CrossRef] [PubMed]
79. Khare, R.; Wei, C.H.; Mao, Y.; Leaman, R.; Lu, Z. tmBioC: Improving interoperability of text-mining tools with BioC. *Database* **2014**, *2014*. [CrossRef] [PubMed]
80. Rinaldi, F.; Clematide, S.; Marques, H.; Ellendorff, T.; Romacker, M.; Rodriguez-Esteban, R. OntoGene web services for biomedical text mining. *BMC Bioinform.* **2014**, *15*. [CrossRef]
81. Torii, M.; Li, G.; Li, Z.; Oughtred, R.; Diella, F.; Celen, I.; Arighi, C.N.; Huang, H.; Vijay-Shanker, K.; Wu, C.H. RLIMS-P: An online text-mining tool for literature-based extraction of protein phosphorylation information. *Database* **2014**, *2014*, bau081. [CrossRef] [PubMed]
82. Casteleiro, M.A.; Demetriou, G.; Read, W.; Prieto, M.J.F.; Maroto, N.; Fernandez, D.M.; Nenadic, G.; Klein, J.; Keane, J.; Stevens, R. Deep learning meets ontologies: Experiments to anchor the cardiovascular disease ontology in the biomedical literature. *J. Biomed. Semant.* **2018**, *9*. [CrossRef]
83. Doğan, R.I.; Kim, S.; Chatr-aryamontri, A.; Chang, C.S.; Oughtred, R.; Rust, J.; Wilbur, W.J.; Comeau, D.C.; Dolinski, K.; Tyers, M. The BioC-BioGRID corpus: Full text articles annotated for curation of protein–protein and genetic interactions. *Database* **2017**, *2017*, baw147. [CrossRef] [PubMed]
84. Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*; Association for Computational Linguistics: Baltimore, MD, USA, 2014; pp. 55–60. [CrossRef]
85. Lu, H.; Kai, Z. How Do General-Purpose Sentiment Analyzers Perform when Applied to Health-Related Online Social Media Data? *Stud. Health Technol. Inform.* **2019**, *264*, 1208–1212. [CrossRef]
86. Weber, L.; Münchmeyer, J.; Rocktäschel, T.; Habibi, M.; Leser, U. HUNER: Improving biomedical NER with pretraining. *Bioinformatics* **2019**, *36*, 295–302. [CrossRef]

87. Weber, L.; Sänger, M.; Münchmeyer, J.; Habibi, M.; Leser, U.; Akbik, A. HunFlair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics* **2021**. [CrossRef]
88. Cabot, C.; Darmoni, S.; Soualmia, L.F. Cimind: A phonetic-based tool for multilingual named entity recognition in biomedical texts. *J. Biomed. Inform.* **2019**, *94*, 103176. [CrossRef] [PubMed]
89. Thomas, P.; Rocktäschel, T.; Hakenberg, J.; Lichtblau, Y.; Leser, U. SETH detects and normalizes genetic variants in text. *Bioinformatics* **2016**, *32*, 2883–2885. [CrossRef]
90. Lee, H.C.; Hsu, Y.Y.; Kao, H.Y. AuDis: An automatic CRF-enhanced disease normalization in biomedical text. *Database* **2016**, *2016*, baw091. [CrossRef]
91. Gupta, S.; Dingerdissen, H.; Ross, K.E.; Hu, Y.; Wu, C.H.; Mazumder, R.; Vijay-Shanker, K. DEXTER: Disease-Expression Relation Extraction from Text. *Database* **2018**, *2018*. [CrossRef]
92. Dingerdissen, H.M.; Torcivia-Rodriguez, J.; Hu, Y.; Chang, T.C.; Mazumder, R.; Kahsay, R. BioMuta and BioXpress: Mutation and expression knowledgebases for cancer biomarker discovery. *Nucleic Acids Res.* **2017**, *46*, D1128–D1136. [CrossRef]
93. Weber, L.; Thobe, K.; Lozano, O.A.M.; Wolf, J.; Leser, U. PEDL: Extracting protein–protein associations using deep language models and distant supervision. *Bioinformatics* **2020**, *36*, i490–i498. [CrossRef]
94. Kim, D.; Lee, J.; So, C.H.; Jeon, H.; Jeong, M.; Choi, Y.; Yoon, W.; Sung, M.; Kang, J. A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. *IEEE Access* **2019**, *7*, 73729–73740. [CrossRef]
95. Malarkodi, C.; Pattabhi, R.; Sobha, L.D. CLRG ChemNER: A Chemical Named Entity Recognizer@ ChEMU CLEF 2020. 2020. Available online: moz-extension://c64046de-9d28-4e46-a199-807c4d6ae096/pdf-viewer/web/viewer.html?file=http%3A%2F%2Fceur-ws.org%2FVol-2696%2Fpaper_236.pdf (accessed on 12 June 2021).
96. Yoon, W.; So, C.H.; Lee, J.; Kang, J. CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinform.* **2019**, *20*. [CrossRef]
97. Dang, T.H.; Le, H.Q.; Nguyen, T.M.; Vu, S.T. D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. *Bioinformatics* **2018**, *34*, 3539–3546. [CrossRef]
98. Wei, C.H.; Kao, H.Y.; Lu, Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed Res. Int.* **2015**, *2015*, 1–7. [CrossRef]
99. Giorgi, J.M.; Bader, G.D. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics* **2018**, *34*, 4087–4094. [CrossRef] [PubMed]
100. Chauhan, G.; McDermott, M.; Szolovits, P. Reflex: Flexible framework for relation extraction in multiple domains. *arXiv* **2019**, arXiv:1906.08318.
101. Giorgi, J.M.; Bader, G.D. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics* **2019**, *36*, 280–286. [CrossRef] [PubMed]
102. Neumann, M.; King, D.; Beltagy, I.; Ammar, W. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv* **2019**, arXiv:1902.07669.
103. Dao, M.H.; Nguyen, D.Q. VinAI at ChEMU 2020: An Accurate System for Named Entity Recognition in Chemical Reactions from Patents. 2020. Available online: https://www.vinai.io/publication-posts/vinai-at-chemu-2020-an-accurate-system-for-named-entity-recognition-in-chemical-reactions-from-patents (accessed on 12 June 2021).
104. Zuo, M.; Zhang, Y. Dataset-aware multi-task learning approaches for biomedical named entity recognition. *Bioinformatics* **2020**, *36*, 4331–4338. [CrossRef]
105. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, i37–i48. [CrossRef]
106. Wei, C.H.; Allot, A.; Leaman, R.; Lu, Z. PubTator central: Automated concept annotation for biomedical full text articles. *Nucleic Acids Res.* **2019**, *47*, W587–W593. [CrossRef]
107. Djekidel, M.N.; Rosikiewicz, W.; Peng, J.C.; Kanneganti, T.D.; Hui, Y.; Jin, H.; Hedges, D.; Schreiner, P.; Fan, Y.; Wu, G.; et al. CovidExpress: An Interactive Portal for Intuitive Investigation on SARS-CoV-2 Related Transcriptomes. 2021. Available online: https://www.biorxiv.org/content/10.1101/2021.05.14.444026v1 (accessed on 12 June 2021). [CrossRef]
108. Wu, M.; Zhang, Y.; Grosser, M.; Tipper, S.; Venter, D.; Lin, H.; Lu, J. Profiling COVID-19 Genetic Research: A Data-Driven Study Utilizing Intelligent Bibliometrics. *Front. Res. Metrics Anal.* **2021**, *6*. [CrossRef]
109. Desterke, C.; Turhan, A.G.; Bennaceur-Griscelli, A.; Griscelli, F. HLA-dependent heterogeneity and macrophage immunoproteasome activation during lung COVID-19 disease. *J. Transl. Med.* **2021**, *19*. [CrossRef]
110. Venkatesan, A.; Kim, J.H.; Talo, F.; Ide-Smith, M.; Gobeill, J.; Carter, J.; Batista-Navarro, R.; Ananiadou, S.; Ruch, P.; McEntyre, J. SciLite: A platform for displaying text-mined annotations as a means to link research articles with biological data. *Wellcome Open Res.* **2016**, *1*, 25. [CrossRef]
111. Palopoli, N.; Iserte, J.A.; Chemes, L.B.; Marino-Buslje, C.; Parisi, G.; Gibson, T.J.; Davey, N.E. The articles.ELM resource: Simplifying access to protein linear motif literature by annotation, text-mining and classification. *Database* **2020**, *2020*. [CrossRef]
112. Firth, R.; Talo, F.; Venkatesan, A.; Mukhopadhyay, A.; McEntyre, J.; Velankar, S.; Morris, C. Automatic annotation of protein residues in published papers. *Acta Crystallogr. Sect. Struct. Biol. Commun.* **2019**, *75*, 665–672. [CrossRef]
113. Müller, H.M.; Kenny, E.E.; Sternberg, P.W. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PLoS Biol.* **2004**, *2*, e309. [CrossRef]

114. Hu, Y.; Chung, V.; Comjean, A.; Rodiger, J.; Nipun, F.; Perrimon, N.; Mohr, S.E. BioLitMine: Advanced Mining of Biomedical and Biological Literature About Human Genes and Genes from Major Model Organisms. *G3 Genes Genomes Genetics* **2020**, *10*, 4531–4539. [CrossRef]
115. Campos, D.; Lourenço, J.; Matos, S.; Oliveira, J.L. Egas: A collaborative and interactive document curation platform. *Database* **2014**, *2014*, bau048. [CrossRef]
116. Nunes, T.; Campos, D.; Matos, S.; Oliveira, J.L. BeCAS: Biomedical concept recognition services and visualization. *Bioinformatics* **2013**, *29*, 1915–1916. [CrossRef]
117. Liu, H.; Hu, Z.Z.; Zhang, J.; Wu, C. BioThesaurus: A web-based thesaurus of protein and gene names. *Bioinformatics* **2005**, *22*, 103–105. [CrossRef]
118. Sernadela, P.; González-Castro, L.; Carta, C.; van der Horst, E.; Lopes, P.; Kaliyaperumal, R.; Thompson, M.; Thompson, R.; Queralt-Rosinach, N.; Lopez, E.; et al. Linked Registries: Connecting Rare Diseases Patient Registries through a Semantic Web Layer. *BioMed Res. Int.* **2017**, *2017*, 1–13. [CrossRef] [PubMed]
119. Liu, Y.; Liang, Y.; Wishart, D. PolySearch2: A significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Res.* **2015**, *43*, W535–W542. [CrossRef]
120. Khan, F.; Radovanovic, A.; Gojobori, T.; Kaur, M. IBDDB: A manually curated and text-mining-enhanced database of genes involved in inflammatory bowel disease. *Database* **2021**, *2021*. [CrossRef]
121. Liu, B.; Bai, C. Regulatory Mechanisms of Coicis Semen on Bionetwork of Liver Cancer Based on Network Pharmacology. *BioMed Res. Int.* **2020**, *2020*, 1–17. [CrossRef]
122. Tsuruoka, Y.; Tsujii, J.; Ananiadou, S. FACTA: A text search engine for finding associated biomedical concepts. *Bioinformatics* **2008**, *24*, 2559–2560. [CrossRef] [PubMed]
123. Tsuruoka, Y.; Miwa, M.; Hamamoto, K.; Tsujii, J.; Ananiadou, S. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* **2011**, *27*, i111–i119. [CrossRef]
124. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119. [CrossRef]
125. Humphreys, B.L.; Lindberg, D.A.B.; Schoolman, H.M.; Barnett, G.O. The Unified Medical Language System: An Informatics Research Collaboration. *J. Am. Med. Inform. Assoc.* **1998**, *5*, 1–11. [CrossRef] [PubMed]
126. Wishart, D.S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A.C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. HMDB: The Human Metabolome Database. *Nucleic Acids Res.* **2007**, *35*, D521–D526. [CrossRef]
127. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef]
128. Wishart, D.S.; Knox, C.; Guo, A.C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2007**, *36*, D901–D906. [CrossRef]
129. Le, N.; Ho, T.; Ho, B.; Tran, D. A nucleosomal approach to inferring causal relationships of histone modifications. *BMC Genom.* **2014**, *15*, S7. [CrossRef]
130. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2018**, *47*, D607–D613. [CrossRef]
131. Szklarczyk, D.; Santos, A.; von Mering, C.; Jensen, L.J.; Bork, P.; Kuhn, M. STITCH 5: Augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **2015**, *44*, D380–D384. [CrossRef]
132. Mendez, D.; Gaulton, A.; Bento, A.P.; Chambers, J.; Veij, M.D.; Félix, E.; Magariños, M.P.; Mosquera, J.F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **2018**, *47*, D930–D940. [CrossRef]
133. Roth, B.L.; Lopez, E.; Patel, S.; Kroeze, W.K. The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches? *Neuroscientist* **2000**, *6*, 252–262. [CrossRef]
134. Burley, S.K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G.V.; Christie, C.H.; Dalenberg, K.; Costanzo, L.D.; Duarte, J.M.; et al. RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* **2020**, *49*, D437–D451. [CrossRef]
135. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2017**, *46*, D1074–D1082. [CrossRef]
136. Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C. GLIDA: GPCR ligand database for chemical genomics drug discovery database and tools update. *Nucleic Acids Res.* **2007**, *36*, D907–D912. [CrossRef]
137. Gunther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E.G.; Gewiess, A.; Jensen, L.J.; et al. SuperTarget and Matador: Resources for exploring drug-target relationships. *Nucleic Acids Res.* **2007**, *36*, D919–D922. [CrossRef]
138. Wang, Y.; Zhang, S.; Li, F.; Zhou, Y.; Zhang, Y.; Wang, Z.; Zhang, R.; Zhu, J.; Ren, Y.; Tan, Y.; et al. Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* **2019**. [CrossRef]
139. Davis, A.P.; Wiegers, T.C.; Wiegers, J.; Grondin, C.J.; Johnson, R.J.; Sciaky, D.; Mattingly, C.J. CTD anatomy: Analyzing chemical-induced phenotypes and exposures from an anatomical perspective, with implications for environmental health studies. *Curr. Res. Toxicol.* **2021**, *2*, 128–139. [CrossRef]

140. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **2020**, *49*, D545–D551. [CrossRef]
141. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2019**. [CrossRef] [PubMed]
142. Karp, P.D.; Billington, R.; Caspi, R.; Fulcher, C.A.; Latendresse, M.; Kothari, A.; Keseler, I.M.; Krummenacker, M.; Midford, P.E.; Ong, Q.; et al. The BioCyc collection of microbial genomes and metabolic pathways. *Briefings Bioinform.* **2017**, *20*, 1085–1093. [CrossRef]
143. Huang, H.Y.; Lin, Y.C.D.; Li, J.; Huang, K.Y.; Shrestha, S.; Hong, H.C.; Tang, Y.; Chen, Y.G.; Jin, C.N.; Yu, Y.; et al. miRTarBase 2020: Updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.* **2019**. [CrossRef]
144. Oughtred, R.; Stark, C.; Breitkreutz, B.J.; Rust, J.; Boucher, L.; Chang, C.; Kolas, N.; O'Donnell, L.; Leung, G.; McAdam, R.; et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **2018**, *47*, D529–D541. [CrossRef] [PubMed]
145. Piñero, J.; Ramírez-Anguita, J.M.; Saüch-Pitarch, J.; Ronzano, F.; Centeno, E.; Sanz, F.; Furlong, L.I. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **2020**, *48*, D845–D855. [CrossRef] [PubMed]
146. Pérez-Rodríguez, G.; Pérez-Pérez, M.; Fdez-Riverola, F.; Lourenço, A. Online visibility of software-related web sites: The case of biomedical text mining tools. *Inf. Process. Manag.* **2019**, *56*, 565–583. [CrossRef]