SURVEY OF NLP IN PHARMACOLOGY: METHODOLOGY, TASKS, RESOURCES, KNOWLEDGE, AND TOOLS

A PREPRINT

Dimitar Trajanov[1], **Vangel Trajkovski**[1], **Makedonka Dimitrieva**[1], **Jovana Dobreva**[1], **Milos Jovanovik**[1], **Matej Klemen**[2], **Aleš Žagar**[2], **Marko Robnik-Šikonja**[2]
[1] Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, North Macedonia
{dimitar.trajanov,jovana.dobreva,milos.jovanovik}@finki.ukim.mk
{vangel.trajkovski,makedonka.dimitrieva}@students.finki.ukim.mk
[2] University of Ljubljana, Faculty of Computer and Information Science, Slovenia
{matej.klemen,ales.zagar,marko.robnik}@fri.uni-lj.si

August 23, 2022

ABSTRACT

Natural language processing (NLP) is an area of artificial intelligence that applies information technologies to process the human language, understand it to a certain degree, and use it in various applications. This area has rapidly developed in the last few years and now employs modern variants of deep neural networks to extract relevant patterns from large text corpora. The main objective of this work is to survey the recent use of NLP in the field of pharmacology. As our work shows, NLP is a highly relevant information extraction and processing approach for pharmacology. It has been used extensively, from intelligent searches through thousands of medical documents to finding traces of adversarial drug interactions in social media. We split our coverage into five categories to survey modern NLP methodology, commonly addressed tasks, relevant textual data, knowledge bases, and useful programming libraries. We split each of the five categories into appropriate subcategories, describe their main properties and ideas, and summarize them in a tabular form. The resulting survey presents a comprehensive overview of the area, useful to practitioners and interested observers.

1 Introduction

Information processing is indispensable to modern drug design, production, and application. A significant amount of information is stored in the textual form and located in scientific papers, clinical notes, ontologies, knowledge bases, social media posts, and newspaper articles. Extraction and retrieval of this information rely on natural language processing (NLP). NLP is a broad scientific area based on computer science, linguistics, and artificial intelligence [99; 100]. As the whole area of artificial intelligence, it has been completely transformed in recent years by deep learning [63]. It has witnessed numerous new techniques and successful applications, such as intelligent search, machine translation, and speech recognition.

Many general NLP techniques and approaches can be applied to the pharmacological area. However, often NLP techniques have to be adapted to the specifics of the field in terms of available knowledge sources, text representation, specific methods, terminology, etc. In this work, we survey modern NLP methodology, tasks, resources, knowledge bases, and tools used and adapted to the area of pharmacology. The review aims to inform practitioners working in the area of the pharmacology of exciting recent development and to give a solid starting reference material to new entrants.

Several surveys summarise NLP in pharmacology but only cover specific areas of NLP methods. One of the first reviews of NLP for clinical decision support (CDS) [46] was published in 2009. The authors observed that many CDS data is textual and reviewed existing NLP developments for CDS. Luo et al. [132] present a structured review of NLP for narratives in electronic health records (EHR) for pharmacovigilance. Dreisbach et al. [52] review NLP

of symptoms from electronic patient-authored text data. A review of NLP in languages other than English for clinic-related texts is presented by Névéol et al. [146]. Chen et al. [39] survey NLP addressing challenges related to COVID-19 pandemic. They present details related to several NLP tasks like information retrieval, named entity recognition, literature-based discovery, question answering, topic modeling, sentiment and emotion analysis, caseload forecasting, and misinformation detection. In contrast to the listed surveys, we aim for a comprehensive overview of NLP in pharmacology.

Recently, the primary methodological approach to NLP has been deep learning. Deep neural networks (DNNs) require that text is transformed (embedded) into numeric vectors in a process called representation learning. We present general text embeddings as well as specific variants relevant to the area of life sciences and pharmacology. As pharmacology is a knowledge-intensive area where relevant information is not stored only in text documents but also in databases, ontologies, and linked data, we survey recent attempts to inject knowledge into DNNs. Due to the need to understand the decisions and biases of DNNs, we discuss techniques that make their output more transparent.

Some NLP tasks are particularly important to the area of pharmacology. While they are often based on general approaches, they are strongly adapted and use specific pharmacological resources. We discuss general tasks such as named entity recognition, relation extraction, literature-based discovery, question answering, and field-specific tasks such as detection of adverse drug reactions and drug discovery.

The basic precondition for applying NLP is the availability of language resources. In multiple studies, EHRs are the main source of information [219; 125; 208; 85; 118]. EHRs contain patient data such as diagnoses, hospital admissions, prescriptions, and adversary drug effects. The data in EHRs are well-structured and can be readily processed; however, different EHR components are difficult to integrate. Many authors use molecular data [197; 153], which can be integrated with diseases [62]. Other important sources of information are clinical data [98] (used in, e.g., drug repurposing [224; 45]), linked data, and the pharmacology-related semantic web.

Linked data and knowledge graphs have recently emerged as general formalisms to represent knowledge in artificial intelligence and the semantic web. Linked (open) data movement introduced new standards for representing, storing, and retrieving data over the web [19; 20; 72; 218; 77], which enabled new distributed data sources and new applications. Knowledge graphs allow generating, consolidating, and contextually linking structured data. We present several knowledge graphs from the biomedical domain and outline several COVID-19-related knowledge graphs.

Software tools and libraries are essential for using NLP in pharmacological research and practice. Mostly, these support the Python language. We present many general NLP tools and libraries as well as life-science and pharmacology-specific variants.

We organize the survey along with five main areas: methodology, common tasks, datasets, knowledge graphs, and software libraries. In Section 2, we structure NLP methodologies into three groups: representation learning (i.e., different embeddings), approaches to inject domain-specific knowledge into deep neural networks, and explainable AI techniques used in pharmacology. The most frequently used NLP tasks in pharmacology are presented in Section 3. We cover the named entity recognition, relation extraction, adverse drug reactions, literature-based discovery, and question answering. In Section 4, we first outline the approaches to finding data resources, followed by a survey of existing data. We organize the overview into five categories: patient data, drug usage data, drug structure data, question answering datasets, and general text processing datasets. Knowledge graphs used in the biomedical domain and a specific example of COVID-19 disease knowledge graphs are covered in Section 5. We give an overview of useful NLP software libraries and tools for the pharmacological domain as well as useful general NLP libraries in Section 6. We conclude the survey in Section 7.

2 NLP Methodology in Pharmacology

Recently, NLP has switched entirely to deep neural networks, mostly large language models (LLMs) that are pretrained on huge quantities of text to capture various linguistic, general, and domain-specific knowledge. LLMs embed the text data into a numeric representation preserving semantic relations between words. To be used for specific tasks, LLMs are fine-tuned with problem-specific data.

In Section 2.1, we give an overview of modern text representations. We present static and contextual embeddings (i.e., LLMs) and specific variants relevant to the area of life sciences and pharmacology. While most of the work is focused on English, we present some notable exceptions in other languages. As pharmacology is a knowledge-intensive area where relevant information is not stored only in text documents but also in databases, ontologies, and linked-data, we survey recent attempts to inject knowledge into deep neural networks in Section 2.2. Unfortunately, deep neural networks often appear as black-box models, lacking transparency on how the decisions are taken. In Section 2.3, we

present general explanation techniques applicable to text prediction and focus on successful applications related to pharmacology.

2.1 Representation Learning

In NLP, text representation is a crucial issue and research direction. Various text embeddings emerged that capture both syntax and semantics of a given text. While traditional approaches were based on sparse representations such as bag-of-words, dense representations such as word2vec [144], ELMo [157], and BERT [48] are based on neural networks and offer much more semantically valid and computationally efficient representations. A common trait of these embeddings is to train a neural network on self-supervised text classification tasks and use the weights of the trained neural network or the whole trained network to represent different text units (words, sentences, or documents). The labels required for training these classifiers originate from large corpora of general texts, e.g., web crawl, news, and Wikipedia. The usual classification tasks used in training these representation models are predicting the next and previous word in a sequence or filling in missing words (also called masked language modeling). Representation learning can be extended with other related tasks, such as prediction if two sentences are sequential. The positive instances for learning are obtained from the text in the given corpus, while the negative instances are mostly sampled from instances that are unlikely to be related.

We first briefly describe the principle of the most frequently used static embeddings, called word2vec, followed by large language models such as contextual BERT. Next, we cover the adaptations of these representation techniques for life sciences and pharmacology domains. We provide a summary of the presented embeddings at the end of the section in Table 1.

2.1.1 Static Embeddings

The word2vec word embedding method [144] trains a shallow (one hidden layer) neural network predicting the neighboring words of a given input word. The trained weights of the hidden layer produce a static embedding in the sense that we get a single vector for each word. For example, the term *bank* may denote a financial institution or land alongside a river, but it is represented with a single vector.

The word2vec method pre-trains a feed-forward neural network on a huge corpus, and the weights of the hidden layer in this network are used as word embeddings. Pretrained word vectors for many languages are publicly available. The published vectors are typically 100 or 300-dimensional, e.g., Google published vectors for 3 million English words and phrases[1]. While the word2vec algorithm consists of two related methods, we describe only the skip-gram method, which mostly produces more favorable results. The method constructs a neural network to classify cooccurring words by taking a word and predicting its d preceding and succeeding words, e.g., $\pm$ 5 words. In the actual neural network, one word is on the input (the central word) and one word is on the output, where both are represented with one-hot encoding. The words and their contexts appearing in the training corpus constitute the training instances of the classification problem. The first word of the training pair is presented at the network's input in the one-hot-encoding representation, and the network is trained to predict the second word. The difference in prediction is evaluated using a loss function. For a sequence of T training words $w_1, w_2, w_3, \ldots, w_T$, the skip-gram model maximizes the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-d\leq j\leq d, j\neq 0}\log p(w_{t+j}|w_t).$$

Once the network is trained with word2vec, vectors for each word in the vocabulary can be generated. As one-hot encoding of the input word only activates one input connection for each hidden layer neuron, the weights on these connections constitute the embedding vector for the given input word.

The resulting word embeddings' properties depend on the context's size. For a small number of neighboring words (e.g., $\pm$ 5 words), we get embeddings that perform better on syntactic tasks. For larger neighborhoods (e.g., $\pm$ 10 words), the embeddings better express semantic properties.

Word2vec has attracted the immense attention of NLP researchers and practitioners. The word2vec precomputed embeddings soon became a default choice for the first layer of many classification deep neural networks. Several domain-specific variants have also been created and made publicly available. For life sciences, a well-known example is the work of Pyysalo et al. [166], who released two sets of word2vec vectors. The first, denoted PubMed-PMC, was trained on 23M PubMed abstracts and 0.7M PubMed Central (PMC) articles. The second model, Wiki-PubMed-PMC, was prepared using the same two corpora combined with 4M English Wikipedia articles. These static embeddings were

[1] `https://code.google.com/archive/p/word2vec/`

successfully used in many life-science applications (the paper received 492 citations by 13 March 2022). For example, Habibi et al. [69] have successfully applied the two embeddings to the biomedical named entity recognition problem to detect genes, chemicals and diseases.

Note that the same technology to represent text can be applied to represent biological sequences, such as DNA, RNA, and proteins [11]. The created bio-vectors (BioVec) refer to biological sequences in general, protein-vectors are called ProtVec, and gene vectors are named GeneVec. A similar attempt to represent biological sequences is dna2vec vectors [147].

Despite the successful use of static embeddings such as word2vec, contextual embedding models such as BERT have become even more successful. Therefore, we skip the detailed review of static embedding models and focus on contextual models.

2.1.2 Contextual Word Embeddings

The problem with word2vec embeddings is their failure to express polysemous words. During its training, all senses of a given word (e.g., *paper* as a material, as a newspaper, as a scientific work, and as an exam) contribute relevant neighboring words in proportion to their frequency in the training corpus. This causes the final vector to be placed somewhere in the weighted middle of all words' meanings. Consequently, rare meanings of words are poorly expressed with word2vec, and the resulting vectors do not offer good semantic representations. For example, none of the 50 closest vectors of the word *paper* is related to science.

The idea of contextual word embeddings is to generate a different vector for each word's context. The context is typically defined sentence-wise. This solves the problems with word polysemy. The context of a sentence is mostly enough to disambiguate different meanings of a word for humans and learning algorithms. Several contextual embeddings have been developed, e.g., ELMo, ULMFit, and BERT. As the latter achieves the best results in most NLP tasks, we describe it below.

Contextual embeddings are based on the idea of language models, which predict either the next, previous or missing word in a sequence. Training often combines several of these and other related tasks. Due to the network's depth, extracting vector representations from the network is no longer trivial, i.e., the trained deep networks store their knowledge in weights spread over several layers. A frequently used approach concatenates weights from several layers into a vector. Still, often it is more convenient to use the whole pretrained neural language model as a starting point and fine-tune its weights further during the training on a specific task.

BERT (Bidirectional Encoder Representations from Transformers) embeddings [48] generalize the idea of language models (LMs) to masked language models, inspired by the gap-filling tests. The masked language model randomly masks some of the tokens from the input. The task of an LM is then to predict each missing token based on its neighborhood. BERT uses the transformer architecture of neural networks [203] in a bidirectional sense (forward and backward). It introduces another task of predicting whether two sentences appear in a sequence. The input representation of BERT is sequences of tokens representing sub-word units. The input is constructed by summing the corresponding token, segment, and position embeddings.

Using BERT for classification requires adding connections between its last hidden layer and new neurons corresponding to the number of classes in the intended task. The fine-tuning process is typically applied to the whole network. All the BERT parameters and new class-specific weights are fine-tuned jointly to maximize the log-probability of the correct labels.

BERT has shown excellent performance on many NLP tasks and is now a de-facto standard in NLP. In the initial evaluation [48], BERT showed improved performance on all eight tasks from the GLUE (general language understanding evaluation) benchmark suite [205], consisting of question answering, named entity recognition, and common-sense inference. A variant of BERT, called RoBERTa [128], which only uses masked language model training but on a larger dataset and for a longer time, has become a popular practical choice due to its improved robustness and better parallel training capability.

Due to its success, BERT has spurred an immense tide of research, analyzing its capabilities and using and adapting it for different purposes. An overview of research on BERT capabilities and inner workings is presented by Rogers et al. [178]. Below, we overview the adaptations and applications relevant to pharmacology.

2.1.3 BERT Variants Relevant to Pharmacology

BERT has many extensions in architecture, training, and fine-tuning. A general improvement for science-related text processing is **SciBERT** [15] that was trained on 1.14M scientific papers (3.17B tokens) from Semantic Scholar instead

of general text. The training data consisted of 18% computer science papers and 82% papers from the biomedical domain. Upon its introduction, the SciBERT was compared to BERT and achieved improved performance in a study involving four classification tasks based on scientific publications: named entity recognition (NER), extraction of participants, interventions, comparisons, and outcomes in clinical trial papers, text classification, relation classification, and dependency parsing (DP). The SciBERT has attracted considerable attention of the scientific community with more than 1000 citations recorded by Google Scholar at the time of this writing.

In life sciences, there are several popular domain adaptations of BERT. **BioMed-RoBERTa**-base [68] (almost 600 Google Scholar citations at the time of this writing) is an adaptation of RoBERTa [128], using long pretraining on 160GB of standard texts and additional 47GB (7.55B tokens from 2.68M papers) of abstracts and full papers randomly sampled from PubMed repository. Using this domain-adapted pretrained model, the authors improved classification for two domain-specific tasks. First, they improved the classification compared to the baseline RoBERTa model for 2.3 micro F_1 percent on the Chem-Prot database [108] that contains chemical-protein-disease annotations enabling the study of systems pharmacology for a small molecule across multiple layers of complexity from molecular to clinical levels. Second, they tested the BioMed-RoBERTa on the PubMed sequential sentence classification task [47] and achieved 0.4 micro F_1 percent advantage over RoBERTa.

The **BioBERT** [114] representation model (almost 2000 Google Scholar citations at the time of this writing) was initialized with BERT weights and then pretrained using domain-specific literature, namely PubMed abstracts (4.5B words) and PubMed Central full-text articles (13.5B words). The resulting model was successfully fine-tuned for three biomedical text mining tasks: biomedical named entity recognition, biomedical relation extraction, and biomedical question answering. The BioBERT model was further pretrained for clinical texts using 2M generic clinical notes and discharge summaries [7]. The resulting **Bio+Clinical BERT** showed superior results on clinical NER tasks and medical natural language inference task.

Clinical BERT [80] is similar to the above Bio+Clinical BERT model, but it is trained on 2 083 180 anonymized clinical notes from the MIMIC III database [92] that consists of the electronic health records of 58 976 unique hospital admissions from 38 597 patients in the intensive care unit between 2001 and 2012. The model performed better than BERT on the clinical readmission prediction problem. A similar model is **BLUE BERT** [154], trained on more than 4B PubMed abstracts and 500M MIMIC-III clinical notes. The model showed good performance on BLUE (Biomedical Language Understanding Evaluation) benchmark that includes several tasks relevant to pharmacology, like named entity recognition (see Section 3.1) and relation extraction (see Section 3.2).

In the light of COVID-19 epidemics, Khadhraoui et al. [104] have prepared a specialized BERT model, called **CovBERT**, intended to improve the COVID-19 literature review. The model, based on BERT, was pretrained on 4304 PubMed abstracts on several topics such as COVID-19 treatment, COVID-19 symptoms, virology, public health, and mental health. CovBERT showed better classification accuracy on this dataset compared to baseline RoBERTa, AlBERT, SciBERT, BioBERT, and Bio+Clinical BERT.

Another popular adaptation to specific terminological areas is named **CharacterBERT** [55]. Instead of using subword tokenization, this approach starts with characters and first constructs words with a convolutional neural network. The pretraining used around 1B tokens from the MIMIC-III clinical dataset and PubMed abstracts. The effectiveness of this approach was originally demonstrated in the biomedical domain using four tasks: medical entity recognition, medical natural language inference, relation extraction (Chem-Prot database and drug-drug interactions), and clinical sentence similarity. The resulting CharacterBERT models performed on par or better than BERT.

As evident from many citations, the BERT enhancements received, these models were successfully applied to many relevant pharmacological problems. We list a sample of works addressing a few relevant problems and approaches in Section 3.

Name	Description	Trained on	Usage
		Static embeddings	
word2vec [144]	General static word embeddings	Any collection of text, e.g., Wikipedia dump	Any general non-contextual text processing.
PubMed-PMC, WikiPubMed-PMC [166]	Word2vec adapted to life-sciences	PubMed abstracts and articles; in combination with Wikipedia	Any non-contextual life-science text processing, e.g., biomedical NER for genes, chemicals and diseases [69]

BioVec, ProtVec, GeneVec [11] dna2vec [147]	Word2vec style embeddings for biological sequences, genes, and proteins	Different biological sequences, e.g., Swiss-Prot	Proteomics and genomics, e.g., structure prediction for proteins.
Contextual embeddings			
BERT [48], RoBERTa [128]	General contextual text embeddings	Large general text corpora such as Wikipedia and Common Crawl.	Any general text processing.
SciBERT [15]	Contextual embeddings for scientific texts.	Scientific papers from Semantic Scholar	NER for clinical use, text classification, relation classification
Character BERT [55]	Character-level input allows for easy adaptation to different areas.	Clinical texts and PubMed abstracts.	Medical NER, NLI, RE, and clinical sentence similarity.
BioMed-RoBERTa [68]	RoBERTa adaptation for life-sciences	Standard texts, abstracts, and full papers from PubMed.	Chemical-protein-disease annotations, sequential sentence classification task
BioBERT [114]	BERT adapted to life-sciences	BERT further trained on PubMed abstracts and papers	biomedical NER, RE, and QA
Bio+Clinical BERT [7]	BioBERT adapted to clinical texts	BioBERT further pretrained with clinical notes and discharge summaries.	Clinical NER and medical NLI.
Clinical BERT [80]	Suitable for clinical texts	Clinical notes from EHR for patients in intensive care units.	Clinical readmission prediction.
BLUE BERT [154]	Suitable for clinical texts	PubMed abstracts and clinical notes.	Good performance on BLUE benchmark, including NER and RE.
CovBERT [104]	BERT adapted to COVID-19.	BERT further pretrained on PubMed abstracts with COVID-19 relevant contents.	Tasks related to COVID-19.

Table 1: Representation models (i.e. embeddings for text and biological sequences) useful for pharmacology.

2.1.4 Languages Other than English

While the majority of NLP in pharmacology is focused on English, there are also some exceptions. Akhtyamova [3] trains a domain-specific BERT model for Spanish on a relatively small dataset (87M tokens) and successfully applies it to the problem of NER in Spanish. In the context of the annual workshop on BioNLP Open Shared Tasks, in 2019[2] one of the tasks, PharmaCoNER (Pharmacological Substances, Compounds and proteins and Named Entity Recognition track), addressed the mentioning of chemicals and drugs in Spanish medical texts. The task included two tracks: one for the NER offset and entity classification and the other one for the concept indexing. In their entry, Xiong et al. [222] devised a system based on BERT for the NER offset and entity classification and Bi-LSTM with max/mean pooling for concept indexing. On the same tasks, Sun et al. [195] compared several BERT variants (see Section 2.1.3): BLUE BERT [154], multilingual BERT [48], SciBERT [15], BioBERT [114], and Spanish BERT [32]. The results show that domain-specific pretraining is successful and better than the language-specific BERT variant.

For the adverse drug reaction relation extraction in Russian, Sboev et al. [184] have preliminary trained multilingual XLM-RoBERTa [43], and Russian RuBERT [111] models on Russian drug review texts, followed by fine-tuning on the created training dataset. The results showed that the former multilingual model is advantageous.

[2] https://2019.bionlp-ost.org/

2.2 Injecting Pharmacological Knowledge into Deep Neural Networks

While large pretrained language models have significantly increased the performance of machine learning approaches for most NLP tasks, many shortcomings still make the approaches less robust as desired. Examples of weaknesses are processing of negation, uncertainty about factual knowledge, and lack of problem-specific knowledge [178].

The knowledge injection approaches attempt to address the shortcomings of large pre-trained models by utilizing external knowledge resources in various forms, such as knowledge graphs (KGs, see Section 5) and other types of knowledge bases. This can reduce the need for ever-larger language models while improving their interpretability. In general, knowledge injection approaches differ in time of injection (during a pretraining phase, as an intermediate task, or in a downstream task), type of injected knowledge (facts, linguistic knowledge, commonsense reasoning, etc.), and type of evaluation (general language, domain-specific language, or probing).

To improve pretrained language models for the biomedical domain, the existing approaches usually use the Unified Medical Language System (UMLS) knowledge base. UMLS is a medical terminology database with hundreds of biomedical vocabulary entries, including definitions of terms and relationships between them. The basic BERT model [48] or any of the specific biomedical BERT models mentioned in Section 2.1.3, are used as a baseline where the knowledge is injected.

Below, we present several approaches to knowledge injection in pharmacology. We divide them into the ones that modify existing pretraining tasks with general improvements in mind and those that focus on better concept representation for a specific task. We summarize the presented models in Table 2.

2.2.1 Modification of Existing Pre-training Tasks for General Improvement

This group of knowledge injection approaches focuses on developing new pre-training tasks or adding new modules to existing pretrained LMs.

Hao et al. [71] improve biomedical LMs for medical downstream tasks by infusing the knowledge base information into the pretraining phase of the Clinical BERT. The authors used the MIMIC-III dataset and continued pre-training on the masked language modeling task and next sentence prediction. They also introduced the task of predicting whether a relationship exists between two concepts in the UMLS knowledge base. Positive instances for this task are taken from the existing relations in ULMS, while negative ones are created through negative sampling as relations in ULMS are very sparse. The final loss function used in training is a combination of all three tasks. The resulting knowledge-enhanced Clinical BERT was evaluated on two named entity recognition datasets and one natural language inference dataset, and the results showed an improvement over the baseline biomedical models BioBERT and Clinical BERT.

UmlsBERT [143] also integrates external knowledge resources to improve biomedical language models. The authors updated the masked language modeling in the pre-training step with the associations between the words specified in the UMLS. Firstly, at the input level, medical terms are enhanced by their semantic types (UMLS contains 44 unique semantic types). For example, the model receives information that 'lungs' are 'body part', 'organ' etc. This represents an additional input layer that must be trained. Words without semantic type are represented by a zero-filled vector. Secondly, the masked language modeling (MLM) task is modified: instead of predicting one missing token, the model predicts all words associated with the same concept unique identifier (CUI). For instance, where the standard MLM task predicts only 'lung', the modified one predicts 'lungs' and 'pulmonary' as well. UmlsBERT achieves the best results in four out of the five tasks (one NLI and four NER tasks). The ablation study checking if semantic type information improves the performance shows that the model performs significantly worse on all tasks without it.

Meng et al. [142] improve biomedical BERTs by partitioning a very large KG into smaller subgraphs and infusing this knowledge into various BERT models using adapters. Adapters [79] are BERT additions that add only a few new trainable parameters while the original weights remain fixed. This reduces the inefficiency of fine-tuning large models for each task and allows a high degree of parameter sharing. Meng et al. [142] construct two KGs from the UMLS knowledge graph. The METIS algorithm [103] partition the knowledge graph into n subgraphs. Following that, they train an adapter module for each sub-graph to predict the tail entity of a triplet from the sub-graph. Finally, they use AdapterFusion mixture layers [158] to combine the knowledge from adapter modules. They experimentally determined that 20 sub-graphs and PubMedBERT yielded the best results. Their approach improves performance on QA, NLI, and document classification tasks in the biomedical domain.

2.2.2 Improved Concept Representation for Specific Tasks

This group of knowledge injection approaches focuses on improving concept representations for specific tasks.

The same medical concepts can be represented by a variety of nonstandard names, misspellings, and abbreviations. Term normalization is a task that addresses this problem. CODER [233] proposes dual contrastive learning simultaneously on both terms and relation triplets from the UMLS KG. The approach is motivated by examples such as that it is better to have "rheumatoid arthritis" closer to "osteoarthritis" than "rheumatoid pleuritis" because both are subtypes of arthritis. Relations between terms express that and thus provide useful information during the training. CODER maximizes similarities between positive term-term pairs and term-relation-term pairs from the KG. They evaluate their approach on datasets in different languages consisting of term normalization, relation classification, and conceptual similarity tasks. Their approach significantly outperforms existing medical embeddings in zero-shot term normalization.

Liu et al. [124] address the problem of entity linking, specifically, the heterogeneous naming of medical concepts. The authors pre-train a transformer-based language model on the UMLS biomedical KG. They propose a metric learning framework that learns to cluster synonyms of the same concept. The goal of a self-alignment pre-training step is to learn such concept embeddings that maximize the similarity between two concepts based on the cosine similarity measure. The learning setup consists of triplets in the form (x_a, x_p, x_n), where x_p is a positive match for x_a and x_n is its negative match. This approach first samples hard triplets (triplets that contain negative pairs closer in space than positive pairs with basic BERT embedding by some margin). It learns to push negative pairs away from each other and positive pairs together by considering the multi-similarity loss function. The resulting SAPBERT improves the accuracy across six medical entity linking tasks (up to 20%) compared to the domain-specific BERT models and achieves state-of-the-art results.

Mao and Fung [136] tackle the problem of measuring semantic relatedness between biomedical concepts (UMLS concepts). Semantic similarity expresses the relatedness of two concepts in their meaning and is an important tool for automatic spelling correction, information retrieval, and word sense disambiguation. Authors use pre-trained word embedding models (e.g., BioWordVec [235], variations of BERT, etc.) to generate concept sentence embeddings from UMLS, and various graph embedding models (e.g., GCNs [106], TransE [24] and its variants). In addition to that, they combined both concept sentence embeddings and graph embeddings by concatenation. The similarity score between two embeddings was computed using the cosine similarity measure. The combined word and graph embeddings produced the best results on three semantic relatedness datasets and a one-word sense disambiguation dataset.

Name	External knowledge	Pretrained model	Evaluation tasks
Hao et al. [71]	MIMIC-III, UMLS	ALBERT	NER, NLI
UmlsBERT [143]	MIMIC-III, UMLS	Bio_ClinicalBERT	NER, NLI
Meng et al. [142]	UMLS	PubMedBERT	document classification, NLI, QA
CODER [233]	UMLS	PubMedBERT, mBERT	term normalization
SAPBERT [124]	UMLS	PubMedBERT	entity linking
Mao and Fung [136]	UMLS	BioWordVec	semantic relatedness, WSD

Table 2: Summary of knowledge enhanced models. All methods are evaluated on more than one pretrained model. Here we report the one that achieved the best results. For CODER, we report the best monolingual in multilingual versions of the models.

2.3 Explainable NLP in Pharmacology

Deep learning models commonly surpass standard machine learning models in terms of predictive performance. However, their decision-making process is typically opaque, meaning that it is difficult to explain why the model made a certain prediction. Understanding models' inner workings are helpful for debugging errors, possibly improving their performance, and gaining scientific insights into the modeled process, e.g., why two drugs interact in the drug-drug interaction identification. Additionally, as pharmacology is concerned with drugs affecting humans, it is essential that predictions are safe and verifiable.

Depending on the time when an explanation is created, there are two types of explanation methods: intrinsic and post-hoc [133]. Intrinsic methods use a model's architecture or its components to construct an explanation. A simple example is a linear regression model using binary bag-of-words features. The learned weights associated with the input words represent an explanation of the prediction for the given input (positive weights indicate the positive impact of words on the decision, and negative weights indicate negative impact). Another commonly used intrinsic method used for large pretrained transformer models is the inspection of attention weights, which intuitively represent the parts of the input the model focuses on. Attention, being the key component of the currently dominant transformer-based

models, is easy to compute. However, multiple attention heads may be difficult to comprehend, and the alignment between attention explanations and the underlying model behavior (i.e. actual explanations) is questionable [86; 214].

Post-hoc explanation methods construct an explanation after a model is trained. While intrinsic methods are based on the design of a specific model, post-hoc methods are typically model-agnostic. An example of such methods are perturbation-based explanation methods such as Local Interpretable Model-agnostic Explanations (LIME) [175], SHapley Additive exPlanations (SHAP) [131], and Interactions-based Method for Explanation (IME) [193]. They work by repeatedly modifying (perturbing) the input, observing the changes in the output, and modeling their associations using a surrogate model. Post-hoc methods are convenient due to their flexibility in the choice of used model architectures. However, the faithfulness of the produced explanations may be poor [190; 58] as they explain the model from an external perspective.

Both intrinsic and post-hoc methods have been successfully applied to general [112] and topic-specific language tasks in bio-medicine [145]. Below, we describe several cases of using explanation methods in pharmacology. We provide an overview of the methods in Table 3.

Jha et al. [88] make pre-trained word embeddings more interpretable by learning a transformation to a more interpretable embedding space with the retained performance. The interpretable word embeddings correspond to categorical embeddings, trained separately using expert-provided definitions and additional knowledge from a biomedical knowledge graph.

Wawrzinek et al. [210] introduce an entity embedding-based explanation method for drug-disease association (DDA) prediction. They construct explanations following the drug-centric and the disease-centric notion of similarity:

- drug-centric: "if two drugs are chemically similar, they likely have a similar relationship with the target disease";
- disease-centric: "a drug has the same relation for similar diseases".

To obtain the explanation, they embed the drug and the disease from a DDA pair and retrieve k intermediate entities (drugs or diseases) using a cosine similarity-based metric. An explanation instance is created based on the relationship between the drug, disease, and the intermediate entity in existing publications. For example, if the intermediate entity is a drug and the intermediate drug treats the input disease, the input drug is assumed to also treat the input disease with confidence proportional to the embedding similarity. The obtained explanations using k intermediate entities are aggregated into the final DDA prediction, e.g., using a majority vote.

Huang et al. [81] include an interpretable component in their drug-drug interaction (DDI) prediction system. The component projects the latent embedding of the input drug pair into a more interpretable subspace, whose basis consists of frequently occurring molecular substructures. The substructures are extracted from a database of drug representations by finding substrings with a high enough frequency. The projection into the subspace aims to capture the relevance of the molecular substructures towards the drug interaction prediction.

Yazdani-Jahromi et al. [230] propose an attention-based drug-target interaction (DTI) prediction system, using the attention weights as an explanation. They demonstrate the high predictive performance of their system on three benchmark datasets, while they demonstrate the interpretation capability of their model on a DTI prediction example via visualization.

Bradshaw et al. [26] present a generator of product molecules from a set of common reactant molecules. It is composed of

- an encoder-decoder model between a latent space and a list of reactant molecules, and
- a reaction prediction model that transforms the reactants into a list of product molecules.

The second component introduces interpretability to the model as it provides some insight on *how* the product molecules are constructed out of the reactants. However, the authors do not put an emphasis on evaluating the interpretability of their approach.

Jiang et al. [90] present an approach for detecting potential adverse medication effects from social media posts. The detection is posed as a word analogy task: given a known possible side effect of a drug, the task is to find similar pairs of drugs and corresponding side effects with a similar relation. The known possible side effects are taken from the SIDER database [109], while the static word embeddings are trained on unlabeled tweets. The found potential side effects are subject to human examination along with relevant tweets expressing the effect.

Rodríguez-Pérez and Bajorath [177] present a usability study of the SHAP explanation method for explaining complex compound activity prediction models. They find that SHAP produces consistent feature attributions across three

complex models. Additionally, they demonstrate how the obtained attributions can be used to find potential biases in the models.

Pope et al. [161] present adaptations of three explanation methods for explaining graph convolutional neural networks: contrastive gradient-based saliency maps, class activation mapping, and excitation backpropagation (EB). They test the methods on molecular graph classification, where the task is to predict whether molecules possess certain properties, such as toxicity. The explanations are salient subgraphs, which can be interpreted as functional groups responsible for the molecular property (according to the model). By analyzing the explanations using automated metrics (fidelity, contrastivity, and sparsity), the authors conclude that the gradient-weighted class activation mapping is the most suitable out of the tested methods, although they emphasize the need for detailed studies of chemical validity of the explanations in future work.

In summary, explanation methods have been adopted across a variety of pharmacology applications. We find that the authors typically use the explanation methods in one of two ways, either using the explanations as a safety mechanism for a semi-automatic use of the model predictions, or as a way to obtain plausible hypotheses that are then manually verified, for example using additional experiments. The proposed explanation methods for pharmacology commonly use a connection to an external knowledge source. We believe that the incorporation of external knowledge into explanation methods is a promising direction for further research as the prediction may not be intuitively explainable to humans in terms of only input components. In addition, external human-curated knowledge may naturally be more intuitive to end-users.

Reference	**Explanation type**	**Short description**	**Downstream tasks**
Jha et al. [88]	intrinsic	Interpretable word embedding transformation	semantic concept categorization
Wawrzinek et al. [210]	intrinsic	Embedding arithmetic (analogies)	drug-disease association prediction
Huang et al. [81]	intrinsic	Interpretable subspace	drug-drug interaction prediction
Yazdani-Jahromi et al. [230]	intrinsic	Interpretable component (attention weights)	drug-target interaction prediction
Bradshaw et al. [26]	intrinsic	Interpretable component (reaction predictor)	molecule generation
Jiang et al. [90]	post-hoc	Present representative examples	detection of potential adverse medication effects
Rodríguez-Pérez and Bajorath [177]	post-hoc	Out-of-the-box method (SHAP)	structure-activity relationship modeling
Pope et al. [161]	intrinsic	Adapt out-of-the-box methods (gradient-based saliency, CAM, EB)	identification of biological molecular properties

Table 3: An overview of the explanation methods used in NLP for pharmacology.

3 Common NLP Tasks and Applications

Several NLP tasks are frequently tackled in the pharmacological context. Some of them are adapted from general NLP tasks (e.g., named entity recognition, relation extraction, and question answering). In contrast, others are specific to pharmacology (e.g., adverse drug reactions and literature-based drug discovery). We have mentioned some successful uses of contextual BERT models on these tasks in Section 2.1.3, but this mainly demonstrated the usability of these models. This section systematically analyzes the most important tasks in life sciences and pharmacology. As hundreds of works tackle these problems exclusively or among other problems, we review a sample of recent works. The overview is presented in Table 4.

3.1 Named Entity Recognition for Pharmacology

Named entity recognition (NER) – called entity identification, entity chunking, or entity extraction, is one of the most popular NLP techniques that classifies named entities in text into pre-defined categories such as person, time,

location, organization, etc. In the biomedical context, the entities of interest can be cells, genes, gene sequences, proteins, biological processes and pathways, diseases, drugs, drug targets, compounds, adverse effects, metabolites, tissues, and organs [155; 23]. NER is often used as the initial stage of analyses to provide semantic interpretations of unstructured text by identifying and categorizing concept references. Various concepts are detected with different degrees of difficulty. The critical issue in recognizing chemicals, for example, is the high variance in concept names and chemical formulas. In contrast, the main challenge in identifying gene functions is the high degree of uncertainty caused by species diversity.

In pharmacology domain, NER is often used as the first step of the relation extraction task (see Section 3.2) [101][67] or adverse drug reactions task (see Section 3.3) [118]. Many authors start with the MADE 1.0 challenge dataset, e.g., Jagannatha et al. [85] finds the medications and their attributes, Chapman et al. [35] apply the conditional random field method for medication recognition, Yang et al. [227] developed the MADEx model based on LSTM networks for the same purpose, and Wunnava et al. [219] apply the Bi-LSTM model.

3.2 Relation Extraction for Pharmacology

The relation extraction task is part of information extraction (IE) and extracts semantic relationships from texts. The extracted relationships connect two or more entities of the same kind that fit into one of many semantic categories (for example, people, organizations, or places). Frequently, extracted relations are related to adverse drug reactions (ADR) and drug-drug interactions (DDI), relations between medications, between their attributes such as dosage, route, frequency, and duration [85]. The ability of NLP models to automatically detect adverse drug event (ADE) related terms in textual data helps avoid ADEs. This results in safer and better quality healthcare services, lower healthcare expenditures, more educated and engaged customers, and improved health outcomes.

In pharmacology, relation extraction typically processes scientific papers that provide novelties from the pharmacology. Classical approaches extracted semantic relationships with a pattern-based approach to find medical relations in pharmaceutical texts [16; 180]. Dep learning approaches brought significant improvements [118; 227]. Lately used approaches apply pretrained language models, e.g., SemRep [105]. The extracted information is sometimes used to construct graphs encoding drug-drug, and disease-drug relationships, representing the similarity between them [238]. Although most approaches are based on textual data, relations are also discovered through the analysis of EHR data [36].

3.3 Adverse Drug Reactions

Adverse drug reaction (ADR) is defined as a considerably damaging or unpleasant reaction occurring from an intervention associated with the use of a pharmaceutical product. Adverse reactions frequently anticipate danger from future administration and demand avoidance, particular therapy, or dose regimen modification [160]. ADRs have traditionally been divided into two categories. Type A responses are dose-dependent and predicted based on the drug's pharmacology (also known as enhanced reactions). In contrast, Type B responses, often known as weird reactions, are distinctive and unpredictable from the pharmacological point of view.

Implementation-wise, ADR extraction is similar to relation extraction, where ADRs connected to to various diseases and drugs are detected. Lately, large pretrained language models,such as BERT, are used in ADR extraction [27; 122; 83]. Again, texts are not the sole source of information, and EHRs are often used as additional information in ADR extraction [118; 219; 35].

3.4 Literature Based Drug Discovery

LBD (literature-based discovery) is an automatic or semi-automatic method for discovering new information from the literature. The amount of scientific literature is steadily growing, driving researchers to become more specialized and making it challenging to track developments even in restricted fields [73]. If text is identified that overtly asserts the knowledge that "A is associated with B" and "B is associated with C" in the Swanson ABC co-occurrence model [198], then the implicit knowledge of "A may be associated with C" is obtained. LBD is essential for biomedical NLP since it allows finding implicit information that can help to enhance biomedical research. A recent study presents the computational strategies utilized for LBD in the biomedical area [65]

LBD applies several NLP tasks to process the pharmacological and medical literature, with the purpose to detect new medical entities [209; 183; 50], extract relations [164; 207] or reactions [239]. Some approaches use scientific texts for protein engineering, and visualization [18]. Frequent information source is the PubMed engine together with the PubTator model [211] for automated annotation. The PharmKE tool [91] labels pharmaceutical entities and the relationships between them. In new diseases, such as COVID-19, LBD technique have proved useful to extract relevant

information [159; 139]. Another frequent task is **drug repositioning** which helps to find another purpose for existing drugs, i.e., to use them in treating similar diseases [223]. Alternatively, novel drug indications can be discovered by analyzing the medical history, as exemplified in the PREDICT model [66].

3.5 Question Answering

Question answering (QA) is an NLP task that takes a question as input and returns an answer in the form of a ranked list of relevant replies, or a summary answer snippet [42]. In a classical (pre-neural) approach, QA incorporates three tasks: information retrieval, retrieving relevant documents or passages for a particular query, and text summarization that summarizes the reply from relevant passages. A related information retrieval task is called "Learning by Doing" and searches the knowledge base for entities most related to the ones mentioned in the question. This task is divided into ranking the texts found in the database and finding the correct answer among the recovered paragraphs.

QA can summarize the pharmacological literature, e.g., for new diseases like COVID-19 [194]. The data are mainly from PubMed articles, and in the case of COVID-19, also news about this disease [115]. To answer pharmaceutical questions, the QA task can be applied in many languages, even in low-resource languages such as Persian [204]. Another source of information can be linked data as used in the GFMed model [137].

Task	Description	Referenced papers
Named Entity Recognition	Identifying pharmaceutical entities in textual data	[85] [35] [219] [67] [101] [118] [227]
Relation Extraction	Finding relation between drugs and diseases from scientific text resources	[16] [118] [36] [105] [227] [180] [238]
Adverse Drug Reactions	Anticipate danger from future administration and demand avoidance, particular therapy, or dose regimen modification	[27] [122] [83] [118] [219] [35]
Literature Based Drug Discovery	Discovering new pharmacological information from existing literature.	[239] [18] [211] [209] [159][139][183] [91] [164] [207] [50] [223] [66]
Question Answering	Answers given question with the most relevant response	[194] [115] [56] [204] [137]

Table 4: Overview of tasks related to the pharmacology.

4 Data Resources

As the application of open science and open data principles is rising [28], the number of publicly available datasets is steadily growing. This makes finding and discovering appropriate datasets increasingly challenging. There are two strategies to find a dataset suitable for a given task. First, a bottom-up approach starts by searching available datasets and evaluating their utility for the given problem. Second, a top-down approach first finds relevant papers for the tackled topic and then explores the available datasets used in the papers.

We first present an overview of specialized search engines for discovering and finding datasets in 4.1. Then we give an overview of the most important datasets utilized in published papers related to NLP in pharmacology. The covered datasets are organized into five groups: patient data, drug usage data, drug structure data, question answering datasets, and general pharmacological data. In Section 4.2, we present datasets containing patients' history and medical notes. The datasets in Section 4.3 contain drug characteristics according to the prescriptions to patients, while in Section 4.4, we cover datasets with information about drugs' chemical composition. Datasets supporting question answering systems in pharmacology are described in 4.5. Section 4.6 describes general resources useful for successful NLP in pharmacology.

We include public and closed (private/commercial) data in the survey. The summary of datasets is contained in Table 5, where for each dataset, we include a list of references where the dataset was used, a short description, the size of the dataset, and its typical usage.

4.1 Finding and Discovering Datasets

As the number of datasets rapidly grows, it becomes essential to have effective tools for finding them. As a solution, there are several specialized search engines for discovering and finding datasets.

Google's Dataset Search[3] currently indexes more than 30 million publicly available datasets. Filters can limit the results based on licensing (free or premium), format (CSV, images, etc.), and update time. Alternatively, a specialized cloud platform **data.world**[4] hosts an enterprise data catalog with over 130,000 datasets and knowledge graphs. Another platform hosting public datasets is **Kaggle**[5], which is primarily a machine learning competition platform, but it also includes a dataset search engine.

The NLP community usually publishes the source code and datasets in the **Github**[6] repository so that this source control platform can be used for dataset discovery. A specialized platform indexing the code and data related to research papers is **Papers with Code**[7]. This platform offers research area-based organization of papers allowing for a convenient discovery and browsing of papers and datasets. One of the most popular development platforms for NLP, the **Huggingface**, offers a good dataset search engine organized by NLP task, category, language, size, and license[8].

A specialized search engine for linked data is the **Linked Open Data (LOD) Cloud**[9] that allows for text-based search and entity lookup. LOD Cloud is a distributed web of interconnected datasets (over 1500 datasets) containing open data in a structured and semantically annotated format from multiple domains - life sciences, publications, government, media, etc. The background on the LOD Cloud is described in Section 5.

4.2 Patient Data

Datasets with the information about patients typically contain patients' medical history or medical notes about them. The main application of these datasets is to find novel relations between drugs and diseases. Below, we briefly describe the most commonly used patient datasets.

MIMIC-III: Medical Information Mart for Intensive Care[92][10] is a dataset that contains data on patients hospitalized in large tertiary care hospitals critical care units. It contains information on vital signs, medicines, laboratory measurements, care providers' observations and notes, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival statistics, etc. This dataset contains data on over 40 000 patients.

MADE1.0 Database [84][11] is a Electronic Health Record (EHR) database that is a part of the MADE1.0 competition. The structured dataset contains information on taken drugs, experienced ADEs (Adverse Drug Events), and indications and symptoms of patients. The competition addressed three tasks: NER, relation identification, and a joint NER-RI task. The dataset contains 1089 patient notes with detailed named entity and relation annotations.

n2c2 NLP Research Database [74][12] is database used for Track 2 of the 2018 National NLP Clinical Challenges shared task. The data is extracted from the MIMIC-III (Medical Information Mart for Intensive Care-III) clinical care database. The records were chosen using a query that looked for ADEs in the description of records' ICD (International Classification of Diseases) code. The retrieved records were manually inspected to ensure that at least one ADE was present and adequately annotated. The dataset contains 505 discharge summaries in textual format.

MarketScan[1][13] dataset is a collection of administrative claims databases that includes information on in-patient and out-patient claims, out-patient prescription claims, clinical usage records, and healthcare costs in US. The three main databases each contain a convenience sample for one of the following patient populations: (1) employees with contributing employers' health insurance, (2) Medicare beneficiaries with employer-paid supplemental insurance, and (3) Medicaid recipients in one of eleven participating states. The data is not in textual format but can be used with NLP applications. The database contains data on approximately 43,6 million persons.

[3] https://datasetsearch.research.google.com/
[4] https://data.world/
[5] https://www.kaggle.com/datasets
[6] https://github.com/
[7] https://paperswithcode.com/datasets
[8] https://huggingface.co/datasets
[9] https://lod.openlinksw.com/
[10] https://mimic.mit.edu/
[11] https://bio-nlp.org/index.php/announcements
[12] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
[13] https://www.ibm.com/products/marketscan-research-databases

4.3 Drug Usage Data

Datasets described in this section provide information on drugs' usage, usage instructions, effects, pharmaceutical properties, and composition.

DailyMed Database [150][14] is a web database provided by the National Library of Medicine (NLM) in US. The US Food and Drug Administration (FDA) updates the material daily. The DailyMed contains prescription and nonprescription medications for human and animal usage, medical gases, gadgets, cosmetics, nutritional supplements, and medical foods. The labeled drugs describe the composition, form, packaging, and other properties of drug products according to the HL7 Reference Information Model (RIM). These details are given in the descriptive text format. The database contains 142 981 labels.

DrugBank Database [216][15] is one of the biggest drug databases. Besides drugs, it contains drug paths which show how the drug travels in the human body and allows search for indications and drug targets. For an individual drug, the database contains all the brand names, background information in the text form, its type, structure, weight, formula, other names it is called by, what it is used for, what therapies it is used in, indications, doses, interactions, etc. All the details for each drug are available online and given in the descriptive text format. The database contains descriptions of 14 665 drug entries.

4.4 Drug Structure Data

Datasets covered in this section contain drug characteristics regarding their chemical composition. Mainly, they are used for discovering new drugs or finding protein-protein interactions between drugs.

ChEMBL Database [61][16] is an open-source database that contains binding, functional, and ADMET (Chemical absorption, distribution, metabolism, excretion, and toxicity) data for a wide range of drug-like bioactive chemicals. These data are regularly manually extracted from the published literature, then selected and standardized to enhance their quality and usability across a variety of chemical biology and drug-discovery research uses. The database includes 2.4 million bioassay measurements spanning 622 824 chemicals, including 24 000 natural products. The contents were produced by sifting through over 34 000 papers published in twelve medicinal chemistry journals. The data from the journals containing details can also be used.

UMLS: The Unified Medical Language Database [21][17] is a database of biomedical vocabularies. The NCBI (National Center for Biotechnology Information) taxonomy, Gene Ontology, MeSH (Medical Subject Headings), OMIM (Online Mendelian Inheritance in Man), and the Digital Anatomist Symbolic Knowledge Base are all included in the UMLS MetaThesaurus. The UMLS is not a textual database but is frequently used in NLP tasks, such as extracting concepts, relationships, or knowledge of pharmacological entities from texts. The UMLS has about 2 million names for over 900 000 concepts from over 60 biomedical vocabularies and 12 million relationships between them.

PDB: The Protein Data Bank Database [165][18] is a global repository of structural data for biological macromolecules. To obtain the data, depositors used X-ray crystal structure determination, NMR (Nuclear magnetic resonance), cryo-electron microscopy, and theoretical modeling. The search queries also return the literature from which the data is extracted, e.g., the abstracts from medical articles that can be further used for NLP. The number of papers accessible in the textual format is not available, but the database contains 133 920 Biological Macromolecular Structures, each accompanied by a related abstract.

ChemProt Database[199][19] is a biology annotated database based on several chemical-protein annotation resources, together with disease-associated protein-protein interactions (PPIs). ChemProt was utilized in the BioCreative VI text mining chemical-protein interactions shared task. The data contains PubMed abstracts in textual format together with annotated entities and interactions. The database has 1820 abstracts.

4.5 Question Answering Data

This section covers some datasets that can be used to build pharmacological question answering models.

[14] https://dailymed.nlm.nih.gov/dailymed/index.cfm
[15] https://go.drugbank.com/
[16] https://www.ebi.ac.uk/chembl/
[17] http://umlsks.nlm.nih.gov
[18] http://www.rcsb.org/pdb/
[19] https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi/

MQP Database[141][20] comprises 3048 question-answer pairs that are categorized as similar or distinct by medical experts (i.e. not particular to COVID-19). Two doctors collaborated on the annotation and their agreement on 836 question pairings in the test set was above 85%.

COVID-Q Database[212][21] is a collection of 1690 COVID-19-related questions divided into 15 general categories and 207 specific question classes. The dataset was annotated in three stages by many curators. First, two curators discussed and categorized the questions. Second, an external curator reviewed the work and, if necessary, proposed adjustments to the categories. Third, questions from more than four different question classes were sampled and allocated to three different AMT (Amazon Mechanical Turk) workers. The validation was based on the majority vote.

CovidQA Database[236][22] is made up of 124 question–article–answer triplets taken from 85 different articles in CORD-19 Kaggle challenge and covers 27 different categories. Five curators created annotations by synthesizing questions from the challenge organizers' categories, then manually discovered relevant articles and replies.

4.6 General Pharmacological Data

In this section, we describe five resources that are general and useful for many tasks.

Wikipedia [215][23] is a well known encyclopedia and web-based collaborative database consisting of over 15 billion articles. Wikipedia contains articles from different scientific fields written in many languages.

PubMed [31][24] is a free web engine for primarily MEDLINE, bibliographic database encompassing medicine, nursing, dentistry, veterinary medicine, the health-care system, and preclinical sciences like molecular biology. More than 4600 biomedical journals are indexed in MEDLINE, together with bibliographic citations and author abstracts. PubMed indexes more than 30 million articles and abstracts.

LitCovid Database [38][25] is a curated literature site for tracking up-to-date scientific knowledge regarding the COVID-19 disease. It is the most comprehensive resource on the topic with central access to more than 255 935 relevant PubMed articles. The articles are updated daily and divided into categories based on research themes and geographical areas.

CORD-19 (COVID-19 Open Research Database) [206][26] contains metadata about papers related to COVID-19. The main sources are PubMed, World Health Organization, bioRxiv and medRxiv. This database contains over 52 000 papers.

DBpedia [12][27] is a structured open-source database with information extracted from Wikipedia articles. For drugs, it contains basic information on uses, contained chemicals, drug type, links to other languages, Wikipedia links, and other links used to extract information. The database contains more than 10 000 drug type entries.

Name	Description	Entries	Usage
	Patient Data		
MarketScan [1][28]	Collection of administrative claims	43,600,000	NER, ADE, Drug-drug interaction
MIMIC-III [92][29]	Data on patients hospitalized	40,000	Drug discovery, ADE, Drug-drug interaction
MADE 1.0 [84][30]	A challenge dataset with 21 EHRs of cancer patients	1,089	NER,ADE

[20] https://github.com/curai/medical-question-pair-dataset
[21] https://paperswithcode.com/dataset/covid-q
[22] https://aclanthology.org/2020.nlpcovid19-acl.18/
[23] https://en.wikipedia.org/wiki/Main_Page
[24] https://pubmed.ncbi.nlm.nih.gov/
[25] https://www.ncbi.nlm.nih.gov/research/coronavirus/
[26] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge
[27] https://www.dbpedia.org/
[28] https://www.ibm.com/products/marketscan-research-databases
[29] https://mimic.mit.edu/
[30] https://bio-nlp.org/index.php/announcements

n2c2 [74][31]	Unstructured notes from the Research Patient Data	505	ADE
Drug Usage Data			
DailyMed [150][32]	Drug label database	142,981	NER, Drug-drug interaction, ADE
DrugBank [216][33]	Database of drugs and drug products	14,665	ADE, pharmacovigilance, standardization, interactions
Drug Structure Data			
ChEMBL [61][34]	Binding, functional, and ADMET data	2,400,000	ADE, pharmacovigilance, standardization, interaction
UMLS [21][35]	Biomedical vocabularies	2,000,000	ADE
PDB [165][36]	biological macromolecules	133,920	ADE, pharmacovigilance, standardization, interaction
ChemProt [199][37]	Biological annotations	1,820	ADE
Question Answering Data			
MQP[141][38]	Collection of medical related pairs of questions and answers	3,048	QA
COVID-Q[212][39]	Collection of COVID-19-related questions divided into 15 general categories and 207 specific question classes	1,690	QA
CovidQA[236][40]	Collection of question–article–answer triplets taken from 85 different articles in CORD-19	124	QA
General Pharmacological Data			
Wikipedia[215][41]	Online free encyclopedia	15,000,000,000	ADE,Drug-drug interaction,Drug discovery, NER
PubMed[31][42]	Web engine for searching health articles	30,000,000	ADE,Drug-drug interaction,Drug discovery, NER

[31] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/
[32] https://dailymed.nlm.nih.gov/dailymed/index.cfm
[33] https://go.drugbank.com/
[34] https://www.ebi.ac.uk/chembl/
[35] http://umlsks.nlm.nih.gov
[36] http://www.rcsb.org/pdb/
[37] https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi/
[38] https://github.com/curai/medical-question-pair-dataset
[39] https://paperswithcode.com/dataset/covid-q
[40] https://aclanthology.org/2020.nlpcovid19-acl.18/
[41] https://en.wikipedia.org/wiki/Main_Page
[42] https://pubmed.ncbi.nlm.nih.gov/

LitCovid[38][43]	Scientific PubMed articles related with COVID-19	255,935	ADE,Drug-drug interaction,Drug discovery, NER
CORD-19 [206][44]	Scientific papers relevant to COVID-19 research	52,000	ADE, Drug-drug interaction, Drug discovery, NER
DBpedia[12][45]	Articles and structured data on e.g., drugs and diseases	10,000	ADE, pharmacovigilance, standardization, interactions

Table 5: Different types of pharmacology-relevant datasets.

5 Knowledge Graphs

The concepts of linked data and knowledge graphs introduced new standards for representing, storing, and retrieving data over the Web, both publicly and privately [19; 20; 72; 218; 77]. As a result of years of adoption of the linked data principles by various data publishers, the Linked Open Data (LOD) Cloud[46] has been created and populated with 1541 interlinked datasets from the domains of geography, government, life sciences, linguistics, media, publications, social networking, user-generated, and cross-domain.

Knowledge graphs, the latest trend in the semantic web and linked data, enable the generation, consolidation, and contextual linking of structured data. The standards and technologies for knowledge graphs solve the problem of having separate 'data silos' in traditional relational database systems, which have to be explicitly mapped to other isolated databases to take advantage of interconnected data [94].

Name	Unique Entities	RDF Statements
Bio2RDF [29]	1,107,871,027	11,895,348,562
HIFM [95; 97]	3,000	21,233
LinkedDrugs [94]	248,746	99,235,032
Covid-19-DS	262,954	69,434,763
KG-Covid-19 [171]	574,778	24,145,556

Table 6: Covered knowledge graphs from the biomedical domain and their characteristics.

The pharmaceutical industry is leading in using knowledge graph-based NLP techniques, especially in patient disease identification, clinical decision support systems, and pharmacovigilance [53]. The problem of identifying patients with specific diseases can be mitigated by knowledge graphs generated from structured and unstructured data from medical records, which capture explicit disease–symptom relationships [36]. Recently, knowledge graphs improved the classification of rare disease patients [121]. In the area of clinical decision support, the combination of NLP and knowledge graphs is employed in inferring drug-related knowledge which is not immediately observed in data, inferring cuisine-drug interactions based on knowledge graphs of drugs and recipes, improving user interaction with relevant medical data, etc. [64; 96; 129; 181; 220; 221]. In pharmacovigilance, the struggles of NLP engines to understand complex language components (e.g., negation, doubt, historical medical statements, family medical history, etc.) from individual case study reports have been significantly mitigated with the use of knowledge graphs [156]. Other examples include the use of knowledge graphs to improve NLP pipelines for detecting medication and adverse drug events from EHRs [148], as well as from Medline abstracts [231].

Section 5.1 presents several knowledge graphs from the biomedical domain used in the mentioned application areas. Given the ongoing COVID-19 pandemic, we outline several recent COVID-19-related knowledge graphs in Section 5.2.

[43] https://www.ncbi.nlm.nih.gov/research/coronavirus/

[44] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge

[45] https://www.dbpedia.org/

[46] https://lod-cloud.net

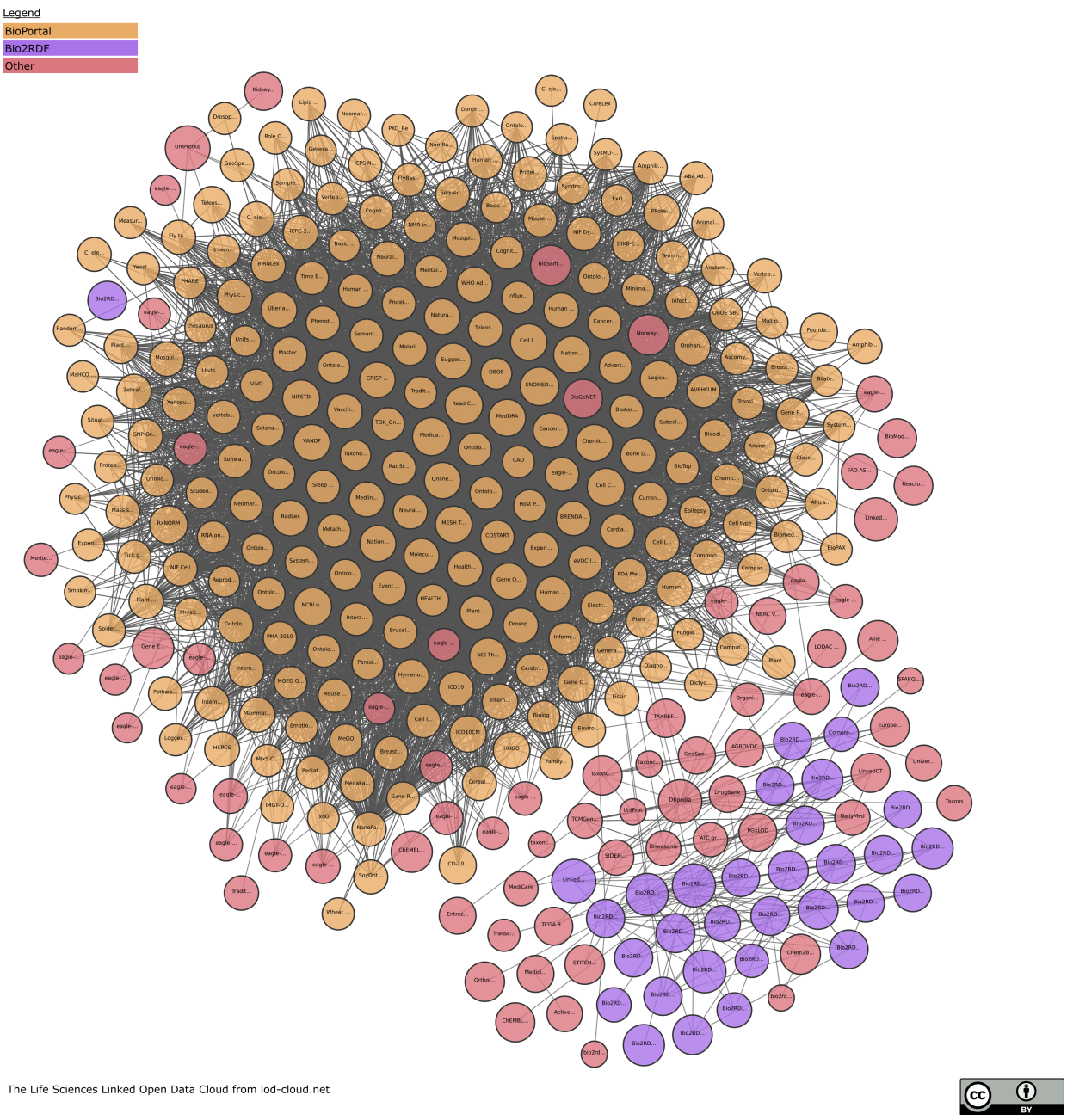

Figure 1: The Life Sciences sub-graph of the Linked Oped Data (LOD) Cloud, as of May 2021. Each node is a dataset. The weighted links denote the amount of RDF triples which link entities from the connected datasets.

5.1 Biomedical Knowledge Graphs

Several projects worked on the transformation of pharmacology-related and healthcare data into linked data and knowledge graphs. Currently, 341 life science datasets are present in the LOD Cloud (see Figure 1). These datasets contain healthcare data from various subdomains, such as drugs, diseases, genes, interactions, clinical trials, enzymes, etc.

Bioportal[47] [213] project hosts ontologies covering drugs, diseases, genes, clinical procedures, etc. With over 980 biomedical ontologies, which define a total of over 13 900 000 classes, it represents the largest such repository in the life-science domain.

[47] `https://bioportal.bioontology.org/`

Bio2RDF[48] [29] is an open-source project which creates RDF datasets from various life science resources and databases and interconnects them into one network [14; 29; 30]. The latest release of Bio2RDF contains around 11 billion triples which are part of 35 datasets. These datasets contain various healthcare data: clinical trials (ClinicalTrials), drugs (DrugBank, LinkedSPL, NDC), diseases (Orphanet), bioactive compounds (ChEMBL), genes (GenAge, GenDR, GOA, HGNC, HomoloGene, MGD, NCBI Gene, OMIM, PharmGKB, SGD, WormBase), proteins (InterPro, iProClass, iRefIndex), gene-protein interactions (CTD), biomedical ontologies (BioPortal), side effects (SIDER), terminology (Resource Registry, MeSH, NCBI taxonomy), mathematical models of biological processes (BioModels), publications (PubMed), etc.

Macedonian Drug Data. Drug data from the Health Insurance Fund of North Macedonia has been transformed into a knowledge graph and linked to other LOD Cloud datasets [95]. This knowledge graph was further extended with linked data about Macedonian medical institutions, and drug availability lists from pharmacies [97].

Cuisine - Drug interactions project used two knowledge graphs for analysis of connections between drugs and their interactions with food, and recipes from different national cuisines, resulting in findings that uncovered the ingredients and cuisines most responsible for negative food-drug interactions in different parts of the world[49] [96].

Global drug data. In this research project, a pipeline-based platform was created to collect, clean, align, consolidate, and create a publicly available knowledge graph of drug products registered in various countries [50] [94]. The source of the data is the official country drug registers. The generated RDF knowledge graph is publicly available through a web-based app[51].

5.2 COVID-19 Knowledge Graphs

COVID-19 pandemic turned the attention of many researchers to life sciences and healthcare domains. Below we list some recent COVID-19-related knowledge graphs.

TypeDB Bio (Covid) knowledge graph[52] contains data extracted from COVID-19 papers and from datasets on proteins, genes, disease-gene associations, coronavirus proteins, protein expression, biological pathways, and drugs. For instance, it allows querying for specific viruses giving associated human proteins related to the virus (e.g., a protein that helps in the replication of the virus). From here, it is possible to identify drugs that inhibit the detected proteins, meaning they can be prioritized in research as potential treatments for patients with the virus. To check the plausibility of this association and the implications, the graph can be used to identify relevant papers in the COVID-19 literature where this protein has been studied (see Figure 2).

Covid-19-DS[53] is an RDF knowledge graph of scientific publications. The base of the graph is the CORD-19 dataset [206] that is regularly updated. The graph generation pipeline applies NER, entity linking, and link discovery to the CORD-19 data. The current version of the resulting graph contains over 69 000 000 RDF triples and is linked to 9 other datasets with over 1 000 000 links.

KG-Covid-19[54] [171] is a framework that allows users to download and transform COVID-19 related datasets and generate a knowledge graph that can be used in machine learning. The project also provides access to pre-built knowledge graphs along with public querying.

6 Tools and Libraries

This section focuses on the technical part of NLP applications in pharmacology. In Section 6.1, we cover software libraries and tools that help to build machine learning models for the tasks mentioned in Sections 2 and 3. For each library, we also mention its recorded use in pharmacology. In Section 6.2, we present general text processing libraries. Most covered libraries and tools are accessible as Python packages. Table 7 gives an overview.

[48] `https://bio2rdf.org`
[49] `http://viz.linkeddata.finki.ukim.mk`
[50] `http://drugs.linkeddata.finki.ukim.mk`
[51] `http://godd.finki.ukim.mk`
[52] `https://github.com/typedb-osi/typedb-bio`
[53] `https://dice-research.org/COVID19DS`
[54] `https://github.com/Knowledge-Graph-Hub/kg-covid-19/wiki`

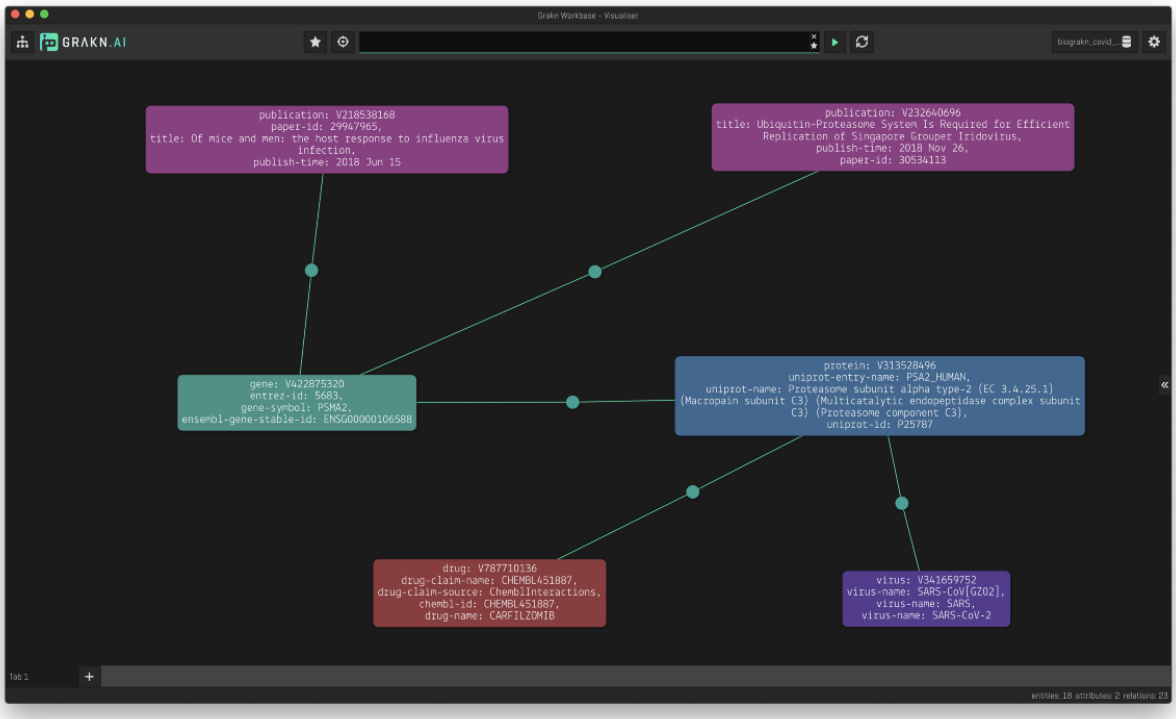

Figure 2: Example usage of the TypeDB Bio (Covid) knowledge graph.

6.1 Machine Learning Libraries

Natural Language Toolkit (NLTK) [17][55] is one of the most powerful and popular NLP libraries. NLTK is a suite of open-source Python modules, data sets, and tutorials on language processing. The toolkit consists of baseline text processing such as sentence splitting, tokenization, and part of speech (POS) tagging. These tools may help in NER, to identify known medications, detect ADEs [35], or in evaluation of entity indicators for relation extraction [167].

MetaMap Transfer (MMTx) [10][56] is an extensively used, Java-based, NER tool that maps biomedical free-form text to UMLS Metathesaurus concepts. In the process of creating the first drug-drug interaction (DDI) corpus that, besides drugs, contains pharmacokinetic DDIs and pharmacodynamic DDIs, the UMLS MetaMap Transfer tool pre-annotates the documents with pharmacological substance entities, i.e., it is used to parse the documents to automatically recognize drug types [75]. MetaMap's intrinsic function - identification of medical concepts - was used for extracting drug indication information from structured product labels [59].

CRFsuite [151][57] implements the Conditional Random Fields machine learning algorithm for labeling sequential data. It is used for NER in the MADEx system for detecting medications and ADEs and their relations from clinical notes [227].

Library for Support Vector Machines (LibSVM) [34][58] is an open-source package that implements the Sequential minimal optimization (SMO) algorithm for kernelized support vector machines (SVMs), supporting both classification and regression. The library was used to classify relation types in the MADEx system [227].

Stanford CoreNLP toolkit [135][59] was initially developed for English, but now supports German, French, Arabic, Chinese and Spanish. The Stanford CoreNLP toolkit is a pipeline of NLP Java tools for linguistic annotations, such as tokenization, sentence splitting, part-of-speech tagging, morphological analysis, NER, syntactic parsing, and coref-

[55] https://www.nltk.org/
[56] https://github.com/theislab/MetaMap
[57] http://www.chokkan.org/software/crfsuite/
[58] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[59] https://stanfordnlp.github.io/CoreNLP/

erence resolution. In pharmacology, CoreNLP was applied in a joint model for entity and relation extraction from biomedical text, providing POS tagging and dependency parsing [117].

BRAT annotation tool [192][60] is an online environment for annotating structured text, i.e. notes in a predefined form. The tool was used to create a corpus from Twitter messages and PubMed sentences to understand drug reports better [8].

SpaCy library [78][61] is a free, open-source library for NLP. It contains ML models for NER, POS tagging, dependency parsing, sentence segmentation, text classification, entity linking, morphological analysis, etc. The library is employed for entity recognition for Pharmaceutical Organizations and Drugs in PharmKE - a text analysis platform focused on the pharmaceutical domain [91].

DOMEO Annotation Toolkit [41][62] (also called SWAN Annotation Tool) is a web application enabling users to manually, semi-automatically, or automatically create ontology-based annotation metadata. DOMEO (Document Metadata Exchange Organizer) can be customized with additional plugins, e.g., for annotation of PDDI mentions in structured product labels [76][63].

Transformers - Hugging Face [217][64] package contains many state-of-the-art NLP models, such as BioBERT [114], RoBERTa [128], CharacterBERT [55], etc. The package offers also tokenizers for several languages and tasks, as well as some popular datasets for NLP tasks such as NER, NLI, QA, etc.

MedCat Tool [107][65] (Medical Concept Annotation Tool) is an open-source tool that uses unsupervised methods for NER and NEL in the biomedical field. The tools were validated with the MIMIC-III program and MedMentions (biomedical papers annotated with mentions from critical care databases). Dobreva et al. [51] highlighted drug entities with the help of this tool in the process of extracting drug-disease relations and drug effectiveness.

AllenNLP [60][66] is an open-soruce research library, built on PyTorch, for developing deep learning models for a wide variety of linguistic tasks. The PharmaKe [91] model uses AllenNLP for NER of drugs and pharmaceutical organizations that appear in texts.

Flair [2][67] is a simple yet powerful framework for NLP, such as NER, POS tagging, and text classification. The framework supports training new models and is used in many research projects and industrial applications, e.g., Sun et al. [195] use FLAIR to find sub-word embeddings.

Gensim [172][68] is a Python library for topic modeling - extraction of unknown topics from a large volume of text (feeds from social media, customer reviews, user feedback, e-mails of complaints, etc.), document indexing, and similarity retrieval from large corpora. The library can handle large text files without having to load the entire file into memory, has efficient multicore implementations of popular algorithms, is platform-independent, and supports distributed computing. Dobreva et al. [50] apply Gensim to NER.

6.2 General NLP libraries

JIEBA tool [196][69] supports Chinese word segmentation based on word frequency statistics with several functions such as POS tagging, TF-IDF weightig and TextRank keyword extraction. It was used to generate POS tags of words [167].

TextBlob [130][70] is a simple Python library, built on top of NLTK and Pattern, that supports complex analysis and operations on text data. The library supports noun phrase extraction, POS tagging, sentiment analysis, classification (Naive Bayes, Decision Tree), tokenization, word and phrase frequencies, parsing, n-grams, word inflection (pluralization and singularization) and lemmatization, spelling correction, etc.

[60] https://brat.nlplab.org/introduction.html
[61] https://spacy.io/
[62] https://github.com/domeo/domeo
[63] https://github.com/rkboyce/DomeoClient
[64] https://huggingface.co/
[65] https://github.com/CogStack/MedCAT
[66] https://allenai.org/demos?o=11
[67] https://github.com/flairNLP/flair
[68] https://radimrehurek.com/gensim/
[69] https://github.com/fxsjy/jieba
[70] https://textblob.readthedocs.io/en/dev/

Polyglot [149][71] is a NLP pipeline that supports multilingual applications and offers a wide range of analyses. It features tokenization (165 languages), language detection (196 languages), NER (40 languages), POS tagging (16 languages), sentiment analysis (136 languages), word embeddings (137 languages), morphological analysis (135 languages), and transliteration (69 languages).

Quepy [9][72] is a Python framework to transform natural language questions to queries in a database query language.

In Table7, we overview the mentioned libraries, together with references from the papers where they are used.

Name	Usage	Referenced Papers
Natural Language Toolkit (NLTK)[17][73]	Tokenization, Lemmatization, POS tagging, NER, Word similarity	[186][104][85][126][5][37][119][35][17][202] [188][162][134][168][173][179][189][169]
MetaMap Transfer tool (MMTx)[10][74]	NER, DD Interaction	[185][10][16][59][66][183][164][105] [225][85][227][155] [102][140][40][89]
CRFsuite library[151][75] [148][127]	NER, Drug Discovery, ADE	[166][35][227][69][13][70][191]
LibSVM[34][76]	Classification, Regression	[227][187] [110][232][82]
Stanford CoreNLP toolkit[135][77]	Tokenization, Lemmatization, POS tagging, NER, Word similarity	[227][205][47][119][117][200] [57][67][105][50][155][91] [44][241]
BRAT[192][78]	Annotating structured text	[229][116]
SpaCy library[78][79]	Tokenization, Lemmatization, POS tagging, NER, Word similarity, SRL	[154][119][91][136][50][128][112][68][37] [80][176][54][201][152][234][87][170]
DOMEO[41][80]	Annotating structured text	[76][25]
Transformers[217][81]	NER, NLI, QA, SRL, Classification, Embeddings	[111][83][222][15][80][55][114][4][32][143] [122][195][3][154][50][178][68][91][158] [27][79][104][7][119][184][136][112][145] [124][155][128][43][5][233][167]
MedCat Tool[107][82]	NER+L	[51][6]
AllenNLP[60][83]	NER,NLI, QA, SRL, Classification, Embeddings	[91][15][205][50][119][154][68][229][123]
Flair[2][84]	NER, POS Tagging, Classification	[195][3][43]
Gensim[172][85]	Text summarization, Embeddings	[50][69] [93][49][240]
JIEBA tool[196][86]	Chinese words: POS tagging, TF-IDF, Text-Rank	[237][228][120][226][113]

[71] https://github.com/aboSamoor/polyglot
[72] https://github.com/machinalis/quepy
[73] https://www.nltk.org/
[74] https://github.com/theislab/MetaMap
[75] http://www.chokkan.org/software/crfsuite/
[76] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[77] https://stanfordnlp.github.io/CoreNLP/
[78] https://brat.nlplab.org/introduction.html
[79] https://spacy.io/
[80] https://github.com/domeo/domeo
[81] https://huggingface.co/
[82] https://github.com/CogStack/MedCAT
[83] https://allenai.org/demos?o=11
[84] https://github.com/flairNLP/flair
[85] https://radimrehurek.com/gensim/
[86] https://github.com/fxsjy/jieba

TextBlob[130][87]	NER, NLI, QA, SRL, Classification, Embeddings	[188][182][174]
Polyglot[149][88]	NER, POS Tagging, Sentiment Analysis, Embedding	[119] [163][33]
Quepy[9][89]	NLP, question transformation to queries	[138]

Table 7: Commonly used machine learning and NLP software libraries and tools.

7 Conclusion

Text is an important source of information in pharmacology. To extract that information from increasingly large collections of structured and unstructured documents, NLP is an essential approach. We present a survey of recent NLP developments relevant to the pharmacological domain.

Our survey comprises five main pillars, each presented in its section: a modern methodology based on pretrained large language models, frequently used tasks, useful datasets, knowledge bases, and software libraries. Each main topic is further split into several components, giving our review a comprehensible hierarchical structure. We compress the main contributions of each section into overview tables at the end of each section. In summary, our survey testifies to swift developments in NLP and a surprising breadth of its use in pharmacology.

While we reviewed over 240 works in our survey, the coverage is by no means exhaustive. In a few years, when next such a survey will be needed, we expect the most exciting developments in the use and integration of multi-modal resources, such as text, images, and 3D structural databases. In artificial intelligence, there is a tendency for large language models, called foundation models [22], to capture as much human knowledge as possible, coupled with the ability for logical and commonsense reasoning. We expect that life sciences and pharmacology will be one of the first areas where domain-specific knowledge will be integrated into such models.

Acknowledgments

This work is also based on COST Action CA18209 – NexusLinguarum "European network for Web-centred linguistic data science", supported by COST (European Cooperation in Science and Technology). The work in this paper was partially financed by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. The work was partially supported by the Slovenian Research Agency (ARRS) core research programme P6-0411 and the young researchers grant.

References

[1] David M Adamson, Stella Chang, and Leigh G Hansen. Health research data for the real world: the marketscan databases. *New York: Thompson Healthcare*, page b28, 2008.

[2] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[3] Liliya Akhtyamova. Named entity recognition in Spanish biomedical literature: Short review and BERT model. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 1–7, 2020. doi: 10.23919/FRUCT48808.2020.9087359.

[4] Jehad Aldahdooh, Markus Vähä-Koskela, Jing Tang, and Ziaurrehman Tanoli. Using BERT to identify drug-target interactions from whole PubMed. bioRxiv, 2021.

[5] Jehad MF Aldahdooh, Ziaurrehman Tanoli, and Jing Tang. R-BERT-CNN: Drug-target interactions extraction from biomedical literature. In *Proceedings of the BioCreative VII Challenge Evaluation Workshop*, 2021.

[6] Anita Alicante, Anna Corazza, Francesco Isgro, and Stefano Silvestri. Unsupervised entity and relation extraction from clinical records in Italian. *Computers in biology and medicine*, 72:263–275, 2016.

[87] `https://textblob.readthedocs.io/en/dev/`

[88] `https://github.com/aboSamoor/polyglot`

[89] `https://github.com/machinalis/quepy`

[7] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019. doi: 10.18653/v1/W19-1909.

[8] Nestor Alvaro, Yusuke Miyao, and Nigel Collier. TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR public health and surveillance*, 3(2):e6396, 2017.

[9] E Andrawos, G García Berrotarán, R Carrascosa, L Alonso i Alemany, and H Durán. Quepy-transform natural language to database queries, 2012.

[10] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[11] Ehsaneddin Asgari and Mohammad RK Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11):e0141287, 2015.

[12] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[13] Michaela Bamburová and Zuzana Neverilová. Structured information extraction from pharmaceutical records. In *RASLAN*, pages 55–62, 2019.

[14] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems. *Journal of Biomedical Informatics*, 41(5): 706–716, 2008.

[15] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: 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)*, pages 3615–3620, 2019. doi: 10.18653/v1/D19-1371.

[16] Asma Ben Abacha and Pierre Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of biomedical semantics*, 2(5):1–11, 2011.

[17] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[18] Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-N protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

[19] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. Linked Data on the Web (LDOW2008). In *Proceedings of the 17th International Conference on World Wide Web*, pages 1265–1266. ACM, 2008.

[20] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[21] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[22] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. ArXiv preprint 2108.07258, 2021.

[23] Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, and William Hamilton. A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. *arXiv preprint arXiv:2102.10062*, 2021.

[24] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*, pages 1–9, 2013.

[25] Richard Boyce, Gregory Gardner, and Henk Harkema. Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, page 206–213, 2012.

[26] John Bradshaw, Brooks Paige, Matt J Kusner, Marwin Segler, and José Miguel Hernández-Lobato. A model to search for synthesizable molecules. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[27] Amy Breden and Lee Moore. Detecting adverse drug reactions from twitter through domain-specific preprocessing and bert ensembling. *arXiv preprint arXiv:2005.06634*, 2020.

[28] Jean-Claude Burgelman, Corina Pascu, Katarzyna Szkuta, Rene Von Schomberg, Athanasios Karalopoulos, Konstantinos Repanas, and Michel Schouppe. Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *Frontiers in Big Data*, 2:43, 2019.

[29] Alison Callahan, Jose Cruz-Toledo, Peter Ansell, and Michel Dumontier. Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. In *Extended Semantic Web Conference*, pages 200–212. Springer, 2013.

[30] Alison Callahan, José Cruz-Toledo, and Michel Dumontier. Ontology-Based Querying with Bio2RDF's Linked Open Data. *Journal of Biomedical Semantics*, 4(1):1–13, 2013.

[31] Kathi Canese and Sarah Weis. PubMed: the bibliographic database. *The NCBI handbook*, 2(1), 2013.

[32] José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained BERT model and evaluation data. In *Proceedings of Practical ML for Developing Countries (PML4DC) at ICLR*, 2020.

[33] W Ceusters and L Bouquet. Language engineering and information mapping in pharmaceutical medicine: Dealing successfully with information overload. *Journal of the Belgian Medical Informatics Association*, 7(1): 26–34, 2000.

[34] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[35] Alec B Chapman, Kelly S Peterson, Patrick R Alba, Scott L DuVall, and Olga V Patterson. Detecting adverse drug events with rapidly trained classification models. *Drug safety*, 42(1):147–156, 2019.

[36] Irene Y Chen, Monica Agrawal, Steven Horng, and David Sontag. Robustly Extracting Medical Knowledge From EHRs: A Case Study of Learning a Health Knowledge Graph. In *Pacific Symposium on Biocomputing 2020*, pages 19–30. World Scientific, 2019.

[37] Long Chen, Yu Gu, Xin Ji, Zhiyong Sun, Haodan Li, Yuan Gao, and Yang Huang. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. *Journal of the American Medical Informatics Association*, 27(1):56–64, 2020.

[38] Qingyu Chen, Alexis Allot, and Zhiyong Lu. LitCovid: an open database of COVID-19 literature. *Nucleic acids research*, 49(D1):D1534–D1540, 2021.

[39] Qingyu Chen, Robert Leaman, Alexis Allot, Ling Luo, Chih-Hsuan Wei, Shankai Yan, and Zhiyong Lu. Artificial intelligence in action: addressing the COVID-19 pandemic with natural language processing. *Annual review of biomedical data science*, 4:313–339, 2021.

[40] Emma Chiaramello, Francesco Pinciroli, Alberico Bonalumi, Angelo Caroli, and Gabriella Tognola. Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *Journal of biomedical informatics*, 63:22–32, 2016.

[41] Paolo Ciccarese, Marco Ocana, and Tim Clark. DOMEO: a web-based tool for semantic annotation of online documents. *Bio-Ontologies 2011*, 2012.

[42] John Coleman and John S Coleman. *Introducing speech and language processing*. Cambridge university press, 2005.

[43] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[44] Alexandre MR Cunha, Kele T Belloze, and Gustavo P Guedes. Recognizing pharmacovigilance named entities in Brazilian Portuguese with CoreNLP. In *Anais do XIII Brazilian e-Science Workshop*, pages 76–79. SBC, 2019.

[45] Spyros N Deftereos, Christos Andronis, Ellen J Friedla, Aris Persidis, and Andreas Persidis. Drug repurposing and adverse event prediction using high-throughput literature analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(3):323–334, 2011.

[46] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772, 2009.

[47] Franck Dernoncourt and Ji Young Lee. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, 2017.

[48] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[49] Anjani Dhrangadhariya, Roger Hilfiker, Roger Schaer, and Henning Müller. Machine learning assisted citation screening for systematic reviews. In *MIE*, pages 302–306, 2020.

[50] Jovana Dobreva, Nasi Jofche, Milos Jovanovik, and Dimitar Trajanov. Improving ner performance by applying text summarization on pharmaceutical articles. In *International Conference on ICT Innovations*, pages 87–97. Springer, 2020.

[51] Jovana Dobreva, Milos Jovanovik, and Dimitar Trajanov. DD-RDL: Drug-Disease Relation Discovery and Labeling. In *International Conference on ICT Innovations*, pages 98–112. Springer, 2022.

[52] Caitlin Dreisbach, Theresa A Koleck, Philip E Bourne, and Suzanne Bakken. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics*, 125:37–46, 2019.

[53] Alexandra Dumitriu, Cliona Molony, and Chathuri Daluwatte. Graph-Based Natural Language Processing for the Pharmaceutical Industry. In Leslie F. Sikos, Oshani W. Seneviratne, and Deborah L. McGuinness, editors, *Provenance in Data Science: From Data Models to Context-Aware Knowledge Graphs*, pages 75–110. Springer International Publishing, Cham, 2021. ISBN 978-3-030-67681-0. doi: 10.1007/978-3-030-67681-0_6. URL https://doi.org/10.1007/978-3-030-67681-0_6.

[54] Sherwyn D'souza, Darlene Nazareth, Cassia Vaz, and Monali Shetty. Blockchain and AI in pharmaceutical supply chain. *Available at SSRN 3852034*, 2021.

[55] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *International Conference on Computational Linguistics*, pages 6903–6915, 2020.

[56] Scott Farrar. The Arizona virtual patient: using question-answering technology to enhance dialogue processing. In *Proceedings of the second international conference on Human Language Technology Research*, pages 222–225, 2002.

[57] Michele Filannino and Özlem Uzuner. Advancing the state of the art in clinical natural language processing through shared tasks. *Yearbook of medical informatics*, 27(01):184–192, 2018.

[58] Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations*, 2021.

[59] Kin Wah Fung, Chiang S Jao, and Dina Demner-Fushman. Extracting drug indication information from structured product labels using natural language processing. *Journal of the American Medical Informatics Association*, 20(3):482–488, 2013.

[60] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*, 2018.

[61] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

[62] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[63] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*. The MIT Press, 2016.

[64] Travis R. Goodwin and Sanda M. Harabagiu. Medical Question Answering for Clinical Decision Support. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 297–306, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983819. URL https://doi.org/10.1145/2983323.2983819.

[65] Vishrawas Gopalakrishnan, Kishlay Jha, Wei Jin, and Aidong Zhang. A survey on literature based discovery approaches in biomedical domain. *Journal of biomedical informatics*, 93:103141, 2019.

[66] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1):496, 2011.

[67] Jinghang Gu, Longhua Qian, and Guodong Zhou. Chemical-induced disease relation extraction with various linguistic features. *Database*, 2016, 2016.

[68] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

[69] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017. doi: 10.1093/bioinformatics/btx228.

[70] Kai Hakala and Sampo Pyysalo. Biomedical named entity recognition with multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, 2019.

[71] Boran Hao, Henghui Zhu, and Ioannis Paschalidis. Enhancing clinical BERT embedding using a biomedical knowledge base. In *Proceedings of the 28th international conference on computational linguistics*, pages 657–661, 2020.

[72] Tom Heath and Christian Bizer. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, 2011.

[73] Sam Henry and Bridget T McInnes. Literature based discovery: models, methods, and trends. *Journal of biomedical informatics*, 74:20–32, 2017.

[74] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, 2020.

[75] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920, 2013.

[76] Harry Hochheiser, Yifan Ning, Andres Hernandez, John R Horn, Rebecca Jacobson, and Richard D Boyce. Using nonexperts for annotating pharmacokinetic drug-drug interaction mentions in product labeling: a feasibility study. *JMIR research protocols*, 5(2):e40, 2016.

[77] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge Graphs. *ACM Comput. Surv.*, 54(4), 2021. doi: 10.1145/3447772.

[78] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python, 2020. URL `https://doi.org/10.5281/zenodo.1212303`.

[79] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[80] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. ArXiv preprint 1904.05342, 2019.

[81] Kexin Huang, Cao Xiao, Trong Hoang, Lucas Glass, and Jimeng Sun. CASTER: Predicting drug interactions with chemical substructure representation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 702–709, 2020.

[82] Ying Huang and Yanda Li. Classifying g-protein coupled receptors with support vector machine. In *International Symposium on Neural Networks*, pages 448–452. Springer, 2004.

[83] Sajid Hussain, Hammad Afzal, Ramsha Saeed, Naima Iltaf, and Mir Yasir Umair. Pharmacovigilance with transformers: A framework to detect adverse drug reactions using BERT fine-tuned with FARM. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

[84] Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). *Drug Safety*, 42(1):99–111, 2019. doi: 10.1007/s40264-018-0762-z.

[85] Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). *Drug safety*, 42(1):99–111, 2019.

[86] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, June 2019. doi: 10.18653/v1/N19-1357. URL `https://aclanthology.org/N19-1357`.

[87] Hyeju Jang, Emily Rempel, Giuseppe Carenini, and Naveed Janjua. Exploratory analysis of COVID-19 related tweets in North America to inform public health institutes. *arXiv preprint arXiv:2007.02452*, 2020.

[88] Kishlay Jha, Yaqing Wang, Guangxu Xun, and Aidong Zhang. Interpretable word embeddings for medical domain. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1061–1066, 2018. doi: 10.1109/ICDM.2018.00135.

[89] Keyuan Jiang and Yujing Zheng. Mining twitter data for potential drug effects. In *International conference on advanced data mining and applications*, pages 434–443. Springer, 2013.

[90] Keyuan Jiang, Tingyu Chen, Liyuan Huang, Ravish Gupta, Ricardo A. Calix, and Gordon R. Bernard. An explainable approach of inferring potential medication effects from social media data. In Mar Marcos, Jose M. Juarez, Richard Lenz, Grzegorz J. Nalepa, Slawomir Nowaczyk, Mor Peleg, Jerzy Stefanowski, and Gregor Stiglic, editors, *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems*, pages 82–92, 2019.

[91] Nasi Jofche, Kostadin Mishev, Riste Stojanov, Milos Jovanovik, and Dimitar Trajanov. PharmKE: Knowledge extraction platform for pharmaceutical texts using transfer learning. *arXiv preprint arXiv:2102.13139*, 2021.

[92] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[93] Chaitali Joshi, Vahida Z Attar, and Shrida P Kalamkar. An unsupervised topic modeling approach for adverse drug reaction extraction and identification from natural language text. In *Advances in Data and Information Sciences*, pages 505–514. Springer, 2022.

[94] Milos Jovanovik and Dimitar Trajanov. Consolidating Drug Data on a Global Scale Using Linked Data. *Journal of Biomedical Semantics*, 8(1):1–24, 2017.

[95] Milos Jovanovik, Bojan Najdenov, and Dimitar Trajanov. Linked Open Drug Data from the Health Insurance Fund of Macedonia. In *10th International Conference for Informatics and Information Technology*, pages 56–61. Faculty of Computer Science & Engineering, Skopje, 2013.

[96] Milos Jovanovik, Aleksandra Bogojeska, Dimitar Trajanov, and Ljupco Kocarev. Inferring Cuisine-Drug Interactions Using the Linked Data Approach. *Scientific Reports*, 5, 2015.

[97] Milos Jovanovik, Bojan Najdenov, Gjorgji Strezoski, and Dimitar Trajanov. Linked Open Data for Medical Institutions and Drug Availability Lists in Macedonia. In *New Trends in Database and Information Systems II*, pages 245–256. Springer International Publishing, 2015.

[98] Jinmyung Jung and Doheon Lee. Inferring disease association using clinical factors in a combinatorial manner and their use in drug repositioning. *Bioinformatics*, 29(16):2017–2023, 2013.

[99] Daniel Jurafsky and James H Martin. *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing, 2nd edition*. Prentice Hall, 2008.

[100] Daniel Jurafsky and James H Martin. Speech and language processingm, 3rd edition draft. Available from: https://web. stanford. edu/˜ jurafsky/slp3, 2022.

[101] Rabiah A Kadir and Behrouz Bokharaeian. Overview of biomedical relations extraction using hybrid rulebased approaches. *Journal of Industrial and Intelligent Information Vol*, 1(3), 2013.

[102] Hennicke Georg KAMP, Tilmann WALK, Gen ISHIKAWA, Niels MOELLER, and Bennard van RAVENZWAAY. The application of metabolomics in vivo for early detection of systemic toxicity in drug safety testing. In *Annual Meeting of the Japanese Society of Toxicology The 40th Annual Meeting of the Japanese Society of Toxicology*, page 150418. The Japanese Society of Toxicology, 2013.

[103] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing*, 20(1):359–392, 1998.

[104] Mayara Khadhraoui, Hatem Bellaaj, Mehdi Ben Ammar, Habib Hamam, and Mohamed Jmaiel. Survey of BERT-base models for scientific text classification: COVID-19 case study. *Applied Sciences*, 12(6):2891, 2022.

[105] Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Dongwook Shin. Broad-coverage biomedical relation extraction with SemRep. *BMC bioinformatics*, 21(1):1–28, 2020.

[106] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[107] Zeljko Kraljevic, Daniel Bean, Aurelie Mascio, Lukasz Roguski, Amos Folarin, Angus Roberts, Rebecca Bendayan, and Richard Dobson. MedCAT–medical concept annotation tool. *arXiv preprint arXiv:1912.10166*, 2019.

[108] Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. ChemProt-3.0: a global chemical biology diseases mapping. *Database*, 2016.

[109] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079, January 2016.

[110] Tannu Kumari, Bhaskar Pant, and KR Pardasani. A SVM model for AAC based classification of class B GPCRs. In *6th World Congress of Biomechanics (WCB 2010). August 1-6, 2010 Singapore*, pages 1607–1610. Springer, 2010.

[111] Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for Russian language. ArXiv preprint arXiv:1905.07213, 2019.

[112] Vivian Lai, Zheng Cai, and Chenhao Tan. Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. 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)*, pages 486–495, November 2019. doi: 10.18653/v1/D19-1046. URL `https://aclanthology.org/D19-1046`.

[113] Wei Lan and Pengcheng Zhang. Research on adaptive learning methods of Chinese medicine based on big data. In *2020 International Conference on Public Health and Data Science (ICPHDS)*, pages 90–93. IEEE, 2020.

[114] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019. doi: 10.1093/bioinformatics/btz682.

[115] Jinhyuk Lee, Sean S Yi, Minbyul Jeong, Mujeen Sung, Wonjin Yoon, Yonghwa Choi, Miyoung Ko, and Jaewoo Kang. Answering questions on COVID-19 in real-time. *arXiv preprint arXiv:2006.15830*, 2020.

[116] BS Levitan, EB Andrews, A Gilsenan, J Ferguson, RA Noel, PM Coplan, and F Mussen. Application of the BRAT framework to case studies: observations and insights. *Clinical Pharmacology & Therapeutics*, 89(2): 217–224, 2011.

[117] Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, 18(1):1–11, 2017.

[118] Fei Li, Weisong Liu, and Hong Yu. Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning. *JMIR medical informatics*, 6(4): e12159, 2018.

[119] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2020.

[120] Mingyue Li, Lixin Du, Jiangying Xu, and Chen Guo. A hypergraph-based method for pharmaceutical data similarity retrieval. In *2021 4th International Conference on Big Data Technologies*, pages 134–140, 2021.

[121] Xuedong Li, Yue Wang, Dongwu Wang, Walter Yuan, Dezhong Peng, and Qiaozhu Mei. Improving Rare Disease Classification Using Imperfect Knowledge Graph. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–2, 2019. doi: 10.1109/ICHI.2019.8904588.

[122] Zhengguang Li, Hongfei Lin, and Wei Zheng. An effective emotional expression and knowledge-enhanced method for detecting adverse drug reactions. *IEEE Access*, 8:87083–87093, 2020.

[123] Zhiheng Li, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. Lexicon knowledge boosted interaction graph network for adverse drug reaction recognition from social media. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2777–2786, 2020.

[124] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, 2021.

[125] Feifan Liu, Abhyuday Jagannatha, and Hong Yu. Towards drug safety surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from electronic health records. *Drug safety*, 42(1):95–97, 2019.

[126] Jing Liu, Rashmie Abeysinghe, Fengbo Zheng, and Licong Cui. Pattern-based extraction of disease drug combination knowledge from biomedical literature. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–7. IEEE, 2019.

[127] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries. *Information*, 6(4):848–865, 2015.

[128] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[129] Ziqing Liu, Enwei Peng, Shixing Yan, Guozheng Li, and Tianyong Hao. T-Know: A Knowledge Graph-Based Question Answering and Information Retrieval System for Traditional Chinese Medicine. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 15–19, 2018.

[130] Steven Loria et al. textblob documentation. *Release 0.15*, 2:269, 2018.

[131] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL `https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf`.

[132] Yuan Luo, William K Thompson, Timothy M Herr, Zexian Zeng, Mark A Berendsen, Siddhartha R Jonnalagadda, Matthew B Carson, and Justin Starren. Natural language processing for EHR-based pharmacovigilance: a structured review. *Drug safety*, 40(11):1075–1089, 2017.

[133] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural NLP: A survey. *CoRR*, abs/2108.04840, 2021. URL `https://arxiv.org/abs/2108.04840`.

[134] Jitendra Mahatpure, Mahesh Motwani, and Piyush Kumar Shukla. An electronic prescription system powered by speech recognition, natural language processing and blockchain technology. *International Journal of Science & Technology Research (IJSTR)*, 8(08):1454–1462, 2019.

[135] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[136] Y Mao and KW Fung. Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts. *Journal of the American Medical Informatics Association: JAMIA*, 27 (10):1538–1546, 2020.

[137] Anca Marginean. GFMed: Question answering over biomedical linked data with grammatical framework. In *CLEF (Working Notes)*, pages 1224–1235, 2014.

[138] Anca Marginean and Oana Marc. Towards querying bioinformatic linked data in natural language. In *2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 23–26. IEEE, 2013.

[139] Matej Martinc, Blaž Škrlj, Sergej Pirkmajer, Nada Lavrač, Bojan Cestnik, Martin Marzidovšek, and Senja Pollak. COVID-19 therapy target discovery with context-aware literature mining. In *International Conference on Discovery Science*, pages 109–123. Springer, 2020.

[140] WB Mattes, HG Kamp, E Fabian, M Herold, G Krennrich, R Looser, W Mellert, A Prokoudine, V Strauss, B van Ravenzwaay, et al. Prediction of clinically relevant safety signals of nephrotoxicity through plasma metabolite profiling. *BioMed research international*, 2013, 2013.

[141] Clara H McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. Effective transfer learning for identifying similar questions: matching user questions to COVID-19 FAQs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3458–3465, 2020.

[142] Zaiqiao Meng, Fangyu Liu, Thomas Clark, Ehsan Shareghi, and Nigel Collier. Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4672–4681, 2021.

[143] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, 2021.

[144] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[145] Milad Moradi and Matthias Samwald. Explaining black-box models for biomedical text classification. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3112–3120, 2021. doi: 10.1109/JBHI.2021.3056748.

[146] Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of biomedical semantics*, 9(1):1–13, 2018.

[147] Patrick Ng. DNA2Vec: Consistent vector representations of variable-length k-mers. *arXiv preprint arXiv:1701.06279*, 2017.

[148] Duy-Hoa Ngo, Alejandro Metke-Jimenez, and Anthony Nguyen. Knowledge-Based Feature Engineering for Detecting Medication and Adverse Drug Events From Electronic Health Records. In *International Workshop on Medication and Adverse Drug Event Detection*, pages 31–38. PMLR, 2018.

[149] Nathaniel Nystrom, Michael R Clarkson, and Andrew C Myers. Polyglot: An extensible compiler framework for Java. In *International Conference on Compiler Construction*, pages 138–152. Springer, 2003.

[150] National Institutes of Health et al. DailyMed database, 2014.

[151] Naoaki Okazaki. CRFsuite: a fast implementation of conditional random fields (CRFs), 2007. URL `http://www.chokkan.org/software/crfsuite/`.

[152] Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Ifeoma Okoh, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, and Vitus Onuigwe. Artificial intelligence for pharmacovigilance in Nigerian social media text. In *AI for Public Health Workshop at ICLR'21*, 2021.

[153] Solip Park, Jae-Seong Yang, Young-Eun Shin, Juyong Park, Sung Key Jang, and Sanguk Kim. Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Molecular systems biology*, 7 (1):494, 2011.

[154] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, 2019.

[155] Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, 8:673, 2020.

[156] Sujan Perera, Amit Sheth, Krishnaprasad Thirunarayan, Suhas Nair, and Neil Shah. Challenges in Understanding Clinical Notes: Why NLP Engines Fall Short and Where Background Knowledge Can Help. In *Proceedings of the 2013 International Workshop on Data Management & Analytics for Healthcare*, pages 21–26, 2013.

[157] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[158] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, 2021.

[159] Bruna GG Pinto, Antonio ER Oliveira, Youvika Singh, Leandro Jimenez, Andre NA Gonçalves, Rodrigo LT Ogava, Rachel Creighton, Jean Pierre Schatzmann Peron, and Helder I Nakaya. ACE2 expression is increased in the lungs of patients with comorbidities associated with severe COVID-19. *The Journal of infectious diseases*, 222(4):556–563, 2020.

[160] Munir Pirmohamed, Alasdair M Breckenridge, Neil R Kitteringham, and B Kevin Park. Adverse drug reactions. *Bmj*, 316(7140):1295–1298, 1998.

[161] Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Charles E. Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10764–10773, 2019. doi: 10.1109/CVPR.2019.01103.

[162] B Prabadevi, Nandyala Srujana Reddy, and B Deepa. Heart rate encapsulation and response tool using sentiment analysis. *International Journal of Electrical and Computer Engineering*, 9(4):2585, 2019.

[163] Srikrishna Prasad and MS Nunifar Sha. NextGen data persistence pattern in healthcare: polyglot persistence. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–8. IEEE, 2013.

[164] Judita Preiss, Mark Stevenson, and Robert Gaizauskas. Exploring relation types for literature-based discovery. *Journal of the American Medical Informatics Association*, 22(5):987–992, 2015.

[165] Protein Data Bank contributors. Protein data bank. *Nature New Biol*, 233:223, 1971.

[166] S Pyysalo, F Ginter, H Moen, T Salakoski, and S Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, pages 39–44, 2013.

[167] Yongbin Qin, Weizhe Yang, Kai Wang, Ruizhang Huang, Feng Tian, Shaolin Ao, and Yanping Chen. Entity relation extraction based on entity indicators. *Symmetry*, 13(4):539, 2021.

[168] Sara Rabhi, Jérémie Jakubowicz, and Marie-Helene Metzger. Deep learning versus conventional machine learning for detection of healthcare-associated infections in French clinical narratives. *Methods of information in medicine*, 58(01):031–041, 2019.

[169] Viju Raghupathi, Yilu Zhou, and Wullianallur Raghupathi. Legal decision support: exploring big data analytics approach to modeling pharma patent validity cases. *IEEE Access*, 6:41518–41528, 2018.

[170] R Ramachandran and K Arutchelvan. Named entity recognition on bio-medical literature documents using hybrid based approach. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–10, 2021.

[171] Justin T. Reese, Deepak Unni, Tiffany J. Callahan, Luca Cappelletti, Vida Ravanmehr, Seth Carbon, Kent A. Shefchek, Benjamin M. Good, James P. Balhoff, Tommaso Fontana, Hannah Blau, Nicolas Matentzoglu, Nomi L. Harris, Monica C. Munoz-Torres, Melissa A. Haendel, Peter N. Robinson, Marcin P. Joachimiak, and Christopher J. Mungall. KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. *Patterns*, 2(1):100155, 2021. doi: https://doi.org/10.1016/j.patter.2020.100155.

[172] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[173] Junqiang Ren. Variability and functions of lexical bundles in research articles of applied linguistics and pharmaceutical sciences. *Journal of English for Academic Purposes*, 50:100968, 2021.

[174] Luiz APA Ribeiro, Daniel Cinalli, and Ana Cristina Bicharra Garcia. Discovering adverse drug reactions from Twitter: A sentiment analysis perspective. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1172–1177. IEEE, 2021.

[175] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1135–1144, 2016. doi: 10.1145/2939672.2939778.

[176] Renzo Rivera and Paloma Martínez. Deep neural model with enhanced embeddings for pharmaceutical and chemical entities recognition in Spanish clinical text. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 38–46, 2019.

[177] Raquel Rodríguez-Pérez and Jürgen Bajorath. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. *Journal of Medicinal Chemistry*, 63 (16):8761–8777, 2020. doi: 10.1021/acs.jmedchem.9b01101.

[178] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349.

[179] Angelo Kenneth S Romasanta, Peter van der Sijde, and Jacqueline van Muijlwijk-Koezen. Innovation in pharmaceutical R&D: mapping the research landscape. *Scientometrics*, 125(3):1801–1832, 2020.

[180] Barbara Rosario and Marti A Hearst. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 430–437, 2004.

[181] Tong Ruan, Yueqi Huang, Xuli Liu, Yuhang Xia, and Ju Gao. QAnalysis: A Question-Answer Driven Analytic Tool on Knowledge Graphs for Leveraging Electronic Medical Records for Clinical Research. *BMC Medical Informatics and Decision Making*, 19(1):1–13, 2019.

[182] Eysha Saad, Sadia Din, Ramish Jamil, Furqan Rustam, Arif Mehmood, Imran Ashraf, and Gyu Sang Choi. Determining the efficiency of drugs under special conditions from users' reviews on healthcare web forums. *IEEE Access*, 9:85721–85737, 2021.

[183] Shengtian Sang, Zhihao Yang, Lei Wang, Xiaoxia Liu, Hongfei Lin, and Jian Wang. SemaTyP: a knowledge graph based literature mining method for drug discovery. *BMC bioinformatics*, 19(1):1–11, 2018.

[184] Alexander Sboev, Anton Selivanov, Ivan Moloshnikov, Roman Rybka, Artem Gryaznov, Sanna Sboeva, and Gleb Rylkov. Extraction of the relations among significant pharmacological entities in Russian-language reviews of internet users on medications. *Big Data and Cognitive Computing*, 6(1), 2022. doi: 10.3390/bdcc6010010.

[185] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

[186] Isabel Segura-Bedmar and Paloma Martínez. Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of biomedical semantics*, 8(1):1–9, 2017.

[187] Zhemin Shan and Hang Song. Research on management decision based on machine learning: Taking the decision of location selection of a pharmaceutical retail enterprise as an example. In *Fuzzy Systems and Data Mining V*, pages 564–574. IOS Press, 2019.

[188] Si Sivasankari, Mu Kavitha, and Gi Saranya. Medical analysis and visualisation of diseases using tweet data. *Research Journal of Pharmacy and Technology*, 10(12):4306–4312, 2017.

[189] Rickard Sjögren, Kjell Stridh, Tomas Skotare, and Johan Trygg. Multivariate patent analysis—using chemometrics to analyze collections of chemical and pharmaceutical patents. *Journal of Chemometrics*, 34(1):e3041, 2020.

[190] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020.

[191] Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336, 2018.

[192] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, 2012.

[193] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665, 2013.

[194] Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung. CAiRE-COVID: a question answering and query-focused multi-document summarization system for COVID-19 scholarly information management. *arXiv preprint arXiv:2005.03975*, 2020.

[195] Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. Deep learning with language models improves named entity recognition for PharmaCoNER. *BMC bioinformatics*, 22(1):1–16, 2021.

[196] J Sun. Jieba Chinese word segmentation tool, 2012.

[197] Silpa Suthram, Joel T Dudley, Annie P Chiang, Rong Chen, Trevor J Hastie, and Atul J Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS computational biology*, 6(2):e1000662, 2010.

[198] Don R Swanson and Neil R Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial intelligence*, 91(2):183–203, 1997.

[199] Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, et al. ChemProt: a disease chemical biology database. *Nucleic acids research*, 39(suppl_1):D367–D372, 2010.

[200] Yixuan Tang, Jisong Yang, Pei San Ang, Sreemanee Raaj Dorajoo, Belinda Foo, Sally Soh, Siew Har Tan, Mun Yee Tham, Qing Ye, Lynette Shek, et al. Detecting adverse drug reactions in discharge summaries of electronic medical records using Readpeer. *International journal of medical informatics*, 128:62–70, 2019.

[201] Amogh Kamat Tarcar, Aashis Tiwari, Vineet Naique Dhaimodker, Penjo Rebelo, Rahul Desai, and Dattaraj Rao. Healthcare NER models using language model pretraining. *arXiv preprint arXiv:1910.11241*, 2019.

[202] Paola Turina, Piero Fariselli, and Emidio Capriotti. ThermoScan: Semi-automatic identification of protein stability data from PubMed. *Frontiers in molecular biosciences*, 8:144, 2021.

[203] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[204] Hadi Veisi and Hamed Fakour Shandi. A Persian medical question answering system. *International Journal on Artificial Intelligence Tools*, 29(06):2050019, 2020.

[205] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

[206] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The COVID-19 open research dataset. *ArXiv*, 2020.

[207] Pengwei Wang, Tianyong Hao, Jun Yan, and Lianwen Jin. Large-scale extraction of drug–disease pairs from the medical literature. *Journal of the Association for Information Science and Technology*, 68(11):2649–2661, 2017.

[208] Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337, 2009.

[209] Xuan Wang, Xiangchen Song, Bangzheng Li, Yingjun Guan, and Jiawei Han. Comprehensive named entity recognition on CORD-19 with distant or weak supervision. *arXiv preprint arXiv:2003.12218*, 2020.

[210] Janus Wawrzinek, Said Ahmad Ratib Hussaini, Oliver Wiehr, José María González Pinto, and Wolf-Tilo Balke. Explainable word-embeddings for medical digital libraries - a context-aware approach. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, page 299–308, 2020. doi: 10.1145/3383583.3398522.

[211] Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, 47(W1):W587–W593, 2019.

[212] Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. What are people asking about COVID-19? a question classification dataset. *arXiv preprint arXiv:2005.12522*, 2020.

[213] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. BioPortal: Enhanced Functionality via New Web Services From the National Center for Biomedical Ontology to Access and Use Ontologies in Software Applications. *Nucleic Acids Research*, 39 (suppl_2):W541–W545, 2011.

[214] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. 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)*, pages 11–20, November 2019. doi: 10.18653/v1/D19-1002.

[215] Wikipedia. *Wikipedia*. PediaPress, 2004.

[216] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

[217] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[218] David Wood, Marsha Zaidman, Luke Ruth, and Michael Hausenblas. *Linked Data*. Manning Publications Co., 2014.

[219] Susmitha Wunnava, Xiao Qin, Tabassum Kakar, Cansu Sen, Elke A Rundensteiner, and Xiangnan Kong. Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. *Drug safety*, 42(1):113–122, 2019.

[220] Eryu Xia, Wen Sun, Jing Mei, Enliang Xu, Ke Wang, and Yong Qin. Mining Disease-Symptom Relation From Massive Biomedical Literature and Its Application in Severe Disease Diagnosis. In *AMIA Annual Symposium Proceedings*, volume 2018, page 1118. American Medical Informatics Association, 2018.

[221] Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. LingYi: Medical Conversational Question Answering System based on Multi-modal Knowledge Graphs, 2022. URL `https://arxiv.org/abs/2204.09220`.

[222] Ying Xiong, Yedan Shen, Yuanhang Huang, Shuai Chen, Buzhou Tang, Xiaolong Wang, Qingcai Chen, Jun Yan, and Yi Zhou. A deep learning-based system for PharmaCoNER. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 33–37, 2019. doi: 10.18653/v1/D19-5706.

[223] Hanqing Xue, Jie Li, Haozhe Xie, and Yadong Wang. Review of drug repositioning approaches and resources. *International journal of biological sciences*, 14(10):1232, 2018.

[224] Hsih-Te Yang, Jiun-Huang Ju, Yue-Ting Wong, Ilya Shmulevich, and Jung-Hsien Chiang. Literature-based discovery of new candidates for drug repurposing. *Briefings in bioinformatics*, 18(3):488–497, 2017.

[225] Hui Yang, Rajesh Swaminathan, Abhishek Sharma, Vilas Ketkar, and Jason D'Silva. Mining biomedical text towards building a quantitative food-disease-gene network. In *Learning structure and schemas from documents*, pages 205–225. Springer, 2011.

[226] Wenxin Yang, Zhiming Zhang, and Yongqiang Gao. Extracting online recruitment information based on BiLSTM-Dropout-CRF model. In *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, pages 1661–1665. IEEE, 2020.

[227] Xi Yang, Jiang Bian, Yan Gong, William R Hogan, and Yonghui Wu. MADEx: a system for detecting medications, adverse drug events, and their relations from clinical notes. *Drug safety*, 42(1):123–133, 2019.

[228] Yang Yang, Qi Li, Zhaoyang Liu, Fang Ye, and Ke Deng. Understanding traditional Chinese medicine via statistical learning of expert-specific Electronic Medical Records. *Quantitative Biology*, 7:210–232, 2019. doi: 10.1007/s40484-019-0173-x.

[229] Yunrong Yang, Zhidong Cao, Pengfei Zhao, Dajun Daniel Zeng, Qingpeng Zhang, and Yin Luo. Extracting impacts of non-pharmacological interventions for COVID-19 from modelling study. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6, 2021.

[230] Mehdi Yazdani-Jahromi, Niloofar Yousefi, Aida Tayebi, Ozlem Ozmen Garibay, Sudipta Seal, Elayaraja Kolanthai, and Craig J. Neal. Interpretable and generalizable attention-based model for predicting drug-target interaction using 3D structure of protein binding sites: SARS-CoV-2 case study and in-lab validation. *bioRxiv*, 2022. doi: 10.1101/2021.12.07.471693. URL `https://www.biorxiv.org/content/early/2022/02/18/2021.12.07.471693`.

[231] SriJyothsna Yeleswarapu, Aditya Rao, Thomas Joseph, Vangala Govindakrishnan Saipradeep, and Rajgopal Srinivasan. A Pipeline to Extract Drug-Adverse Event Pairs From Multiple Data Sources. *BMC Medical Informatics and Decision Making*, 14(1):1–16, 2014.

[232] Farzana Yesmin. *Identification of Pharmaceutical Substances With Raman Spectroscopy*. PhD thesis, Ruhr Universität Bochum Germany, 2016.

[233] Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, page 103983, 2022.

[234] Jia Zeng, Christian X Cruz-Pico, Turçin Saridogan, Md Abu Shufean, Michael Kahle, Dong Yang, Kenna Shaw, and Funda Meric-Bernstam. Natural language processing–assisted literature retrieval and analysis for combination therapy in cancer. *JCO Clinical Cancer Informatics*, 6:e2100109, 2022.

[235] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, 6(1):1–9, 2019.

[236] Wen-Ming Zhao, Shu-Hui Song, Mei-Li Chen, Dong Zou, Li-Na Ma, Ying-Ke Ma, Ru-Jiao Li, Li-Li Hao, Cui-Ping Li, Dong-Mei Tian, et al. The 2019 novel coronavirus resource. *Yi chuan= Hereditas*, 42(2):212–221, 2020.

[237] Zufeng Zhong. Internet public opinion evolution in the COVID-19 event and coping strategies. *Disaster Medicine and Public Health Preparedness*, 15(6):e27–e33, 2021.

[238] Renyi Zhou, Zhangli Lu, Huimin Luo, Ju Xiang, Min Zeng, and Min Li. NEDD: a network embedding based method for predicting drug-disease associations. *BMC bioinformatics*, 21(13):1–12, 2020.

[239] Zhenpeng Zhou, Xiaocheng Li, and Richard N Zare. Optimizing chemical reactions with deep reinforcement learning. *ACS central science*, 3(12):1337–1344, 2017.

[240] Hongting Zhu, Ashwin Pothukuchi, and Joel Guo. Doc2Vec on similar document suggestion for pharmaceutical collections. Technical report, College of Engineering, University of Michigan, USA, 2020.

[241] Anastazia Žunić, Padraig Corcoran, and Irena Spasić. Improving the performance of sentiment analysis in health and wellbeing using domain knowledge. In *Healthcare Text Analytics Conference - HealTAC 2020, London, UK*, 2020.