OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–20
https://doi.org/10.1093/bib/bbac543
Review

A comprehensive review on knowledge graphs for complex diseases

Yang Yang †, Yuwei Lu† and Wenying Yan

Corresponding author. Wenying Yan, Department of Bioinformatics, School of Biology and Basic Medical Sciences, Medical College of Soochow University, and Center for Systems Biology, Soochow University, Suzhou 215123, China. Tel.: +86-138-1484-2809. E-mail: wyyan@suda.edu.cn

†Yang Yang and Yuwei Lu contributed equally to this work.

Abstract

In recent years, knowledge graphs (KGs) have gained a great deal of popularity as a tool for storing relationships between entities and for performing higher level reasoning. KGs in biomedicine and clinical practice aim to provide an elegant solution for diagnosing and treating complex diseases more efficiently and flexibly. Here, we provide a systematic review to characterize the state-of-the-art of KGs in the area of complex disease research. We cover the following topics: (1) knowledge sources, (2) entity extraction methods, (3) relation extraction methods and (4) the application of KGs in complex diseases. As a result, we offer a complete picture of the domain. Finally, we discuss the challenges in the field by identifying gaps and opportunities for further research and propose potential research directions of KGs for complex disease diagnosis and treatment.

Keywords: biomedicine, complex disease, entity extraction, knowledge graphs, relation extraction

Introduction

Complex diseases, such as cancer and neurodegenerative diseases, are caused by the interaction of multiple genes and environmental factors with high morbidity and mortality worldwide. According to the latest global cancer burden data released by the International Agency for Research on Cancer, there were about 19.29 million new cancer cases and 9.96 million cancer deaths worldwide in 2020 [1]. With the rapid and stable development of modern medicine and informatics, extensive information on complex diseases has been obtained and recorded in the form of biomedical literature, electronic medical records (EMRs) and biomedical databases. Extracting and integrating this information from the above media to facilitate clinical diagnosis and therapies have become significant challenges.

Recently, knowledge graph (KG) technologies have emerged as promising strategies to overcome these challenges by understanding the interconnections among the biomedical terms. A KG consists of nodes and edges, where nodes represent entities or concepts and edges are used to connect two nodes and represent the relations between entities or the attributes of entities. As the ensemble of human knowledge, KGs have attracted increasing research attention [2]. At present, KGs have greatly benefitted various biomedical studies, such as medical question-answering systems [3, 4], retrieval systems [5, 6], data analysis systems [7] and drug repurposing [8, 9]. In particular, complex disease knowledge graphs (CDKGs) have been constructed and developed. For example, Li *et al.* [10] constructed a comprehensive KG of hepatocellular carcinoma, including drugs, diseases, proteins, DNA and other entities. Xiu *et al.* [11] constructed a knowledge graph of digestive system tumors. CDKGs have shown potential strong capacity to provide applications that are more efficient in complex disease theoretical research and clinical practice. In a recent related review, David *et al.* [12] made a review on the construction and application of knowledge graphs in the general biomedical field, but without detail information on how to build the knowledge graph. In addition, its focus is on the biomedical field, and it does not specifically focus on the disease field. Abu-Salih *et al.* [13] provide a detailed overview of multiple domain-specific knowledge graphs, such as Education, Healthcare, Finance and Healthcare. Alshahrani *et al.* [14] only provide a detailed overview of knowledge embedding methods in the biomedical domain.

Compared with traditional knowledge graphs in the general field, CDKG has higher requirements on the quality of data content, and it is more difficult to obtain, preprocess and integrate biomedical data. In addition, it is of great significance and value to develop medical applications based on the constructed CDKG. In what follows, we offer a comprehensive overview of the construction and application of CDKGs. First, two strategies for constructing a KG are introduced, namely, constructing a KG by extracting knowledge from text and constructing a KG by merging databases. The process of building a KG from scratch is outlined in detail, and this process is summarized into seven steps. Second, we provide an overview of the knowledge sources for CDKGs and

Yang Yang is an associate professor with School of Computer Science & Technology, Soochow University, China. His research interests include biomedicine data analysis and machine learning.
Yuwei Lu is currently pursuing the master's degree with the School of Computer Science & Technology, Soochow University, Suzhou, China. His research interest is biomedical natural language processing.
Wenying Yan is currently an associate professor with the Department of Bioinformatics, School of Biology and Basic Medical Sciences, Medical College of Soochow University, and Center for Systems Biology, Soochow University, China. Her research interests include bioinformatics, biological network analysis and machine learning.
Received: August 22, 2022. **Revised:** November 2, 2022. **Accepted:** November 10, 2022

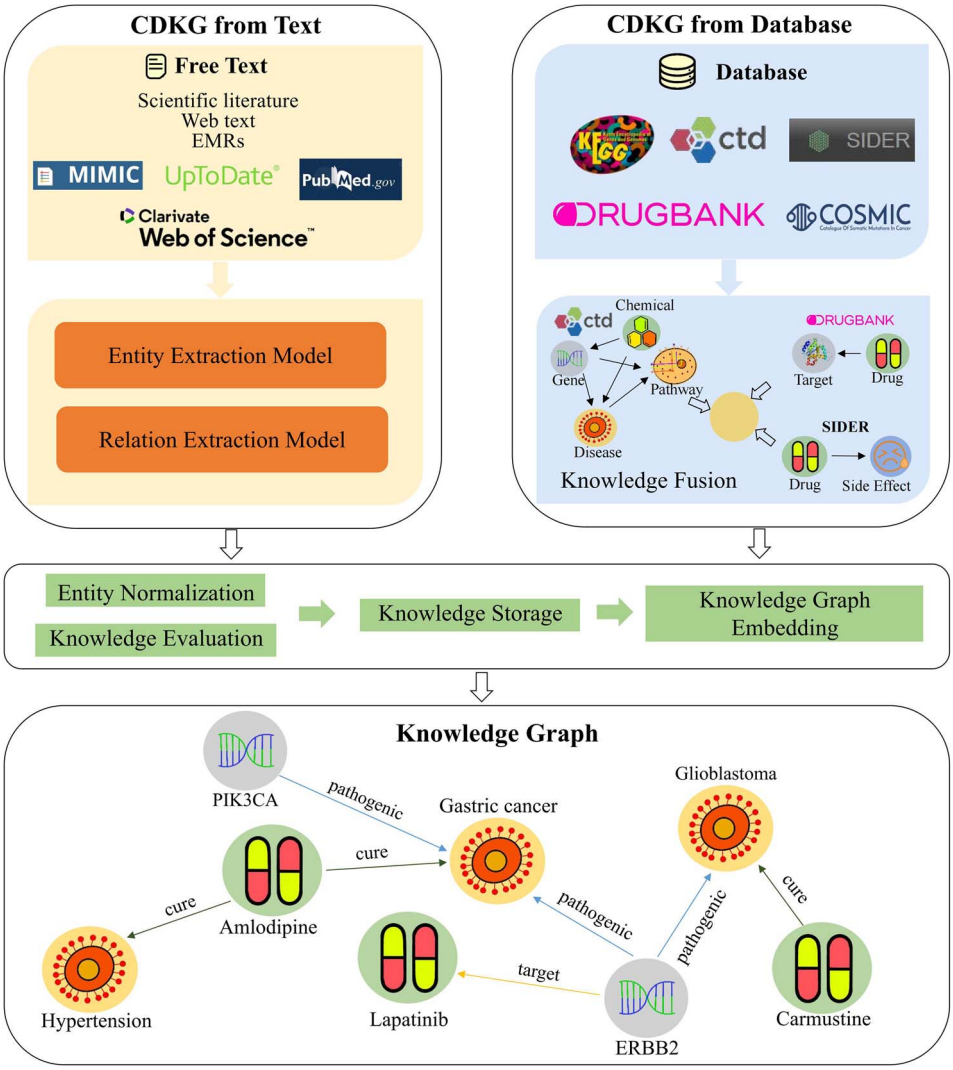

Figure 1. Pipeline of knowledge graph construction from text and databases.

introduces open shared information extraction tasks. Third, the key technologies for constructing KGs including entity extraction and relationship extraction were summarized. Four, details the applications for KGs in regarding complex diseases. At last, we discuss current challenges and opportunities.

Construction process of complex disease knowledge graphs

Since there are a variety of ways to construct KGs, we divide them into two categories according to knowledge source: (1) constructing a KG by extracting knowledge from text and (2) constructing a KG by merging databases. Figure 1 shows the pipelines of construction of CDKG from text and the construction of CDKG from databases. As shown in the figure, the main difference between the two categories is knowledge extraction with different emphases: the former construction pipeline focused on the entities and relation extraction, while the later one pay more attention on knowledge fusion of interactions from different databases.

Constructing a KG by extracting knowledge from text means that the data used for KG construction are extracted from medical text and databases, and the main source is medical text. A database usually stores structured data, which usually can be directly used. By contrast, text is difficult to use directly, so information extraction technologies are employed to build structured data from text. This is the most common strategy for KG construction, and the difficulty lies in extracting the target knowledge from the text. Table 1 summarizes the methods for constructing CDKGs by extracting knowledge from text.

Constructing a KG by merging databases means that data from multiple databases or knowledge bases are merged to construct a large comprehensive knowledge graph. The difficulty is eliminating the ambiguity between data from different sources for subsequent knowledge integration.

Constructing a knowledge graph by extracting knowledge from text

In this review, we summary CDKG construction into seven steps: (1) preprocessing, (2) data schema design, (3) knowledge extraction, (4) entity normalization, (5) knowledge evaluation, (6) knowledge storage and (7) knowledge embedding. Figure 2 shows the process of constructing a knowledge graph by extracting knowledge from text.

Preprocessing

The first step in constructing a CDKG is preprocessing, which usually consists of two parts: selecting knowledge sources and preprocessing the data. The knowledge sources should be authoritative to ensure the correctness of knowledge. Frequently used

Table 1. Complex disease related KGs that were constructed by extracting knowledge from text

Name	Description	Knowledge sources	Entity extraction	Relation extraction	Entity normalization	Knowledge evaluation
KnowLife	Health and life sciences KG	Structured data: UMLS Unstructured data: 1 451 299 biomedical literatures	UMLS 7 entity types	Pattern-based methods 13 relation types	–	Manual sampling
HKG	Disease-symptom KG	Structured data: ICD-9 Unstructured data: 273 174 EMRs	UMLS and GHKG 2 entity types	Logistic Regression Naive Bayes Noisy OR 1 relation types	–	Manual sampling Compared with Google health KG
ATOM	Anti-tumor biomaterial KG	Unstructured data: 100 biomedical literature abstracts from Web of Science	Pubtator Stanford CoreNLP 6 entity types	Stanford NLP –	–	–
KGHC	Hepatocellular carcinoma KG	Structured data: SemMedDB Unstructured data: PubMed literature and web text	Attention-BiLSTM+CRF SemRep 9 entity types	SemRep 22 relation types	JaccardSimilarity	Manual sampling
DSTKG	Digestive system tumor KG	Unstructured data: 731 EMRs of digestive tumors	BilLSTM+CRF 7 entity types	BiGRU+attention 16 relation types	Dictionary-based	Manual sampling
TBKG	Tumor-biomarker KG	Structured data: UMLS Unstructured data: biomedical literature from MEDLINE	cTAKES 4 entity types	Naive Bayes methods 6 relation types	–	Sampling match
MKG	Clinical KG	Structured data: 16 217 270 EMRs	BiLSTM-CRF dictionary-based 9 entity types	Rule-based methods 9 relation types	Dictionary-based	Manual sampling
DSKG	Autism spectrum disorder KG	Unstructured data: 24 687 autism spectrum disorder-related article abstracts from PubMed	UMLS dictionary-based methods Not mentioned	Kmeans++ Manually annotated 6 relation types	MeSH	Compared with SemRep
PKG	Pubmed KG	Structured data: Authority, Semantic Scholar, NIH ExPORTER, ORCID, MapAffil 2016, Affiliation Parser Library Unstructured data: PubMed literature	BioBERT 12 entity types	–	Dictionary-based	–
DRKF	Drug repurposing KG of Parkinson's disease	Structured data: DrugBank,PharmGKB, KEGG, DRUG TTD, DID, SIDER Unstructured data: 54 100 biomedical literatures from PubMed	SemRep 5 entity types	SemRep 6 relation types	UMLS	Downstream mission
GNBR	Large and heterogeneous KG comprising drug, disease and gene	Unstructured data: ∼28.6 million PubMed abstracts	PubTator 3 entity types	Ensemble biclustering algorithm (EBC) with hierarchical clustering 4 relation types	–	Compared with biomedical database

knowledge sources for CDKG construction are introduced in later section.

The knowledge source is mainly text data. However, the original text always contains noise, and preprocessing the text will make the information extraction model more efficient at extracting structured data. Text preprocessing includes word segmentation, sentence splitting, parts-of-speech (POS) tagging, dependency parsing, etc., which can provide grammatical and semantic features for the information extraction model, thus helping to improve model performance. For example, Wang *et al.* [15] employed Stanford CoreNLP [16] for word segmentation, sentence segmentation and POS tagging. Xiu *et al.* [11] performed word segmentation and POS tagging on text sequences using ICTCLAS tools. Rossanez *et al.* [17] simplified the sentence structure by performing dependency parsing on the text.

Data schema design

The second step is designing the data schema. The data schema is the core of the knowledge graph and used to determine and standardize the entity type and relation type of the knowledge graph, such as to stipulate certain types of relations that only appear between specific entity pair types.

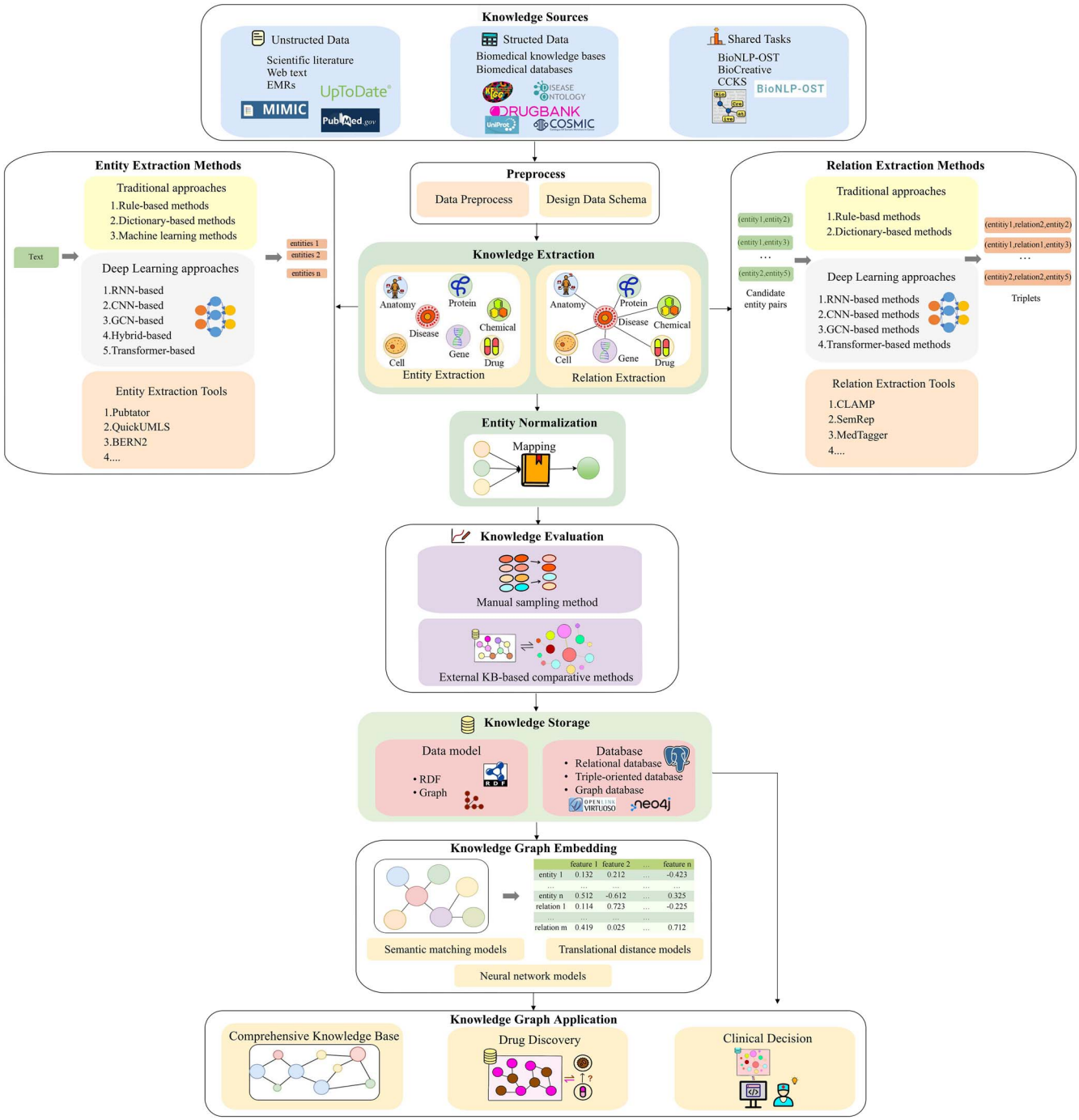

Figure 2. Knowledge graph construction and application from text.

The data schema usually needs to be designed according to the content and scale of the knowledge graph. For example, Rotmensch *et al.* [18] constructed a health knowledge graph (HKG) to model the relation between diseases and symptoms. Therefore, the entity types were set as disease and symptoms, and there was only one relation type, indicating the correlation between disease entities and symptom entities. But in a comprehensive clinical knowledge graph (CKG), which was constructed by merging multiple databases, there were 33 types of entities and 51 types of relations [19].

Knowledge extraction

The third step in constructing a CDKG is knowledge extraction, which is the core step in the process. The goal is to extract correct knowledge from unstructured data in a variety of sources. In this review, we divide knowledge extraction into two tasks: entity extraction and relation extraction. The former task extracts all target entities from the data source, and the relation extraction task extracts knowledge triples from the source. Figure 3A shows the process of knowledge extraction.

Entity normalization

The next step of knowledge extraction is entity normalization. In the field of biomedicine, the same entity usually has more than one name. For example, Alzheimer's disease is also known as senile dementia. Moreover, constructing a CDKG often involves extracting knowledge from multiple different sources, where different standard names are used for the same entity. Therefore,

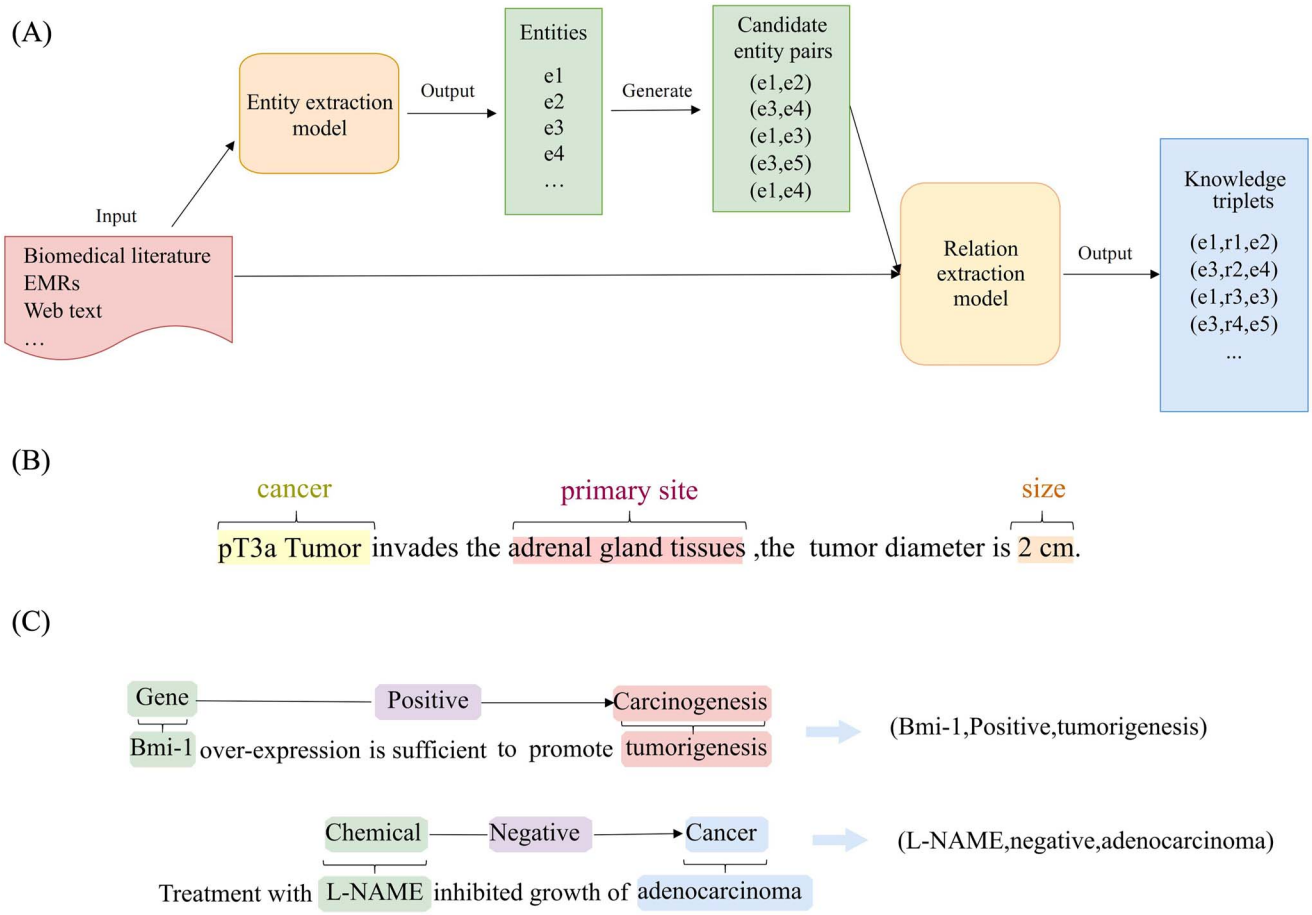

Figure 3. Knowledge extraction. (**A**) Knowledge extraction process. (**B**) An example of an entity extraction task. (**C**) An example of extracting knowledge triples from literature.

it is necessary to carry out entity normalization to reduce the redundancy and ambiguity of the knowledge in the KG.

Dictionary-based approaches are often used for entity normalization in the field of biomedicine. In this approach, the standard name of the entity is set, and the other names of the entity are set as synonyms. Then the original terms of the entity in the text are mapped to the standard name in the dictionary.

To construct a drug repurposing for Parkinson's disease, Zhang *et al.* [20] used the concept identifier in the Unified Medical Language System (UMLS) as the standard name for the entity of the knowledge graph to unify the different names from the medical knowledge base (KB) and from biomedical literature. Yuan *et al.* [21] performed entity extraction and entity normalization at the same time. In entity normalization, an unambiguous biomedical dictionary was constructed based on medical subject headings (MeSH) and the UMLS, and then dictionary-based approaches and heuristic rules were employed to eliminate entity ambiguity. The heuristic rules included singular/plural forms, linguistic semantic patterns and co-occurring semantic types. The above studies used a unified normalization model to normalize multiple entity types. However, the normalization model depends on the entity type. Thus, a more reasonable option is to adopt distinct entity normalization methods for different entity types. For example, there are five types of biomedical entities in PKG [22], and a distinct entity normalization model is used for each entity type (e.g. tmChem [23] is used to normalize chemical type entities). Pubtator [24] is another commonly used as an ad hoc controlled vocabulary for entity extraction and entity normalization in building KG [25], and was updated and extended as PubTator Central (PTC) [26].

Recently, deep learning-based approaches are also in continuous development and have become the main approach to entity normalization. The neural network has been used to calculate the entity mention representation and the concept representation. Ji *et al.* [27] used BERT to obtain entity representations containing contextual information and regarded the entity normalization task as a binary classification task. In addition to using BioBERT to obtain the dense representation, Sung *et al.* [28] proposed the BioSyn model, which employ Term frequency-Inverse Document Frequence (TF-IDF) to calculate the sparse representation, and then the weighted sum of the two to obtain the entity representation. Liu *et al.* [29] proposed a pre-trained language model SapBERT, so that entity representations contain not only contextual semantic features, but also synonym semantic features, and achieves very high performance on multiple datasets.

Knowledge evaluation

Knowledge evaluation is a relatively difficult task, since there is no unified method, and an appropriate evaluation method should be preferred according to the specific situation. In order to measure the content quality of CDKG, accuracy is often used as an evaluation metric [30].

In this review, knowledge evaluation methods are classified as: manual sampling evaluation method and external KB-based comparative evaluation methods.

In manual sampling evaluation methods, samples are randomly selected from the KG, and then the accuracy of the samples is manually evaluated. This method is simple, easy and precision, although it is costly. Since CDKGs always require highly accurate knowledge, this method is often preferred. For example, Knowlife [30] used four people to evaluate the knowledge in the KG, with a voting method to determine accuracy of each piece of knowledge. Knowlife has 542 689 pieces of knowledge with the average accuracy of 93.3%. KGHC [31] randomly sampled 1000 triples from a hepatocellular carcinoma KG, and manually calculated the accuracy rate as 81.2%.

In external KB-based comparative evaluation methods, a high-quality external KB that has a high degree of overlap with the target KG is selected and then automatically compared with the KG. The advantage of this approach is that it can automatically, efficiently and accurately complete knowledge evaluation. However, the disadvantage is that the external KB and the target knowledge base cannot be completely consistent. Inconsistent ways of knowledge expression and inconsistent knowledge coverage lead to bias when evaluating of the target KG. For example, Lu *et al*. [18] used the Google Health Knowledge Graph (GHKG) as an external knowledge base. Since knowledge in the GHKG does not cover all aspects, some knowledge will be marked as false positives when evaluating the target KG according to the GHKG. The knowledge in the target KG is not necessarily wrong. Rather, it may be only due to the absence of this knowledge in the GHKG.

Knowledge storage

Knowledge in the KG is usually expressed in the form of triples. It is important to select an appropriate database to store triples, which will affect applications of KGs. Types of databases, it can be divided into (1) relational databases, (2) triple-oriented databases and (3) graph databases.

Relational databases are mature and offer a complete theoretical and practical system. Before storing the triples in a relational database, a large number of triples need to be reorganized into two-dimensional tables. The triple table is the simplest and most direct way to store KGs in a relational database. Besides the triple table, there are various other organization methods, such as, property table, vertical partitioning etc. However, when querying complex knowledge from the KG, multiple tables must be connected to search for target information, and the query cost is relatively high. Yu *et al*. [32] used a MySQL database to store triples to construct a medicine KB for interpreting the relationships between diseases, genes, variants and drugs.

A triple-oriented database is specially developed for storing large-scale triple data and managing the SPARQL language. Virtuoso, a triple-oriented database, has been used to store CDKGs [7, 33]. Hasan *et al*. [7] compared the query speed of the relational database PostgreSQL and the triple database Virtuoso. The results showed that Virtuoso is 76% faster than PostgreSQL when querying the sequences from breast cancer patients with different treatment types.

Graph database are another way to store knowledge graphs using a graph structure. Neo4j, an open-source NoSQL graph database based on Java, is the most widely used graph databases in biomedicine [10, 11, 15, 19, 34, 35]. Besides Neo4j, the graph database OrientDB is also used to store CDKGs [36].

Knowledge embedding

By using knowledge embedding, knowledge graphs can be applied to many more tasks. Knowledge embedding projects the knowledge in the KG to a low-dimensional vector space. As such, vectors represent the KG. Knowledge embedding approaches are divided into three categories: semantic matching models, translational distance models and neural network models [20].

Semantic matching models use a similarity-based scoring function to project the entities and relations of the fact triples in the KG to the vector space. Then the similarity between them is calculated. For example, RESCAL [37] uses a vector to represent the underlying semantic information of an entity, and a matrix is used to represent the semantic information of a relation. The relation matrix is multiplied by the entity vector to obtain the score of the fact triple.

Translation distance models use a distance-based scoring function inspired by the word vector method. For example, TransE [38] assumes that a triplet (e1,r,e2) fits to the relation of $\mathrm{vec(e1)} + \mathrm{vec(r)} \approx \mathrm{vec(e2)}$, and the correctness of the fact triplet is represented by the distance between the vectors. Subsequent models such as TransH [39] and TransR [40] have improved the shortcomings of TransE.

Neural network models are the current mainstream approach to knowledge embedding. They can learn more knowledge information from KGs. Based on convolutional neural networks, ConvE [41] captures latent semantic features between entity vectors and relation vectors through convolution operations. In addition, there are other methods, such as ConvTransE [42], which is based on a CNN and RSN [43] that is a recurrent neural network (RNN). With the widespread use of pre-trained language models (PLMs), semantic features are used to predict whether triples are effective [44,45]. Li *et al*. [46] built a PLM dedicated to knowledge embedding, which achieved good results in multiple experiments.

Constructing a knowledge graph by merging databases

This way involves merging the information from multiple existing databases to construct a comprehensive KG, whose construction process consists mainly of database selection and knowledge fusion. Knowledge fusion is an important and challenging step, because the data comes from multiple databases, and each database may not use the same standard name for the same entity. The method of knowledge fusion is the same as that introduced in the previous section, and mainly using rule-based or dictionary-based methods. Furthermore, entity names in medical databases are more canonical than those in medical texts, thus this task is relatively easy.

PreMedKB [32] is a comprehensive and precise medicine knowledge base for interpreting the relationships between diseases, genes, variants and drugs. It integrates more than 20 public databases, which contain information about the relations between diseases, genes, mutations and drugs. In PreMedKB, there is a corresponding meta-database for each entity type, to store the standard names and synonyms of entities. To ensure that the meta-database contains more complete vocabulary information, the content of the meta-database often comes from multiple databases. For example, for gene meta-databases, vocabulary information is obtained from HGNC, UniProtKB, TCGA and other databases.

Su *et al*., combined 18 public biomedical databases to form a biomedical knowledge base called iBKH [41]. First, a database was selected to initialize the standard dictionary. Then a linkage pool was constructed to map the entities of other databases to the standard dictionary, and new entity name will be added to update the content of the standard dictionary. For example, for a drug type entity, DrugBank [2] is used to initialize the standard

dictionary and the linkage pool includes MeSH terms, MeSH term IDs and UMLS Identifiers.

Knowledge sources

The knowledge sources for CDKG construction are mainly grouped into four categories: (1) EMRs; (2) biomedical literature, (3) public biomedical databases and (4) public datasets.

Electronic medical records

EMRs provide rich medical resources regarding patients, including various examinations, hospitalized medical records, admission records, course records, and pathology reports. EMRs mainly record data in text form, as a more natural and vivid way to promote communication between medical staffs. But it is difficult to use text data directly in computers, so it is necessary to convert text data into structured data. The following public databases store clinical data information from EMRs.

The Surveillance, Epidemiology and End Results program (SEER, https://seer.cancer.gov/) collects pathology reports of cancer patients, and organizes and analyzes the content for cancer treatment and research. Therefore, the cancer pathology report in SEER can be used as for cancer information extraction tasks.

MIMIC [47] is a publicly available database developed by the Computational Physiology Laboratory at the Massachusetts Institute of Technology. It also contains comprehensive information for each hospitalized patient: laboratory measurements, drug management, vital signs records, etc. To date, fourth versions of the MIMIC has been released.

Biomedical literature

Biomedical literature is an important source for CDKGs, and the latest research findings are usually published in the form of medical literature. The knowledge in biomedical literature can be used to update and supplement databases, thereby promoting the application of the latest research. Biomedical literature is the main source of knowledge when constructing CDKGs [10, 15, 20–22, 30, 48]. PubMed is a mainstream literature database in biomedicine, especially for literature on complex diseases including cancer. Cancer-related literatures on PubMed accounts for about 13% of the total.

Biomedical databases

Biomedical databases are generally constructed by experts manually extracting data from literatures, textbooks and other text data. This helps to ensure the accuracy and reliability of the content in the database. Biomedical databases can help with the construction of CDKGs and information extraction in two ways. First, the structured data in the database can be directly used. Subsequently, the content of the database can be used to construct a dictionary for entity identification and normalization. Table 2 summarizes commonly used databases in CDKGs. Among them, UMLS is the most frequently used and the most often-referenced resource for entity extraction [21] and normalization [11] in CDKG construction. It is composed of three parts: a Metathesaurus, a semantic network, and a specialist lexicon [49]. The Metathesaurus is the largest collection of biomedical dictionaries, containing 2.9 million entities and 11.4 million entity names and synonyms. The semantic network is a collection of the relations of all concepts in Metathesaurus, representing various semantic relations between biomedical concepts in UMLS.

Public datasets

The task of knowledge extraction is inseparable from the need for high-quality annotated data. More and more research groups have published related datasets through conferences, such as Critical Assessment of Information Extraction systems in Biology (BioCreative), National NLP Clinical Challenges (N2C2, formerly known as i2b2 NLP shared task), BioNLP-OST, China Conference on Knowledge Graph and Semantic Computing (CCKS) and International Workshop on Semantic Evaluation (SemEval).

BioCreative focused on the development of information extraction in biomedical domains and has published several datasets, such as CHEMDNER [50] from BioCreative IV track 2. The goal of CHEMDNER is to improve a performance of chemical compound and drug mention extraction model from text. The BioNLP Shared Task (BioNLP-ST) series aims to develop and share computational tasks in biomedical text mining. In 2013, BioNLP proposed the Cancer Genetics shared task, which aims to advance the development of such event extraction methods and the capacity for automatic analysis of texts on cancer biology [51].

The N2C2, SemEval and CCKS series workshops have helped to create high-quality annotated datasets for clinical data processing. The 2018 N2C2 shared task involved extracting adverse drug events (ADEs) and medication from electronic health records. From 2015 to 2017, SemEval conducted tasks to extract temporal information from clinical notes and pathology reports of brain cancer and colon cancer patients at the Mayo Clinic [52–54]. CCKS pursued a Chinese EMR entity recognition task from 2017 to 2020 [3, 55–57].

Knowledge extraction methods

Knowledge extraction is the most significant step in constructing a CDKG. This step consists of entity and relation extractions. This section provides an overview of entity and relation extraction including the content of the task, solutions and evaluation metrics.

Entity extraction

Entity extraction aims to extract target entity information from text and is the basic task in knowledge extraction. For entity extraction, the entity boundary and entity type must be correctly identified from the text. An example of an entity extraction task is shown in Figure 3B. In this review, the entity extraction task takes cancer as a research topic, since cancer is a representative disease among complex diseases.

Entity extraction tasks in clinical and biomedical fields are more challenging than in other fields for the following reasons. Firstly, there is no standard entity naming nomenclature and different researchers may adopt different nomenclature. Secondly, abbreviations in biomedical documents are pervasive. Thirdly, many biomedical entity names are descriptive and as such, entity names are often too long [58]. Thus, numerous methods have been proposed to try to overcome these challenges. In this review, we focus on the methods of entity extraction for cancer studies. They are roughly classified as traditional methods and deep learning-based methods. Table 3 summarizes the methods of extracting cancer-related entities from EMRs.

Traditional methods

Traditional entity extraction methods include dictionary-based, rule-based and machine learning-based methods.

Dictionary-based methods were the primary solutions for entity extraction in the early days of research and remain effective

Table 2. The commonly used databases in the CDKG

Name	Create (Update)	Description	Statistics	URL
KEGG	1995 (2022)	Resource for understanding high-level functions and utilities of the biological system from molecular-level information	551 pathways, 40 539 572 genes, 18 907 compounds, 1408 Disease-related network elements, 2551 diseases, 11 873 drugs	https://www.kegg.jp/
ORPHANET	1997 (2021)	Information on rare diseases	6172 diseases, 5835 genes, 45 734 diagnostic tests	https://www.orpha.net/consor/cgi-bin/index.php
GO	1999 (2022)	The world's largest source of information on the functions of genes	43 917 gene ontological terms, 1 553 323 gene products from 4990 species and 7 898 497 gene annotations	http://geneontology.org/
STRING	2000 (2021)	A database of known and predicted protein–protein interactions	14 094 species, 67 592 464 proteins and more than 20 052 394 041 interactions were included.	https://string-db.org/
PharmGKB	2001 (2021)	A pharmacogenomics knowledge resource	784 Drug Label Annotations,165 Clinical Guideline Annotations,153 Curated Pathways, 715 Annotated Drugs	https://www.pharmgkb.org/
CDT	2004 (2022)	A digital ecosystem that relates toxicological information for chemicals, genes, phenotypes and diseases.	11 interaction types, 53 744 genes, 7270 diseases, 17 045 chemicals	http://ctdbase.org/
UniProt	2002 (2022)	A resource of protein sequence and functional information	565 254 manually annotated and reviewed data, 219 174 961 computationally analyzed data.	https://www.uniprot.org/
DO	2003 (2022)	A standardized ontology for human disease providing human disease terms, phenotype characteristics and related medical vocabulary disease concepts	10 862 standard disease terms	https://disease-ontology.org/
PubChem	2004 (2022)	A collection of chemical information including chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations and more.	110 758 106 compounds, 279 531 956 substances, 294 028 347 bioactivities, 33 864 958 literature, 41 796 860 Patents	https://pubchem.ncbi.nlm.nih.gov/
UMLS	2004 (2021)	A set of files and software that brings together health and biomedical vocabularies and standards to enable interoperability between computer systems.	Supports 25 languages, 220 vocabularies, 4 536 653 concepts	https://www.nlm.nih.gov/research/umls/index.html
COSMIC	2004 (2021)	A comprehensive resource for exploring the impact of somatic mutations in human cancer	1 491 089 samples, 9 215 470 gene expression variants, 28 175 papers	https://cancer.sanger.ac.uk/cosmic
DrugBank	2006 (2022)	Contain information about drugs, drug mechanisms of action, drug interactions and drug targets, providing an up-to-date list of approved and investigational drugs in clinical trials.	14 528 unique drugs, 4894 unique targets and 18 974 drug target pairs	https://go.drugbank.com/
ChEBI	2008 (2022)	A freely available dictionary of molecular entities focused on 'small' chemical compounds.	59 214 chemical entities that have been manually reviewed	https://www.ebi.ac.uk/chebi/
FunCoup	2009 (2020)	A comprehensive functional association network databases focused on discovering novel functional associations between proteins.	22 species, 60 999 771 functional associations	https://funcoup5.scilifelab.se
SemMedDB	2012 (–)	A repository of semantic predications (subject-predicate-object triples) extracted by semrep, a semantic interpreter of biomedical text	93.6 million predications from all of PubMed citations (about 32.9 million citations) and forms the backbone of the Semantic MEDLINE application.	https://skr3.nlm.nih.gov/SemMedDB/
DISEASES	2015 (2022)	A resource that integrates evidence on disease-gene associations	17 606 genes, 4610 diseases, 543 405 disease–gene associations	https://diseases.jensenlab.org/Search
SIDER	2010 (2015)	Contains information on marketed medicines and their recorded adverse drug reactions	1430 drugs, 5868 side effects, 139 756 drug-side effect pairs	http://sideeffects.embl.de/
DrugCentral	2016 (2021)	Provides information on active ingredients chemical entities, pharmaceutical products, drug mode of action, indications, pharmacologic action.	4714 Drugs, 129 975 pharmaceuticals	https://drugcentral.org/

Table 3. Methods of extracting cancer related entities from EMRs

Name	Cancer type	Dataset	Input feature	Embedding layer	Model	Entity types	micro-F1
Weegar *et al.* [59]	Breast cancer	40 pathology reports	Input text	–	Rule-based methods	10 entity types	0.86
Yala *et al.* [60]	Breast cancer	91 505 pathology reports	Input text	N-gram representation	boosting classification	20 entity types	0.96
Si *et al.* [61]		>260 000 pathology reports	Input text	Character Embedding Glove Embedding MIMIC-III Embedding	BiLSTM-CRF	Diagnosis Therapeutic procedure Tumor description	0.94 0.96 0.87
Gao *et al.* [62]	Breast cancer Lung cancer	942 pathology reports	Input text	Word2vec Embedding Glove Embedding	HANs+BiLSTM/GRU	Primary site (12 classes) Histological grade	0.80 0.90
Alawad *et al.* [63]	Breast cancer Lung cancer	942 pathology reports	Input text	Word2vec Embedding	Coarse-to-Fine Multi-task CNN	Primary Site (2 classes) Laterality (4 classes) Histological grade	0.77 0.96 0.79
Yoon *et al.* [64]	Breast cancer Lung cancer	942 pathology reports	Input text	Word2vec Embedding	Multi-task Text GCN	Primary site (6 classes) Laterality (2 classes) Behavior (2 classes) Histological grade (3 classes)	0.99 0.99 0.99 0.98
Zhang *et al.* [65]	Breast cancer	600 clinical records	input text	–	BERT-BiLSTM-CRF	41 entity types	0.94
Wang *et al.* [66]	Lung cancer	2311 clinical records	input text	Word2vec Embedding	CNN	Histology Stage Histological grade Therapies	0.93 0.79 0.85 1.00
Liu *et al.* [67]	Liver cancer	1089 abdomen and pelvic CT radiology reports	Input text Lexicon feature sequence	Word2Vec Embedding	BiLSTM-CRF	Location Morphology Enhancement Density Modifier	0.95 0.96 0.89 0.85 0.81
Alawad *et al.* [68]	Pan-cancer	71 223 pathology reports	Input text	–	Multi-task CNN	Primary site (65 classes) Laterality (4 classes) Behavior (4 classes) Histological grade(63 classes) Histological type(5 classes)	0.94 0.93 0.97 0.81 0.8
Wu *et al.* [69]	Pan-cancer	3622 pathology reports	Input text	Glove Embedding	TextGCN+ Attention+ BiLSTM+CRF	Cancer types Laterality Cancer subtypes Histopathological grade TNM staging Diagnosis	0.91 0.87 0.80 0.91 0.96 0.71
Wang *et al.* [70]	Intestinal cancer	8818 EMRs	Input text	Tencent Embedding Word2vec Embedding	BiLSTM+CNN +CRF MT-MI-BILSTM-ATT	Specimen name, tumor size, infiltration depth, cancer metastatic ratio, adenocarcinoma shape Cancer involved in upper margin, cancer involved in bottom margin, cancer involved in base margin, nerve invasion, vascular invasion	0.70/0.79/0.71/ 0.79/0.72 0.87/0.83/0.84/ 0.80/0.81
Alawad *et al.* [31]	Pan-cancer	878 864 pathology reports	Input text Concept CUIs	Word Embedding BOE Embedding	text-CNN-CUIs-BOE	Histology (556 classes) Sub-site (317 classes)	0.79 0.68
Liu *et al.* [71]	Liver cancer	1089 abdomen and pelvic CT radiology reports	Input text	Word2vec Embedding	BERT-BiLSTM-CRF	APHE PDPH	0.89 0.82
Solarte Pabón *et al.* [72]	Lung cancer	Clinical report contains 14 750 sentences	Input text lemmas POS	Medical Embedding Char Embedding	BiLSTM-CRF	Cancer entity, stage, TNM, date, family member, events, treatments, drug	0.90

POS: Part of Speech; CUIs: concept unique identifiers.

to this day. The quality of the dictionary and the entity matching algorithm affect the efficiency of the extraction results. The content of the dictionary should be rich, and it needs to be maintained and updated periodically. Matching algorithms are used to match candidate entities in the text with the entities in the dictionary.

As noted in Knowledge sources Section, UMLS is the most widely used public dictionary resource in the biomedical field. Knowlife [30] and DSKG [21] used UMLS as a dictionary, and the types of entities, such as gene, protein, anatomy, physiology etc., were extracted from the text. In Knowlife, in order to efficiently process large dictionaries and a large number of input corpora, the matching algorithm uses locality sensitive hashing (LSH) with min-wise independent permutations (MinHash) to quickly find candidate entities from the text. DSKG uses minimum hashing as a matching algorithm to determine whether a candidate entity in a biomedical text is in the UMLS dictionary. The dictionary-based entity extraction approach is simple and easy to use, but it relies heavily on the dictionary. This approach can only extract the entities that exist in the dictionary, and it is difficult to identify unknown entities.

With rule-based methods, rules are manually constructed for the specific task and the corpus, and then these rules are used to extract entities from the text. Weegar *et al.* [59] used a rule-based method to extract entities from breast cancer pathology reports. Firstly, the written form and contextual identities of the target entities in the pathology reports were observed. Subsequently, the corresponding rules were constructed and entities were extracted from the pathology reports according to the rules. Their method achieved an F1-score of 86%. Rule-based methods need to manually construct rules according to specific tasks, and as such they incur higher costs, with poor generalization. But the advantage of this approach is that there is no need to label data.

The learning ability of machine learning-based models is limited. When the amount of data is too large, it is difficult for machine learning approaches to learn more features from the data for entity extraction tasks. Traditional machine learning approaches rely on feature engineering, and handcrafted features sometimes need to be added to the model to achieve good performance.

Deep learning-based methods

More recently, deep learning has also contributed to the field of entity extraction. The deep learning-based approach refers to entity extraction by constructing a deep neural network model that automatically learns more useful features from the data. The processing of a neural network model to entity extraction is shown in Figure 4A. It is divided into three parts: an input layer, embedding layer and neural networks layer.

Input layer

The neural network model automatically extracts grammatical and semantic features from the text without relying on feature engineering, but sometimes the features automatically extracted by the neural network are not effective. Consequently, features can be constructed through prior knowledge, which helps to improve the performance of the model. Entity extraction mainly involves obtaining information from the words in the text, and additional related features of the words can help to improve the performance. Liu *et al.* [67] additionally added lexicon feature sequence to enhance the model's ability to recognize entity boundaries. Alawad *et al.* [31] used biomedical concept unique identifiers (CUIs) as input data, which were subsequently represented as vectors. Pabón *et al.* [72] additionally input the lemma and part of speech (POS) of the text sequence, so that the model could learn the semantic and grammatical features underlying text sequences.

Embedding layer

In the early days, pre-trained word embeddings (PWEs) were mainly used as embedding layer to obtain word representations of the input text. Chiu *et al.* [73] trained word embedding in the biomedical field for the first time. Habibi *et al.* [74] compared the performance of training word embedding on four corpora (clinical notes, biomedical literature, Wikipedia and news articles). The results showed that the word embedding trained on a corpus of biomedical literatures had the best performance. In addition, there are also methods such as character embedding, bag-of-words embedding, etc. However, PWE obtained from pre-trained word embedding is non-contextual and cannot address the issue of polysemy.

Currently, PLMs are often used as embedding layers. In order to solve the problems of PWEs, PLMs have begun to receive widespread attention and applications, because they can obtain the contextual embedding of each word in a text sequence to address the issue of polysemy. BERT [75] made each word has global semantic information through a deep network structure and self-attention mechanism based on Transformer [76].

Using professional domain specific corpora to train the model can improve its performance. BioBERT [77] is a language representation model in the biomedical field pre-trained on a large-scale biomedical corpus. It performed better than BERT and previous state-of-the-art models. Similarly, the PLMs [78–80] constructed by training on a corpus in the clinical domain were used to improve the effect of text mining in the clinical domain.

Neural network layer

Neural network model architectures can be classified into CNN-based (Figure 4B), RNN-based (Figure 4C), graph convolutional neural network (GCN)-based (Figure 4D), hybrid-based (Figure 4E) and transformer-based approaches (Figure 4F).

CNN is one of the most widely used neural network architectures. Figure 4B shows a general CNN model used for cancer entity extraction. Wang *et al.* [66] used a normal CNN to extract four types of cancer entities from the clinical records of lung cancer patients. Alawad *et al.* [52] proposed a coarse-to-fine multi-task CNN model to extract three types of entities from cancer pathological reports. Their model consists of two phases. In the first stage, an MT learning with hard parameter sharing (HPS) approach trains an MT-CNN model to learn the shared features. In the second stage, pre-trained MT-CNN model parameters are used to initialize a CNN model for each individual task, and then each CNN model is retrained separately in the corresponding dataset. They found that the MT-CNN outperformed single-task CNN, especially for imbalanced data. Recently, Alawad *et al.* [68] compared two multi-task learning methods, cross-stitch and HPS. HPS is suitable for the similar subtasks so that features can be shared between subtasks to improve the performance of the model. The cross-stitch method is suitable for subtasks with large differences. The CNN-based model can train the model quickly, but the receptive field of the convolutional layer is limited, so it is difficult to extract correctly when the medical entity name is long. In addition, the entity extraction model based on multi-task learning has achieved good experimental results and prevent overfitting and reduce the need for data.

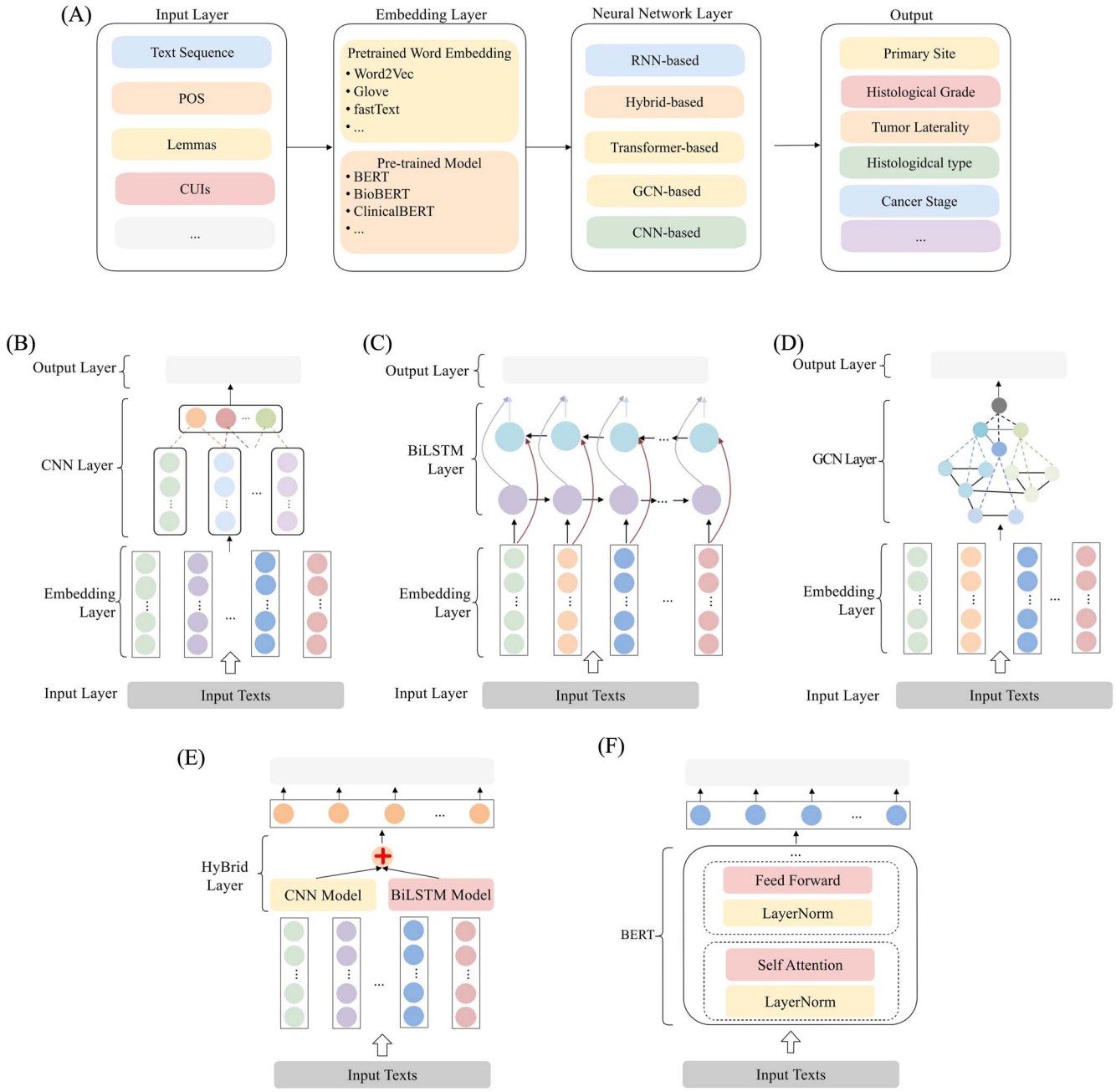

Figure 4. Illustrations of deep learning-based methods for entity extraction. (**A**) Processing a neural network model for entity extraction. (**B**) CNN-based model. (**C**) RNN-based model. (**D**) GCN-based model. (**E**) Hybrid-based model. (**F**) Transformer-based model.

RNN is a neural network with short-term memory. In order to overcome the problem of long-sequence dependence., a gating mechanism is added on the basis of the RNN model, such as long short-term memory (LSTM) [81]. Si *et al.* [61] proposed a system based on bidirectional long short-term memory conditional random field (BiLSTM-CRF) to extract frame semantic information from clinical narratives. It achieved *F*1-scores of 93.7% for cancer diagnosis, 96.33% for therapeutic procedure and 87.18% for tumor description. Liu *et al.* [67] used the BiLSTM-CRF model to extract entities from radiology reports. Their model introduces a lexicon feature sequence which can effectively promote the learning of entity boundaries. The model with the lexicon improved the *F*-score by 2.22% compared to the model without the lexicon. For entity extraction, the BiLSTM-CRF model [82] has attracted much attention. BiLSTM-CRF is a classic entity extraction model, which is often used as a baseline for simple biomedical entity extraction tasks. Compared with the CNN-based model, it takes a long time to train the model.

Recently, the emerging GCN has shown promise in biomedicine. GCNs are specifically used to process graph structure data and extract feature information through convolution operations. So when applying GCNs to entity extraction, it is necessary to transform the input data from a text sequence into a graph. Yoon *et al.* [64] proposed a multi-task GCN. First, the input text sequence was transformed into a graph with document nodes and word nodes. Graph edges were based on the word occurrence and word co-occurrence in the documents. The graph is then fed to the text GCN model to extract useful features through convolution. Wu *et al.* [69] proposed the encoder-GCN-decoder model, which consists of three parts, namely the encoder, the GCN module

and the decoder. The encoder uses BiLSTM to encode the input text sequences to obtain sentences representations. Then these sentence representations are input into the GCN module. In this module, three graphs are constructed according to the input text, a semantic-based graph, a syntax-based graph and a sequence-based graph. The graphs share the same nodes, but their edges are different. The decoder is based on the BiLSTM-CRF model and feeds the output of the GCN module to the decoder. Finally, the decoder completes the entity extraction. GCN can extract features from global information to discover the connections between words. However, GCN needs to transform the existing input text into a graph structure according to the rules, which depend on experience.

Hybrid-based methods generally use multiple neural network architectures (e.g. the combination of a CNN and BiLSTM), combining the advantages of each to achieve better performance. Figure 4E shows a hybrid neural network architecture using a CNN and BiLSTM, which, respectively, extract features from the input sequence, and then concatenate the two results into the final classifier. Wang *et al.* [70] proposed a model to extract entities, in which BiLSTM and a CNN were used to learn text features, and CRF was used as the last layer of the model.

Transformer [76] has become the mainstream model in the field of NLP. BERT is a bidirectional transformer-based model pre-trained over a large general English domain corpus. Based on BERT, Liu *et al.* [71] and Zhang *et al.* [65] proposed models for entity extraction tasks and achieved good performance. Liu *et al.* used BERT trained on a corpus of the general domain. Zhang *et al.* trained BERT on a corpus of the clinical domain, so that BERT can learn the semantic and grammatical features of texts in the clinical domain. In addition to BERT, Lee *et al.* developed BioBERT, a BERT model trained on biomedical literature from PubMed [77]. Using BioBERT, Xu *et al.* constructed a PubMed knowledge graph to understand the trends over time of researcher-centric and bio-entity-centric activity [22]. Transformer-based models are the mainstream methods for current entity extraction tasks, which can extract deep semantic features from texts and achieve better experimental results. However, the model structure is large, and it requires more computing resources in the training and inference stages.

Tools for entity extraction

Many powerful tools have been developed for entity extraction to greatly facilitate researchers. Table 4 summarizes the entity and relation extraction tools in the biomedical and clinical domains. These tools are in the form of web services, software clients or Python packages. Some entity extraction tools additionally provide entity normalization capabilities.

BERN2 [83] and PTC [26] are web tools for entity extraction and entity normalization in the biomedical field. Users can use these tools quickly extract biomedical entities from PubMed abstracts or PMC full-text articles. QuickUMLS [84], HunFlair [85], BERN2, etc., provide python packages, which are convenient for users to use with highly extensible. BERN2, PTC and HunFlair use deep learning-based methods with high accuracy and generalization, while CLAMP and cTAKES mainly integrate traditional machine learning algorithms, dictionary-based methods and rule-based methods.

Relation extraction

Relation extraction aims to extract the relations between entities from the knowledge source. The relations are usually in the form of triples (e_1, r, e_2), which indicate that there is a relation r between the entity e_1 and the entity e_2. Relation extraction tasks in the biomedical and clinical domains include (but are not limited to) protein–protein interactions (PPIs), drug–drug interactions (DDIs), chemical-protein interactions (CPIs) [86], adverse drug events (ADEs) [87] and chemical-induced disease (CID) [88]. Figure 3C shows an example of extracting knowledge triples from literature. In this review, we regard relation extraction as relation classification, which identifies whether the candidate entity pair has a semantic relation within a text sequence and pre-defines the relation type. Relation classification tasks can be binary classification or multi-classification. For example, CID relation extraction is regarded as a binary classification task to predict whether candidate entity pairs have CID relations [88]. But Zhang *et al.* [89] regarded the relation extraction as a multi-classification task. Relation extraction models are summarized in Table 5.

Traditional methods

Traditional relation extraction methods are rule-based methods, which directly extract triples from corpora. These methods use manually constructed rules based on the characteristics of text data. Li *et al.* [5] used a method based on heuristic rules to extract nine types of disease-associated relations for KG construction. Rule-based methods can achieve high accuracy but they have low recall. The rules constructed can only be used in the current task, and they are not easily transferred to other tasks. Moreover, rule-based methods require experts to build the rules manually, which is costly. Pattern-based methods automatically build rules based on the text, instead of manually building rules.

Deep learning-based methods

Deep learning-based methods are also widely used in relation extraction. This section takes the biomedical relation classification and clinical relation classification as examples to introduce specific deep learning-based methods. Figure 5A shows a general framework for relation extraction based on neural networks.

Input layer

There are many features in relation extraction that can effectively improve the performance of the model in biomedicine, such as position features [89,91,92], semantic entity types [65,91,92] and external knowledge [95]. The most commonly used external knowledge is from published knowledge bases. Qi *et al.* [92] first manually constructed triples on the original dataset. They used knowledge embedding to obtain knowledge representations, and then added knowledge representations when training the model.

Neural network layer

In the relation extraction task, the methods used in the neural network layer can be divided into CNN-based (Figure 5B), RNN-based (Figure 5C), GCN-based (Figure 5D), and transformer-based methods (Figure 5E) according to the neural network architecture similar to those for the entity extraction methods.

CNN-based relation extraction methods have been proposed for many years [96]. Zhang *et al.* [89] proposed a CNN-based neural network model called ResNet-PAtt to extract relational triples from Chinese EMRs. In the model, relation extraction was regarded as a multi-class relation classification, and position features were added to enable the model to directly obtain the location information of the candidate entity pair in the input sequence. The CNN-based model can only extract context information from short text sequences. When the distance between entity pairs is large, the model can not extract the contextual information of the full text, resulting in poor performance.

Table 4. The entity and relation extraction tools in the biomedicine

Name	Year	Description	Type	Platform	URL
MetaMap	2001	Mapping biomedical text to the UMLS Metathesaurus or to discover Metathesaurus concepts referred to in text.	EE	Java	metamap.nlm.nih.gov
QuickUMLS	2016	A tool for unsupervised biomedical concept extraction from medical text.	EE	Python	github.com/Georgetown-IR-Lab/QuickUMLS
DeepPhe	2017	An entity extraction tool to extract cancer-related entity information from EMRs	EE	Java	github.com/DeepPhe/DeepPhe-Release
OGER++	2019	A hybrid system for named entity recognition and concept recognition (linking)	EE/EN	Python	github.com/OntoGene/OGER
Pubtator Central	2019	Providing automatic annotations of biomedical concepts such as genes and mutations.	EE/EN	Web	www.ncbi.nlm.nih.gov/research/pubtator
BERN2	2022	Recognition of 9 biomedical entity types (Gene, Disease, Chemical, Species, etc.)	EE/EN	Web/Python	bern2.korea.ac.kr
DTranNER	2020	A deep-learning-based method suited for biomedical named entity recognition on the five biomedical benchmark corpora (BC2GM, BC4CHEMD, BC5CDR-disease, BC5CDR-chemical and NCBI-Disease).	EE	Python	github.com/kaist-dmlab/BioNER
HunFlair	2021	A NER tagger for biomedical texts. It comes with models for genes/proteins, chemicals, diseases, species and cell lines.	EE/EN	Python	github.com/flairNLP/flair/blob/master/resources/docs/HUNFLAIR.md
SemRep	2003	Extracts semantic predications (subject-relation-object triples) from biomedical free text.	RE	C/C++/Web	semrep.nlm.nih.gov
cTAKES	2010	Information extraction from the clinical narrative that can discover codable entities, temporal events, properties and relations.	EE/RE	Java	ctakes.apache.org
MedTagger	2013	A pipeline based on UIMA framework for indexing based on dictionaries, information extraction and machine learning-based named entity recognition from clinical text	EE/RE	Java	github.com/OHNLP/MedTagger
CLAMP	2017	A clinical NLP toolkit that can be used for cancer information extraction	EE/RE	Java/Web	clamp.uth.edu/get-clamp.php
BiOnt	2020	Performing relation extraction using multiple biomedical ontologies regarding gene-products, phenotypes, diseases and chemical compounds, respectively.	EE/RE	Python	github.com/lasigeBioTM/BiONT
GNormPlus	2015	A system that handles both gene/protein name and identifier detection in biomedical literature, including gene/protein mentions, family names and domain names.	EE/EN	Java/Perl	www.ncbi.nlm.nih.gov/research/bionlp/Tools/gnormplus
tmChem	2015	Identifying chemical names in biomedical literature	EE/EN	Java/Perl/C++	www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmchem
tmVar 2.0	2013	An approach for extracting sequence variants in biomedical literature	EE/EN	Java	www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar
DNorm	2013	A method for determining which diseases are mentioned in biomedical text	EE/EN	Java	www.ncbi.nlm.nih.gov/research/bionlp/Tools/dnorm/

EE: Entity Extraction; EN: Entity Normalization; RE: Relation Extraction

RNN-based models were summarized in detail above in the section on entity extraction. These models are suitable for processing sequence data and model text sequences. In particular, relation extraction needs to use the information of the entire text sequence. Xiu *et al.* [11] regarded relation extraction as a multi-class relation classification task and defined 16 relation types utilizing a bidirectional gated regression unit neural network and dual-attention mechanism at the word and sentence levels to extract the relation. The final *F*1-score of the model was 51.67%. Zhang *et al.* [94] proposed a BiLSTM-based model with a multi-hop self-attention mechanism to extract medical knowledge from Chinese medical literature. The model achieved *F*1-score of 93.19% for therapeutic relation tasks and 73.47% for causal relation tasks. Compared with the CNN-based model, the RNN-based model can extract semantic information from a longer text, but when the text sequence is too long, there is a problem of gradient vanishing or exploding.

GCN-based methods have also been applied in relation classification tasks. Zeng *et al.* [90] proposed the CID-GCN model, a GCN-based model with a gating mechanism, to predict whether a candidate entity pair has a chemical-induced-disease (CID) relation from the context of a global document. Wang *et al.* [97] proposed a document-level relation classification model, which mainly consists of BiLSTM, GCN and Multihead self-attention. BiLSTM

Table 5. Summarized of relation extraction models

Name	Dataset	Input feature	Embedding Layer	Model	Task type	External mechanism	F1
Zhang *et al.* [56]	Chinese EMRs	Input text Position features	Word2vec Embedding	ResNet-PAtt: Residual CNN networks+Attention	Multi-classification	Attention mechanism	0.78
Zeng *et al.* [90]	BioVCDR	Input text	Biomedical pre-trained word2vec Embedding	CID-GCN: GCNs with gating mechanism	Binary classification		0.65
Christopoulou *et al.* [91]	2018 n2c2 challenge	Input text Position features Semantic entity types	Biomedical pre-trained word2vec Embedding	Weighted BiLSTM+ Walk-based model Transformer model	Binary classification	Attention mechanism	0.95
Qi *et al.* [92]	Chinese medicine instructions	Input text Position features Semantic entity types External knowledge		KeMRE: BERT-CNN-LSTM based framework	Multi-classification	External knowledge	0.99
Xiu *et al.* [11]	Chinese electronic clinical reports	Input text	Word2vec Embedding	BiGRU with attention	Binary classification	Attention mechanism	0.52
Zhang *et al.* (2019)	Clinical records	Input text Semantic entity types	–	BERT	Binary classification	Attention mechanism	0.97
Hebbar *et al.* (2021)	BioVCDR	Input text	–	Covid-BERT	Binary classification	–	0.91
Yang *et al.* [93]	2018 MADE1.0 2018 n2c2	Input text Cross-sentence distance	–	RoBERTa; BERT; XLNet	Binary classification	Attention mechanism	MADE1.0: 0.90 2018 n2c2: 0.96
Zhang *et al.* [94]	Chinese medical literature abstracts	Input text Position features	Word2vec Embedding	BiLSTM with multi-hop self-attention	Binary classification	Attention mechanism	therapeutic: 0.93 causal: 0.74

is used to extract contextual representations from the text. The document-level dependency graph is constructed from the text, and then GCN model extracts global representations from the graph. Multihead self-attention extract the most useful features for relation classification tasks from contextual representations. Finally, the above-extracted features are concatenated for relation classification. The GCN-based models convert text into graphs according to certain rules and then extract global feature from the graphs through convolution operations, which solves the task of document-level relationship extraction to some extent.

Transformer-based models have achieved state-of-the-art performance for relation extraction in biomedicine. Yang *et al.* [93] compared three mainstream transformer-based models for clinical relation extraction: XLNet, BERT and RoBERTa. Among these three transformer-based models, RoBERTa and XLNet outperformed BERT for clinical relation extraction. Chen *et al.* [95] used BERT to extract contextual features, additionally introduced prior knowledge features in the knowledge graph, and then fused the features for relation classification. Transformer-based models have achieved satisfactory results in both entity extraction tasks and relation extraction tasks. But the restriction on the input text length makes it cannot solve the document-level relation extraction task.

As with entity extraction, hybrid model frameworks are also good choice for relation extraction in biomedicine. Qi *et al.* [92] proposed a hybrid model frame that combined BERT, CNN and LSTM to extract medical knowledge from medicine instructions. The framework includes four modules. The first module uses the BERT-CNN-BiLSTM model to obtain contextual text representations from the input text. The second module to obtain the relatedness between entities. The third module adds external medicine knowledge for relation extraction. The fourth module is the relational classifier, which predicts the relation category based on global contextual entity representations and knowledge embedding.

The deep learning-based methods presented above are supervised learning methods that rely on manually annotated datasets to train models. Current unsupervised learning-based relationship extraction methods are also being developed. For example, Percha *et al.* [25] integrated ensemble biclustering algorithm (EBC) with hierarchical clustering to learn the relationships.

Tools for relation extraction

Table 4 summarizes the relation extraction tools. These tools are mainly in the form of software clients. SemRep [98] is the most commonly used information extraction tool in the biomedicine. It can extract semantic triples from biomedical free text. The extracted entities and semantic relationships respectively come from metathesaurus and semantic network of the UMLS. It both provides a web service and a software installation package. The web service is easy to use, but slow to process, while the local package installation is relative complicated because the tool relies on UMLS. BiOnt is also a tool for relation extraction from the medical literature [99]. CLAMP (Clinical Language Annotation, Modeling and Processing) [100], cTAKES and MedTagger [101] are information extraction tools for clinical text that developed based on Java and provides a user friendly GUI interface. CLAMP-cancer [102] is a component of CLAMP that focuses on extracting cancer-related information from EMRs. MedTagger and SemRep provide only a Character User Interface (CUI) for local use, while BiOnt provides users with a python package that is highly scalable.

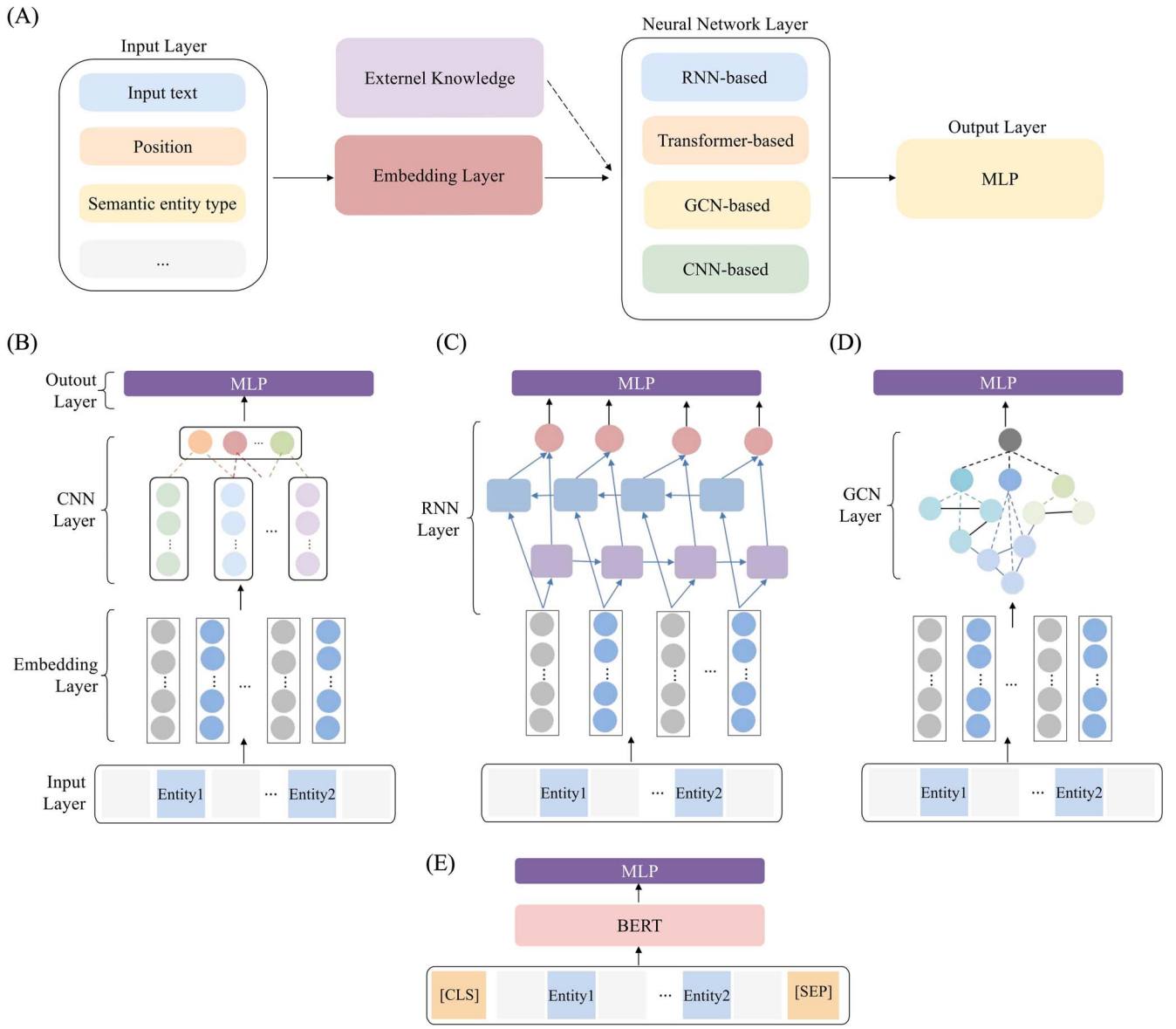

Figure 5. Illustrations of deep learning-based methods for relation extraction. (**A**) Processing a neural network model for relation extraction. (**B**) CNN-based model. (**C**) RNN-based model. (**D**) GCN-based model. (**E**) Transformer-based model.

SemRep, CLAMP, cTAKES and MedTagger integrates several traditional machine learning methods, dictionary-based methods and pattern matching-based methods. BiOnt employs deep learning-based methods using multiple biomedical ontologies.

Applications of knowledge graph in complex disease

Recently, KGs have already attracted much attention and are widely used in biomedicine. In this section, we review the applications of KGs for complex disease in three aspects including knowledge base, drug discovery and clinical decision. Knowledge Base is the most basic application, which can facilitate researchers to see the complex relationship among entities more intuitively. In the application of CDKG in drug discovery and clinical decision, they mainly use the information in the CDKG to build corresponding prediction models. Moreover, drug discovery relies more knowledge in the biomedical field, while clinical decision needs more knowledge in the clinical field. We also provided an example to show how CDKG could help to drug discovery for gastric cancer in Supplementary Materials available online at http://bib.oxfordjournals.org/.

Knowledge base

To build a KB, knowledge graphs can be classified into two types: professional KGs and comprehensive KGs.

The goal of a professional KG is to serve as a knowledge base for specific aspects, such as a certain disease. KGHC [10] obtained knowledge from SemMedDB and literature to construct a professional KG of hepatocellular carcinoma to improve research efficiency. Rossanez *et al.* [17] constructed a KG about Alzheimer's disease from biomedical literature. Domingo-Fernández *et al.* [36] extracted knowledge from 160 biomedical publications to construct a Covid-19 KG with relevant research information. Currently, the professional KBs are still at the initial stage. Very few complex diseases have a corresponding KB.

Comprehensive KB are generally constructed by merging a large number of existing medical databases to provide more comprehensive knowledge of different diseases. For example, PreMedKB [32] integrates more than 20 public databases and establishes the relation between diseases, genes, mutations and drugs through the application of semantic network technology. The goal of PreMedKB is to promote the interpretation of the clinical significance of a patient's genetic variations.

Drug discovery

KGs with rich knowledge and logical reasoning can be also applied to assist with drug design, such as drug target prediction, adverse drug reaction (ADR) discovery and drug repositioning.

Drug target prediction is a critical part of drug discovery in pharmaceutical research and plenty of computational methods have been developed for it, as well as KG-based methods. Wang *et al.* [103] proposed a knowledge graph-based deep learning method called KG-DTI for drug target interaction predictions from the DrugBank database using DistMult [104] embedding methods and a fully connected neural network.

ADR discovery involves finding the relations between potential drugs and adverse reactions to speculate on the possibility of whether a drug will cause an adverse reaction. A tumor-biomarker knowledge graph (TBKG) [48] was used to discover potential adverse drug reactions of anti-tumor drugs. The TBKG was built from biomedical literature. A naive Bayesian model was used to calculate the weight of the edge in the KG. The depth first search (DFS) algorithm was used to calculate all the paths between drugs and ADRs in the KG to predict potential ADRs under the hypothesis that the shorter the distance between drug and ADR in the path, the greater the possibility of the drug causing that ADR.

Drug repurposing, also known as drug repositioning/rediscovery, involves identifying novel indications of approved drugs to new disease. KG-based methods can be an important supplement to existing silicon models. Zhu *et al.* [105] proposed a KG-based strategy from biomedical literature to mine information pertaining to predicting drugs for repurposing against Parkinson's disease. Drug repurposing is regarded as a binary classification task: given a drug and a disease, the task is to predict whether or not the drug can treat the given disease. They improved the KG-based strategy by fusing the literature-based KG with a medical knowledge base and employing KG completion methods [20]. Sose *et al.* [8] identified drug repurposing opportunities in rare disease using embeddings learned from the CDKG with high performance.

Clinical decision

With the popularization of EMRs in hospitals, a large volume of patient-level data has accumulated in the clinical domain. These data can be used to extract structured data for a KG. Li *et al.* [5] constructed a medical knowledge graph (MKG) based on EMRs, which contained 22 508 entities and 579 094 quadruplets. It was later applied to a decision support system and medical information retrieval. Gong *et al.* [106] proposed a safe medicine recommendation (SMR) framework for recommending safe medicines for patients using a high-quality heterogeneous graph generated by linking EMRs and an MKG. In order to achieve precision medicine, Santos *et al.* [19] combined public databases, literatures and proteomics data to construct a clinical knowledge graph (CKG) composed of more than 16 million nodes and 220 million relations. They used this CKG to identify biomarkers for cancers.

Discussion and future prospects

In this review, we conducted a comprehensive comparison and analysis of the construction and application of CDKG. Current application scenarios for KGs in biomedicine show the wide applicability and potential of KGs for a variety of different uses and systems. However, it is still at an early stage with multiple open challenges.

Challenge one: insufficient labeled data

Most approaches in knowledge extraction are supervised learning, which require a large amount of manually annotation data. Generally, the more labeled data, the better the performance of the model. However, labeling data in biomedical fields is more difficult than in other fields because of the following reasons:

(1) Data in biomedicine are usually complex and noisy, and experts are needed to label the data.
(2) The original corpus is difficult to obtain and share, such as EMRs that contain the patient's private information.
(3) We still lack of comprehensive and accurate ontology in biomedicine due to its extreme complexity, although several ontology systems have been proposed such as UMLS.

Besides development of new methods to get enough labeled data, another strategy is to reduce the dependence on labeled data, transfer learning has received extensive attention. Transfer learning involves transferring model parameters that have been trained in other tasks to a new model which no longer needs to be trained from scratch. In addition, few-shot learning and unsupervised learning are developing and applying rapidly.

Challenge two: complex problem

Most current entity extraction tasks and relation extraction tasks consider simple problems, but lack a consideration of complex problems. For example, for entity extraction, complicated situations include (1) nested entities, where one entity name is nested in another entity name, such as '<DISEASE> <CELL TYPE> liver cell </CELL TYPE> carcinoma </DISEASE>' and (2) conjunction and disjunction, where two or more entity names share a head noun. For example, in 'fibrous and globular proteins' there are two entities 'fibrous proteins' and 'globular proteins'. For relation extraction, it is usually assumed that knowledge triples are extracted from a single sentence. The actual situation is, however, more complicated. For example, some candidate entity pairs are separated by multiple sentences, but still have a semantic relation. Moreover, when predicting the relation type of a candidate entity pair, the context or even the whole document should be considered, even if the candidate entity pair appears in the same sentence.

Knowledge extraction with CDKG methods is divided into two tasks: entity extraction and relation extraction. All entities are extracted from the text and then the relations are extracted and classified. This method is a so-called pipeline-based method, which decomposes a complex task into multiple simple tasks. However, this method leads to cascading errors. For example, errors generated by the entity extraction task will be transferred to the subsequent relation classification task. In addition, there is a connection between entity recognition and relation classification in knowledge extraction tasks, and the above method cuts off the connection between the two tasks. A joint learning method is one solution, whereby the entity recognition task and the relation classification task are learned at the same time [107].

Moreover, recently several nature language models have been popular due to their enhanced capabilities and performance, such as GPT-3 [108] and OPT [109], and is starting to be applied to biomedicine [110]. However, these models are usually considerably larger in terms of memory requirements and is more computationally intensive, which is one of the main challenge for the application in complex disease.

Challenge three: standardization procedure

KG development has so far been carried out without clear and uniform guidelines, and each team has developed its own solutions. Hence, the standardization of CDKG development should be an important direction for future research in the field of biomedicine. In addition, some efficient model should be employed in the standardization, such as PLMs which have been widely used in entity extraction tasks, entity standardization tasks and relation extraction tasks, because of they can address the issue of polysemy.

Challenge four: contexts and contradictions

Unspecified contexts and contradicting research claims in biomedical literature can lead to inaccuracy of information extraction for CDKG and then make its applications ineffective in practical problems [9]. Sometimes relationships between entities can be extracted that are only valid in a particular context; sometimes there are either direct contradictions or relative contradictions information in the literatures. Therefore, it is a challenge to the field to make sure that KGs are appropriately context-sensitive and can manage the contradictions in different literature. In addition, comprehensive and suitable evaluation methods for CDKG from multiple perspective, such as accuracy, completeness, relevance, consistency, usability, verifiability, etc., could also help to address the challenge.

Challenge five: application of CDKG

At present, most CDKG-related works mainly focus on the construction method of KG and lack the development of medical applications based on KG. However, the applications of CDKG should be pay more attention, and they are the important clinical implications and future directions, such as drug discovery, clinical decision-making, medical question answering systems, medical knowledge retrieval systems, etc. The goal of precision medicine is to diagnose and treat diseases based on differences in patients' genetics, living environment and lifestyle. The CDKG which integrates massive data will provide high-quality biomedical knowledge to contribute to the precision medicine.

Key Points

- CDKG technologies have emerged as promising strategies to overcome these challenges by understanding the interconnections among the biomedical terms to provide new insights in complex disease diagnosis and treatment.
- The construction methods of CDKG can be divided into extracting knowledge from text and from merging databases. The main steps consist of entity and relation extractions.
- At present, the applications scenario of CDKG are predominantly knowledge base, drug discovery and clinical decision.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

None.

Funding

Key project of the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (20KJA520010) ; Collaborative Innovation Center of Novel Software Technology and Industrialization at Soochow University; Key Research and Development Program of Jiangsu Province (BE2020656); Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Sung H, Ferlay J, Siegel RL, *et al.* Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021;**71**:209–49.
2. Ji SX, Pan SR, Cambria E, *et al.* A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Trans Neural Netw Learn Syst* 2022;**33**:494–514.
3. Han X, Wang Z, Zhang J, *et al.* Overview of the CCKS 2019 knowledge graph evaluation track: entity, relation, event and QA. arXiv preprint, arXiv:2003.03875, 2020.
4. Sheng M, Li A, Bu Y, *et al. DSQA: A Domain Specific QA System for Smart Health Based on Knowledge Graph*. Cham: Springer International Publishing, 2020, 215–22.
5. Li L, Wang P, Yan J, *et al.* Real-world data medical knowledge graph: construction and applications (MKG). *Artif Intell Med* 2020;**103**:101817.
6. Tran V, Tran V-H, Nguyen P, *et al.* CovRelex: a COVID-19 retrieval system with relation extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Online, 2021. p. 24–31. Association for Computational Linguistics.
7. Hasan SMS, Rivera D, Wu XC, *et al.* Knowledge graph-enabled cancer data analytics. *IEEE J Biomed Health Inform* 2020;**24**: 1952–67.
8. Sosa DN, Derry A, Guo M, *et al.* A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. *Pac Symp Biocomput* 2020;**25**: 463–74.
9. Sosa DN, Altman RB. Contexts and contradictions: a roadmap for computational drug repurposing with knowledge inference. *Brief Bioinform* 2022;**23**:1–13.
10. Li N, Yang Z, Luo L, *et al.* KGHC: a knowledge graph for hepatocellular carcinoma. *BMC Med Inform Decis Mak* 2020;**20**:135.
11. Xiu X, Qian Q, Wu S. Construction of a digestive system tumor knowledge graph based on chinese electronic medical records: development and usability study. *JMIR Med Informatics* 2020;**8**:e18287.
12. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J* 2020;**18**: 1414–28.
13. Abu-Salih B. Domain-specific knowledge graphs: a survey. *J Netw Comput Appl* 2021;**185**:103076.
14. Alshahrani M, Thafar MA, Essack M. Application and evaluation of knowledge graph embeddings in biomedical data. *PeerJ Comput Sci* 2021;**7**:e341.

15. Wang T, Duan L, He C, *et al.* ATOM: construction of anti-tumor biomaterial knowledge graph by biomedicine literature. In: *2019 IEEE International Conference on BIBM*. San Diego, CA, USA: IEEE, 2019, p. 1256–8.
16. Manning CD, Surdeanu M, Bauer J, *et al.* The stanford CoreNLP natural language processing toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: ACL, 2014, p. 55–60.
17. Rossanez A, dos Reis JC, Torres RS, *et al.* KGen: a knowledge graph generator from biomedical scientific literature. *BMC Med Inform Decis Mak* 2020;**20**:314.
18. Rotmensch M, Halpern Y, Tlimat A, *et al.* Learning a health knowledge graph from electronic medical records. *Sci Rep* 2017;**7**:5994.
19. Santos A, Colaço AR, Nielsen AB, *et al.* Clinical knowledge graph integrates proteomics data into clinical decision-making. *bioRxiv* 2021;1–35. https://doi.org/10.1101/2020.05.09.084897.
20. Zhang X, Che C. Drug repurposing for parkinson's disease by integrating knowledge graph completion model and knowledge fusion of medical literature. *Future Internet* 2021;**13**:14.
21. Yuan J, Jin Z, Guo H, *et al.* Constructing biomedical domain-specific knowledge graph with minimum supervision. *Knowl Inf Syst* 2020;**62**:317–36.
22. Xu J, Kim S, Song M, *et al.* Building a PubMed knowledge graph. *Sci Data* 2020;**7**:205.
23. Leaman R, Wei C-H, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. *J Chem* 2015;**7**:S3.
24. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res* 2013;**41**:W518–22.
25. Percha B, Altman RB. A global network of biomedical relationships derived from text. *Bioinformatics* 2018;**34**:2614–24.
26. Wei C-H, Allot A, Leaman R, *et al.* PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* 2019;**47**:W587–93.
27. Ji Z, Wei Q, Xu H. BERT-based ranking for biomedical entity normalization. *AMIA Jt Summits Transl Sci Proc* 2020;**2020**: 269–77.
28. Sung M, Jeon H, Lee J, *et al.* Biomedical entity representations with synonym marginalization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: ACL 2020, 3614–50.
29. Liu F, Shareghi E, Meng Z, *et al.* 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. Online: ACL* 2021;4228–38.
30. Ernst P, Siu A, Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinformatics* 2015;**16**:157.
31. Alawad M, Gao S, Alamudun FT, *et al.* Multimodal data representation with deep learning for extracting cancer characteristics from clinical text. In: *IEEE International Conference on Big Data, Oak Ridge, TN (United States)*. Georgia, USA: Oak Ridge National Lab. (ORNL), 2020.
32. Yu Y, Wang Y, Xia Z, *et al.* PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. *Nucleic Acids Res* 2018;**47**:D1090–101.
33. Zhang Y, Sheng M, Zhou R, *et al.* HKGB: an inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians' expertise incorporated. *Inf Process Manag* 2020;**57**:102324.
34. Himmelstein DS, Lizee A, Hessler C, *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* 2017;**6**:e26726.
35. Su C, Hou Y, Guo W, *et al.* CBKH: the cornell biomedical knowledge hub. *medRxiv* 2021; 1–15. https://doi.org/10.1101/2021.03.12.21253461.
36. Domingo-Fernández D, Baksi S, Schultz B, *et al.* COVID-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology (COVID-19 KG). *Bioinformatics* 2021;**37**:1332–4.
37. Nickel M, Tresp V, Kriegel H-P. A three-way model for collective learning on multi-relational data. In: *The 28th International Conference on International Conference on Machine Learning*. Washington, USA: Omnipress, 2011, 809–16.
38. Bordes A, Usunier N, Garcia-Duran A, *et al. Translating Embeddings for Modeling Multi-relational Data*. USA: NIPS, 2013, 1–9.
39. Wang Z, Zhang J, Feng J, *et al.* Knowledge graph embedding by translating on hyperplanes. In: *AAAI'14*. USA: AAAI Press, 2014, p. 1112–9.
40. Lin Y, Liu Z, Sun M, *et al.* Learning entity and relation embeddings for knowledge graph completion. In: *AAAI'15*. USA: AAAI Press, 2015, 2181–7.
41. Su C, Hou Y, Guo W, *et al. Biomedical Discovery through the integrative Biomedical Knowledge Hub (iBKH)*. *medRxiv*; 2021;1–44. https://doi.org/10.1101/2021.03.12.21253461.
42. Shang C, Tang Y, Huang J, *et al.* End-to-end structure-aware convolutional networks for knowledge base completion. In: *Proceedings of the AAAI Conference on AI. USA: AAAI Press* 2019;**33**: 3060–7.
43. Guo L, Sun Z, Hu W. Learning to exploit long-term relational dependencies in knowledge graphs. arXiv preprint, arXiv:1905.04914.
44. Yao L, Mao C, Luo Y. KG-BERT: BERT for knowledge graph completion. arXiv preprint, arXiv:1909.03193, 2019.
45. Wang B, Shen T, Long G, *et al.* Structure-augmented text representation learning for efficient knowledge graph completion. In: *Proceedings of the Web Conference 2021*. Ljubljana, Slovenia: Association for Computing Machinery, 2021, 1737–48.
46. Li D, Yang S, Xu K, *et al.* Multi-task pre-training language model for semantic network completion. arXiv preprint, arXiv:2201.04843, 2022.
47. Johnson AEW, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data* 2016;**3**:160035.
48. Wang M, Ma X, Si J, *et al.* Adverse drug reaction discovery using a tumor-biomarker knowledge graph. *Front Genet* 2020;**11**:625659.
49. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
50. Krallinger M, Leitner F, Rabal O, *et al.* CHEMDNER: the drugs and chemical names extraction challenge. *J Chem* 2015;**7**:S1.
51. Pyysalo S, Ohta T, Ananiadou S. *Overview of the Cancer Genetics (CG) task of BioNLP Shared Task 2013*. Sofia, Bulgaria: Association for Computational Linguistics, 2013, 58–66.
52. Bethard S, Savova G, Palmer M, *et al.* SemEval-2017 Task 12: clinical TempEval. In: *SemEval 2017*. Canada:ACL 2017, p. 565–72.
53. Bethard S, Savova G, Chen W-T, *et al.* Semeval-2016 task 12: clinical tempeval. In: *SemEval 2016*. San Diego, California: ACL 2016, p. 1052–62.
54. Bethard S, Derczynski L, Savova G, *et al.* Semeval-2015 task 6: clinical tempeval. In: *SemEval 2015*. Denver, Colorado: ACL, 2015, p. 806–14.

55. Li X, Wen Q, Jiao Z, *et al.* Overview of CCKS 2020 Task 3: named entity recognition and event extraction in Chinese electronic medical records. *Data Intelligence* 2021;3: 1–13.
56. Zhang J, Li J, Jiao Z, *et al.* Overview of CCKS 2018 Task 1: named entity recognition in Chinese electronic medical records. In: *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding*. Singapore: Springer Singapore, 2019, 158–64.
57. Xia Y, Wang Q. Clinical named entity recognition: ECUST in the CCKS-2017 shared task 2. In: *CEUR Workshop Proceedings*. Chengdu, China 2017;1976:43–8.
58. Zhou G, Zhang J, Su J, *et al.* Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics* 2004;**20**: 1178–90.
59. Weegar R, Dalianis H. Creating a rule based system for text mining of Norwegian breast cancer pathology reports. In: *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*. Lisbon, Portugal: ACL, 2015, p. 73–8.
60. Yala A, Barzilay R, Salama L, *et al.* Using machine learning to parse breast pathology reports. *Breast Cancer Research and Treatment* 2016;**161**:203-11.
61. Si Y, Roberts K. A frame-based NLP system for cancer-related information extraction. *AMIA Annu Symp Proc* 2018;**2018**: 1524–33.
62. Gao S, Young MT, Qiu J, *et al.* Hierarchical attention networks for information extraction from cancer pathology reports. *J Am Med Inform Assoc* 2017;**25**:321-30.
63. Alawad M, Yoon HJ, Tourassi GD. Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports. In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. USA:IEEE 2018, 218–21.
64. Yoon H, Gounley J, Young MT, *et al.* Information extraction from cancer pathology reports with graph convolution networks for natural language texts. In: *2019 IEEE International Conference on Big Data*. USA: IEEE, 2019, p. 4561–4.
65. Zhang X, Zhang Y, Zhang Q, *et al.* Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int J Med Inform* 2019;**132**:103985.
66. Wang L, Luo L, Wang Y, *et al.* Information extraction for populating lung cancer clinical research data. *IEEE Int Conf Healthc Inform* 2019;**2019**. https://doi.org/10.1109/ICHI.2019.8904601.
67. Liu H, Xu Y, Zhang Z, *et al.* A natural language processing pipeline of chinese free-text radiology reports for liver cancer diagnosis. *IEEE Access* 2020;**8**:159110–9.
68. Alawad M, Gao S, Qiu JX, *et al.* Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *J Am Med Inform Assoc* 2020;**27**:89–98.
69. Wu J, Tang K, Zhang H, *et al.* Structured information extraction of pathology reports with attention-based graph convolutional network. In: *2020 IEEE International Conference on BIBM*. Seoul, Korea: IEEE, 2020, p. 2395–402.
70. Wang S, Pang M, Pan C, *et al.* Information extraction for intestinal cancer electronic medical records. *IEEE Access* 2020;**8**: 125923–34.
71. Liu H, Zhang Z, Xu Y, *et al.* Use of BERT (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. *J Med Internet Res* 2021;**23**:e19689.
72. Solarte Pabón O, Torrente M, Provencio M, *et al.* Integrating speculation detection and deep learning to extract lung cancer diagnosis from clinical notes. *Appl Sci* 2021;**11**:865.
73. Chiu B, Crichton G, Korhonen A, *et al.* How to train good word embeddings for biomedical NLP. In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Berlin, Germany: ACL, 2016, p. 166–74.
74. Habibi M, Weber L, Neves M, *et al.* Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* 2017;**33**:i37–48.
75. Devlin J, Chang M-W, Lee K, *et al.* Bert: pre-training of deep bidirectional transformers for language understanding. In: *NAACL* Minneapolis, Minnesota: ACL, 2019; **2019**:4171–86.
76. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. arXiv preprint, arXiv:1706.03762, 2017.
77. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40.
78. Huang K, Singh A, Chen S, *et al.* Clinical XLNet: modeling sequential clinical notes and predicting prolonged mechanical ventilation. arXiv preprint, arXiv:1912.11975, 2019.
79. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv preprint, arXiv.1904.05342, 2019.
80. Alsentzer E, Murphy JR, Boag W, *et al.* Publicly available clinical BERT embeddings. arXiv preprint, arXiv.1904.03323, 2019.
81. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.
82. Lample G, Ballesteros M, Subramanian S, *et al.* Neural architectures for named entity recognition. In: *NAACL*. San Diego, California: ACL, 2016, 260–70.
83. Sung M, Jeong M, Choi Y, *et al.* BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* 2022;**38**:4837–9.
84. Soldaini L, Goharian N. Quickumls: a fast, unsupervised approach for medical concept extraction. In: *MedIR Workshop, Sigir*. ACM, 2016, p. 1–4.
85. Weber L, Sänger M, Münchmeyer J, *et al.* HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics* 2021;**37**:2792–4.
86. Krallinger M, Rabal O, Akhondi SA, *et al.* Overview of the BioCreative VI chemical-protein interaction track. In: *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop*. Bethesda, MD USA: BioCreative, 2017, p. 141–146.
87. Henry S, Buchan K, Filannino M, *et al.* 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* 2019;**27**:3–12.
88. Li J, Sun Y, Johnson RJ, *et al.* BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016;**2016**:baw068.
89. Zhang Z, Zhou T, Zhang Y, *et al.* Attention-based deep residual learning network for entity relation extraction in Chinese EMRs. *BMC Med Inform Decis Mak* 2019;**19**:55.
90. Zeng D, Zhao C, Quan Z. CID-GCN: an effective graph convolutional networks for chemical-induced disease relation extraction. *Front Genet* 2021;**12**:624307.
91. Christopoulou F, Tran TT, Sahu SK, *et al.* Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *J Am Med Inform Assoc* 2019;**27**:39–46.
92. Qi T, Qiu S, Shen X, *et al.* KeMRE: knowledge-enhanced medical relation extraction for Chinese medicine instructions. *J Biomed Inform* 2021;**120**:103834.

93. Yang X, Yu Z, Guo Y, *et al.* Clinical relation extraction using transformer-based models. arXiv preprint, arXiv:2107.08957, 2021.
94. Zhang T, Lin H, Tadesse MM, *et al.* Chinese medical relation extraction based on multi-hop self-attention mechanism. *Int J Mach Learn Cybern* 2021;**12**:355–63.
95. Chen J, Hu B, Peng W, *et al.* Biomedical relation extraction via knowledge-enhanced reading comprehension. *BMC Bioinformatics* 2022;**23**:20.
96. Liu C, Sun W, Chao W, *et al.* Convolution neural network for relation extraction. In: *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, 2013, 231–42.
97. Wang J, Chen X, Zhang Y, *et al.* Document-level biomedical relation extraction using graph convolutional network and multihead attention: algorithm development and validation. *JMIR Med Inform* 2020;**8**:e17638.
98. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 2003;**36**:462–77.
99. Sousa D, Couto FM. BiOnt: deep learning using multiple biomedical ontologiesc. In: *Advances in Information Retrieval*, Springer, Cham, 2020;**12036**:36774.
100. Soysal E, Wang J, Jiang M, *et al.* CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2017;**25**:331–6.
101. Liu H, Bielinski SJ, Sohn S, *et al.* An information extraction framework for cohort identification using electronic health records. *AMIA Jt Summits Transl Sci Proc* 2013;**2013**: 149–53.
102. Soysal E, Warner JL, Wang J, *et al.* Developing customizable cancer information extraction modules for pathology reports using CLAMP. *Stud Health Technol Inform* 2019;**264**: 1041–5.
103. Wang S, Du Z, Ding M, *et al.* KG-DTI: a knowledge graph based deep learning method for drug-target interaction predictions and Alzheimer's disease drug repositions. *Appl Intell* 2022;**52**:846–57.
104. Yang B, Yih WT, He X, *et al.* Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint, arXiv:1412.6575, 2015.
105. Zhu Y, Jung W, Wang F, *et al.* Drug repurposing against Parkinson's disease by text mining the scientific literature. *Library Hi Tech* 2020;**38**:741–50.
106. Gong F, Wang M, Wang H, *et al.* SMR: medical knowledge graph embedding for safe medicine recommendation. *Big Data Res* 2021;**23**:100174.
107. Luo L, Yang Z, Cao M, *et al.* A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *J Biomed Inform* 2020;**103**:103384.
108. Brown TB, Mann B, Ryder N, *et al.* Language models are few-shot learners. arXiv preprint, arXiv:2005.14165, 2020.
109. Zhang S, Roller S, Goyal N, *et al.* OPT: open pre-trained transformer language models. arXiv preprint, arXiv:2205.01068, 2022.
110. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. *JMIR Med Inform* 2022;**10**:e32875.