Received: 17 April 2023 | Revised: 29 May 2023 | Accepted: 31 May 2023 | Published online: 31 May 2023

Journal of Data Science and Intelligent Systems
2023, Vol. 1(2) 65–82
DOI: 10.47852/bonviewJDSIS3202973

REVIEW

Models and Techniques for Domain Relation Extraction: A Survey

Jiahui Wang[1,2], Kun Yue[1,2,*] and Liang Duan[1,2]

[1]*School of Information Science and Engineering, Yunnan University, China*

[2]*Yunnan Key Laboratory of Intelligent Systems and Computing, Yunnan University, China*

Abstract: As the significant subtask of information extraction, relation extraction (RE) aims to identify and classify semantic relations between pairs of entities and is widely adopted as the foundation of downstream applications including knowledge graphs, intelligent question answering, text mining, and sentiment analysis. Different from general knowledge, domain knowledge is pertinent to specific fields which include a wealth of proprietary entities and relations. Besides, most of the data are formed as documents rather than sentences. In this paper, the task of domain RE is defined, and the common domains are presented. Furthermore, we provide a systematic review of state-of-the-art techniques as well as the latest trends. We survey different neural network-based techniques for RE and describe the overall framework, training procedures, as well as the pros and cons of these techniques. Then, we introduce and compare the corpus and metrics used for domain RE tasks. Finally, we conclude and discuss future research issues of domain RE.

Keywords: relation extraction, domain knowledge, tasks, sentence-level models, document-level models, corpus

1. Introduction

Relation extraction is the preliminary task of natural language processing. It is challenging to extract different relations among the entities in various sentences and documents due to the specialty of domain knowledge. The tasks, techniques, corpus, and metrics of domain relation extraction are surveyed in this paper.

1.1. Background

Domain knowledge (Alexander, 1992) is the knowledge in specific, specialized disciplines or fields. In contrast to the general or domain-independent knowledge, domain knowledge is often associated with a particular profession, field of study, interest, community, or cultural circle. In recent years, with the exploration of domain data, there has been an increasing need for automated methods to extract knowledge from data. In several domains like biomedical research, financial analysis, and legal documentation, much of the knowledge is implied in unstructured texts, making it difficult to access and analyze. Relation extraction (RE) w.r.t. domain knowledge refers to the process of automatically identifying and extracting relations between entities in a specific domain (E et al., 2019). This involves analyzing the unstructured natural language texts to identify the relation types among entities, which can be any objects or concepts within the domain, such as genes and proteins in the biomedical domain, or companies and executives in the business domain. The entities and relations could be used to build knowledge graphs, which are further used for a variety of applications such as knowledge graphs (Wang et al., 2017), question answering (Xu et al., 2016), sentiment analysis (Xu et al., 2021), information retrieval (Krallinger et al., 2017), and decision support (Wang et al., 2018b).

In general, domain knowledge refers to the specialized knowledge and terminology that is unique to a particular field or industry. For example, the biomedical domain knowledge may include the information about specific genes, proteins, diseases, and treatments, as well as the corresponding relations (Zhou et al., 2014). The financial domain knowledge may include the information about companies, stock prices, financial indicators, and economic trends (Vela & Declerck, 2009). Existing surveys of RE (Akkasi & Moens, 2021; Han et al., 2020; Li et al., 2019c; Nasar et al., 2021; Pawar et al., 2017; Smirnova & Cudré-Mauroux, 2018; Wang et al., 2022a) focus on different models and techniques in general domains or a separate one. It is still worthwhile to comprehensively survey the models and techniques for domain RE in response to the specific challenges in various potential domains and characteristics of domain-specific data.

To fulfill RE in specific domains, it is crucial to have a thorough comprehension of domain-specific knowledge and terminologies. This often involves collaborating closely with domain experts like biomedical researchers or financial analysts to identify relevant entities and relations. Additionally, performing RE necessitates the knowledge of natural language processing (NLP) techniques and machine learning (ML) algorithms. This may entail developing customized models and algorithms that are tailored to the specific domain and data type in unstructured texts or structured data. The speciality of domain knowledge for RE varies according to different domains and different types of relations. Thus, an

*Corresponding author: Kun Yue, School of Information Science and Engineering, Yunnan University, China. Email: kyue@ynu.edu.cn

in-depth understanding of domain knowledge-oriented models and techniques is necessary for developing accurate and effective RE methods. Besides, the quality of domain-specific data and the external knowledge are also the significant aspects relevant to domain RE.

1.2. Challenges

There are several key challenges of RE in terms of domain knowledge.

- **Domain-specific knowledge.** Developing effective RE models requires a deep understanding of the domain-specific knowledge, and many domains have specialized terminologies and jargons that may not be familiar to those outside the field. It is difficult for ML models to accurately identify and extract relations between entities, for which the collaboration with domain experts and development of customized models are probably required.
- **Complex relations and difficulties in data annotation.** In some domains, the relations between entities may be complex and multifaceted. For example, in the biomedical domain, a gene may have multiple interactions with other genes and proteins, each of which is associated with different functional implications. Obtaining a large amount of annotated data is another difficult task w.r.t. domain knowledge, which limits the performance of supervised learning methods.
- **Data sparsity and quality.** In many domains, there may be limited annotated data available for training ML models, which makes it difficult to develop accurate and effective models for RE. The quality of data can also be a challenge, especially in the domains whose data are unstructured or come from a variety of sources, leading to errors and inaccuracies in the extracted relations.
- **Language diversity and ambiguity.** The language used in domain knowledge could be highly diverse and domain specific, which limits the generalization ability of ML models across different domains. Natural language is often ambiguous and could be interpreted in multiple ways. This can make it difficult to accurately identify the relations between entities, especially when dealing with complex or abstract concepts.

1.3. Ideas

The ideas to address the above challenges are listed as follows.

- **Domain knowledge preprocessing and understanding.** One idea to address the domain-specific knowledge is to develop the domain-specific dictionaries or ontologies that capture the relevant terminologies and jargons. These resources could be used to help ML models identify and extract relations between entities accurately. Another idea is to use domain-specific preprocessing techniques, such as terminology normalization or entity linking, to facilitate the ML models better understanding the domain-specific terminologies and jargons.
- **Inferences and automatic generation.** To address the challenge of complex relations, graph-based models could be used to represent entities and their relations as a graph, helping capture the complex and multifaceted relations between entities. The idea to address the difficulties in data annotation is to use active learning, by which the model is trained on a small amount of annotated data and then the most informative examples are selected for annotation by domain experts. Another idea is to use semi-supervision, by which the model is trained on noisy or incomplete labels that are generated from domain-specific rules or heuristics automatically.
- **Pre-trained language model and data curation.** To address the challenge of limited annotated data, one idea is to use transfer learning or pre-trained language models, such as BERT or GPT, that have been trained on large general language corpora. This can help improve the performance of RE models even with limited annotated data. To address the challenge of data quality, careful data curation and preprocessing are essential, which may involve data cleaning and standardization, as well as removing irrelevant or noisy data.
- **Transfer learning and natural language understanding.** One idea to address the language diversity is to use the cross-lingual transfer learning, where models are trained on multilingual data to improve their ability to handle diverse languages. The idea to address the ambiguity is to use context-sensitive models that can incorporate the surrounding text to disambiguate the meaning of words or phrases. Another idea is to use the rule-based methods that incorporate domain-specific knowledge and heuristics to disambiguate and identify relations.

1.4. Models

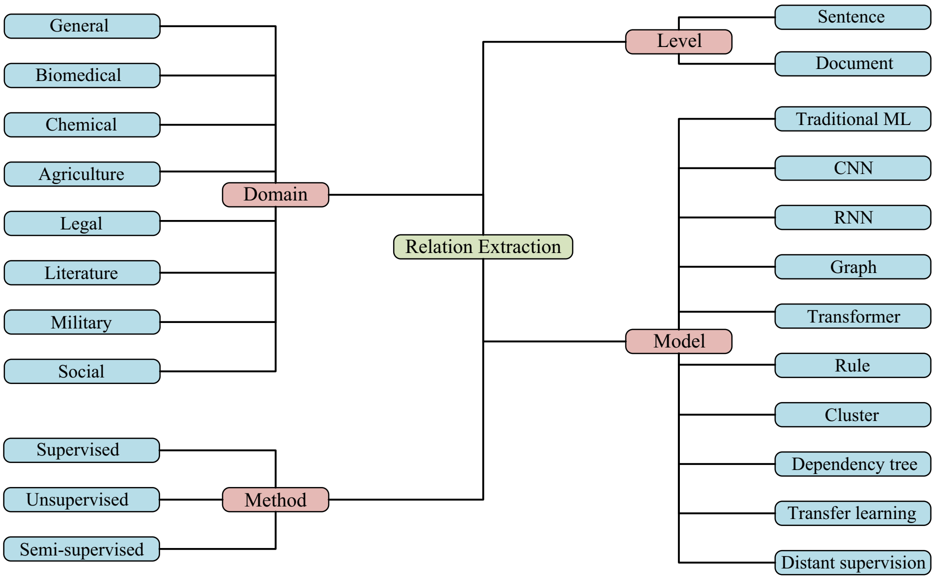
To implement the aforementioned ideas, there are several models commonly used for domain-specific RE.

- **Rule-based models** could be used to incorporate domain-specific knowledge and heuristics into the RE process. These models typically involve the development of a set of rules or patterns that capture the relevant relations between entities (Ravikumar et al., 2017).
- **Convolutional neural network (CNN)**-based models work by using convolutional filters to extract features from the input data, which are then fed into a classifier to predict the relations between entities (Li et al., 2019b).
- **Recurrent neural network (RNN)**-based models process the input data sequentially, one word at a time, and use a hidden state to capture the context and dependencies between words (Geng et al., 2021).
- **Graph-based models** represent entities and their relations as a graph to capture complex and multifaceted relations between entities. These models include the techniques of graph convolutional networks (GCNs) or knowledge graph embeddings (Xue et al., 2021).
- **Transformer-based models**, such as BERT or GPT, have been shown to be highly effective for RE in text data (Gutiérrez et al., 2022; Lai et al., 2022). These models use self-attention mechanisms to capture the context and dependencies between words and could be fine-tuned for specific RE tasks (Xiao et al., 2022).

The choices of RE models depend on the specific characteristics of the domain-specific data, as well as the performance requirements of the domain RE task. In addition, the category of the original data also determines the appropriate models for domain RE tasks. Figure 1 shows an overview of levels, models, and methods of domain RE.

The rest of this paper is organized as follows. Section 2 shows various RE tasks in typical domains and the sentence/document-level tasks. Section 3 introduces sentence-level and document-level RE models and techniques from the aspects of unsupervised, semi-supervised, and supervised methods. Section 4 gives the commonly used RE datasets and evaluation metrics. Section 5 concludes and discusses future research issues.

Figure 1
An overview of levels, models, and methods of domain relation extraction



2. RE Tasks in Typical Domains

Generally, RE tasks involve identifying and extracting semantic relations between entities in unstructured texts, and the specific domains vary widely in various applications. Typical domains where RE tasks are commonly performed include biology, medical, chemical, agriculture, legal, literature, finance, military, and social. The principal difference of these RE tasks lies in the types of entities and relations, as well as the downstream applications as shown in Table 1.

Table 1
RE tasks on various domains

Domain	Type of entities	Type of relations	Downstream applications
Biology	Genes	Protein–protein interactions	Drug discovery
	Proteins	Gene–disease associations	Disease diagnosis
		Drug–target interactions	
Medical	Diseases	Patient–treatment relations	Clinical decision support
	Symptoms	Disease–symptom associations	Pharmacovigilance
	Drugs	Drug–adverse event relations	
	Treatments		
Chemical	Compounds	Compound–synthesis relations	Drug discovery
	Reactions	Compound–properties relations	Material science
	Properties	Drug–target interactions	
Agriculture	Crops	Crop–pest relations	Crop management
	Pests	Crop–growth factors relations	Disease control
	Weather	Soil–plant relations	
Legal	Laws	Defendant–crime relations	Case analysis
	Regulations	Plaintiff–evidence relations	Compliance monitoring
	Cases	Contract–term relations	
Literature	Characters	Character relations	Genre classification
	Settings	Plot–event relations	Authorship attribution
	Themes	Author–influence relations	
Finance	Companies	Company–stock relations	Market prediction
	Stocks	Market–trend relations	Risk management
	Financial events	Investor–portfolio relations	
Military	Weapons	Troop–mission relations	Threat analysis
	Equipment	Equipment–use relations	Mission planning
	Operations	Strategy–tactics relations	
Social	People	Person relationships	Sentiment analysis
	Organizations	Sentiment–opinion relations	Trend prediction
	Events		

The specific tasks and challenges involved in RE may vary widely depending on the original domain-specific data and require specialized models and techniques. Sentence-level RE and document-level RE are two different methods to identify and extract semantic relations between entities in unstructured texts (Han et al., 2020).

The sentence-level RE methods focus on extracting relations between entities within individual sentences, without considering the broader context of the entire document. Sentence-level RE is typically more precise than document-level RE, as it allows for more granular analysis of the relations between entities within each sentence.

The document-level RE method, on the other hand, involves identifying and extracting relations between entities across an entire document. This kind of method considers the broader context of the document, including the relations between entities across multiple sentences. Document-level RE is typically less precise than sentence-level RE, as it involves analyzing a larger amount of data and may include noise from the irrelevant text. The key to improve the performance of document-level RE is to design and adopt efficient inference methods (Song et al., 2022; Wang et al., 2022b).

Let $D = \{S_j | 1 \leq j \leq m\}$ denote the documents, and $S_j = \{w_i | 1 \leq i \leq n_j\}$ represent a sentence as the set of words, where w_i is the i-th word, and m and n_j represent the number of the sentences and that of words, respectively. For ease of expression, we use $\mathcal{S}$ to denote the set of sentences ($S_j \in \mathcal{S}$), and $\pi = \{< h, r, t >\}$ to denote the set of triplets, where $< h, r, t >$ is a triplet, and h, r, and t represent head entity, relation, and tail entity, respectively. Sentence-level and document-level RE methods are to extract the set of triplets representing entities and relations, denoted as $f_S : S \rightarrow \pi$ and $f_D : D \rightarrow \pi$, respectively.

Figure 2 shows the two RE tasks, which have their own advantages and disadvantages. The choice of the method depends on the specific task and goals of the analysis. Sentence-level RE is useful for the tasks that require a more granular understanding of the relations between entities within individual sentences, such as question answering and information retrieval. Document-level RE is useful for the tasks that require a broader understanding of the relations between entities across an entire document, such as document classification and summarization.

In some cases, a combination of both methods may be used to achieve the improved results. For example, document-level RE may be used to identify the relations between entities across a document, and then sentence-level RE may be used to further analyze the relations within individual sentences.

3. Domain RE Methods

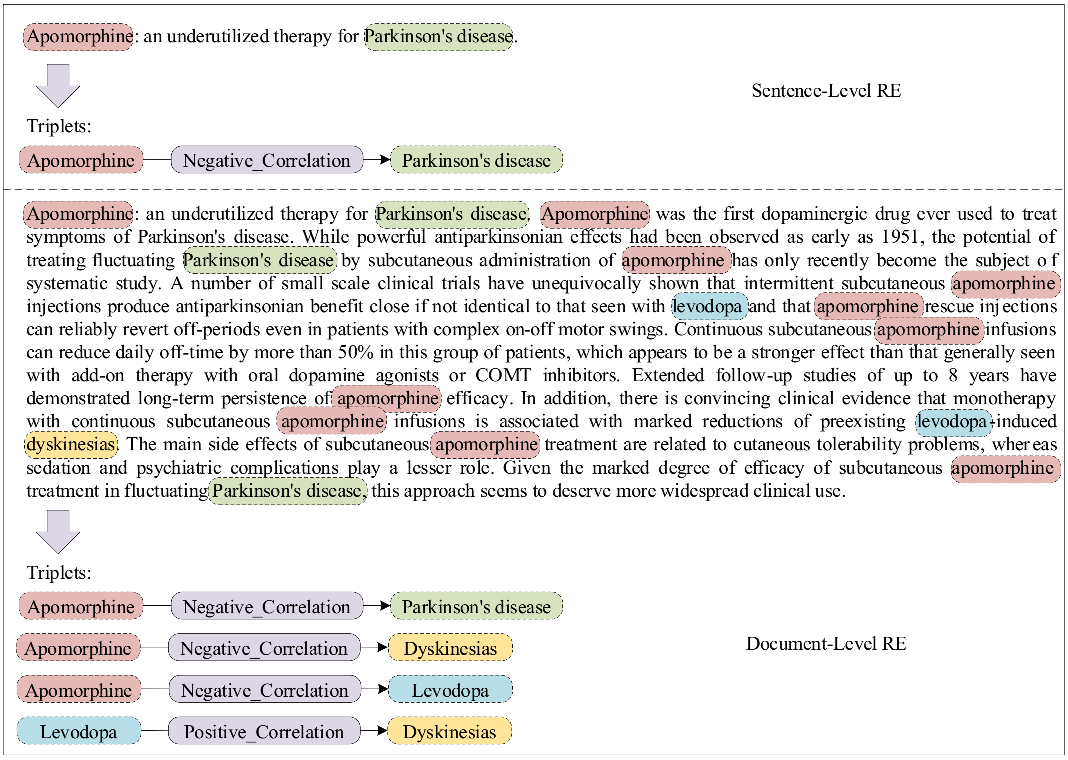
RE methods of domain knowledge are roughly categorized into three groups: *supervised*, *unsupervised*, and *semi-supervised*. For each category, various models in *sentence level* and *document level* are stated separately in the following subsections.

3.1. Supervised methods

3.1.1. Overview

Supervised domain RE involves using supervised learning algorithms to identify and extract relations between entities in a specific field of knowledge.

Figure 2
Tasks of sentence/document-level RE (BioRED)



Apomorphine: an underutilized therapy for Parkinson's disease.

Apomorphine: an underutilized therapy for Parkinson's disease Apomorphine was the first dopaminergic drug ever used to treat symptoms of Parkinson's disease. While powerful antiparkinsonian effects had been observed as early as 1951, the potential of treating fluctuating Parkinson's disease by subcutaneous administration of apomorphine has only recently become the subject o f systematic study. A number of small scale clinical trials have unequivocally shown that intermittent subcutaneous apomorphine injections produce antiparkinsonian benefit close if not identical to that seen with levodopa and that apomorphine rescue inj ections can reliably revert off-periods even in patients with complex on-off motor swings. Continuous subcutaneous apomorphine infusions can reduce daily off-time by more than 50% in this group of patients, which appears to be a stronger effect than that generally seen with add-on therapy with oral dopamine agonists or COMT inhibitors. Extended follow-up studies of up to 8 years have demonstrated long-term persistence of apomorphine efficacy. In addition, there is convincing clinical evidence that monotherapy with continuous subcutaneous apomorphine infusions is associated with marked reductions of preexisting levodopa-induced dyskinesias. The main side effects of subcutaneous apomorphine treatment are related to cutaneous tolerability problems, wher eas sedation and psychiatric complications play a lesser role. Given the marked degree of efficacy of subcutaneous apomorphine treatment in fluctuating Parkinson's disease, this approach seems to deserve more widespread clinical use.

Traditional ML methods could be used to classify relations between entities by analyzing various features in the data. Supervised learning models are trained on annotated data and then used to predict the relation types in the test data. Feature vector-based methods and kernel function-based methods are the two predominant methods used for this purpose. Feature vector-based methods involve extracting feature vectors from key elements such as context information, part of speech and syntax, and then using classification algorithms like Naive Bayes, support vector machines (SVMs), and maximum entropy model to predict relation categories (Minard et al., 2011; Sureshkumar & Zayaraz, 2015). On the other hand, kernel-based methods train classification models by calculating the similarity between entities, in which a designed kernel function (Nguyen et al., 2015; Zhou et al., 2010) is used. The accuracy of supervised learning methods is directly related to the quantity and quality of annotated data available in the corpus, which limits their ability to identify new relations.

In recent years, there has been significant progress in supervised RE based on neural networks (NNs), with the development of deep learning models such as CNN and RNN. In supervised deep learning, the RE task is viewed as a classification problem. To accomplish this, informative features are designed based on the training data, and a range of classification models are learned. Trained classifiers are then adopted to predict relations, which can be achieved through either the pipeline or joint learning method. The pipeline method involves performing named entity recognition (NER) and RE tasks in series (Cai et al., 2016). Firstly, RE is carried out on sentences containing marked entity pairs, and then entities with relations are combined into triples and output as the prediction result. In this group of methods, NER and RE tasks are performed separately, and the performance of RE is easily affected by the results of NER. The joint learning method (Li et al., 2017; Zheng et al., 2017) involves identifying and extracting multiple types of relations between entities simultaneously. For both methods, the basic vector representation of characters is obtained initially, and then sentence features are extracted using different NN models. Finally, nonlinear classifiers are used to complete relation classification.

Figure 3 shows the framework of supervised sentence-level and document-level RE, followed by the brief introduction to the procedure.

Given a sentence $S = \left\{w_1, w_2, \cdots, w_{n_j}\right\}$, the framework first obtains the basic corresponding representations $\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{n'}\}$ of all characters in the sentence by the pre-representation model. Then, the representation vectors in the sentence are the input of the NNs to learn deeper features, such as context and mutual relations, and obtain better embeddings as follows:

$$\mathbf{b}_i = \mathrm{NN}(\mathbf{e}_i)\left(1 \le i \le n_j\right) \tag{1}$$

Then, the sentence S could be represented by the matrix $\mathbf{S} = \left[\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_{n_j}\right]$, where $\mathbf{S} \in \mathbb{R}^{d \times n'}$ and d is the dimension of the embeddings. Then, the framework generally uses an attention layer to automatically capture the semantic information in S. The final representation of the sentence is as follows:

$$\mathbf{S}' = \tanh(\mathbf{S}) \tag{2}$$

$$\boldsymbol{\alpha} = \mathrm{Softmax}(\boldsymbol{\rho}^{\mathrm{T}}\mathbf{S}') \tag{3}$$

$$\mathbf{s}^* = \tanh(\mathbf{S}\boldsymbol{\alpha}^{\mathrm{T}}) \tag{4}$$

where $\boldsymbol{\rho}$ is the training parameter vector with dimension d, $\boldsymbol{\alpha}$ is the attention weight parameter vector with dimension n_j, and $\mathbf{s}^*$ is the vector representation of the sentence obtained upon the sum of attention weights.

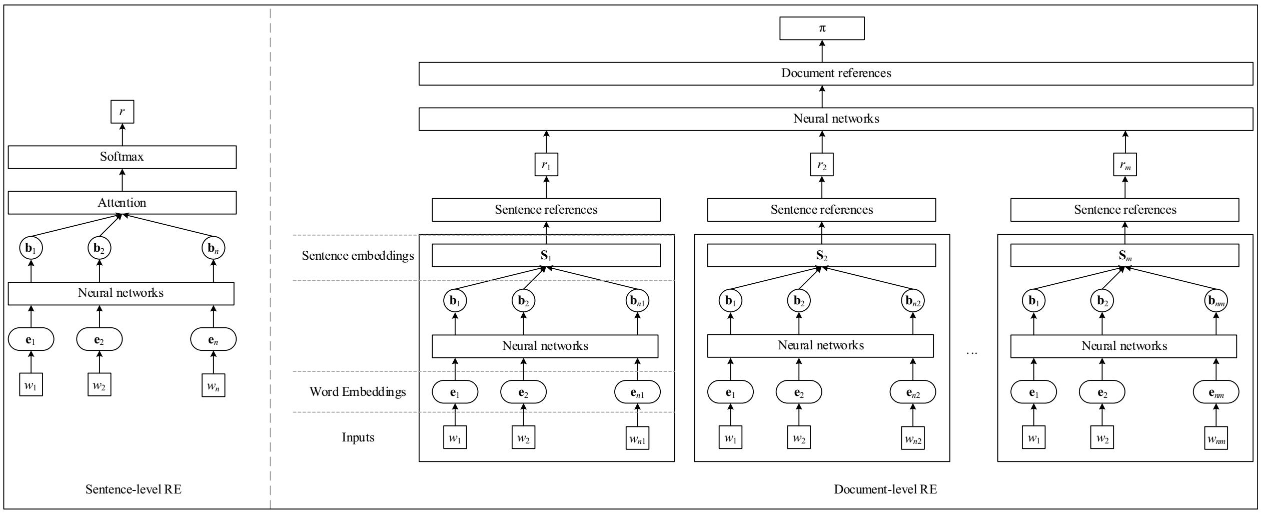
Finally, the model uses a Softmax classifier to calculate the probabilities of various types of relationships that may exist between entity pairs in the sentence and outputs the predicted relationship category corresponding to the maximum probability.

$$P(r|S) = \mathrm{Softmax}\left(\mathbf{W}^{(S)}\mathbf{s}^* + b^{(S)}\right) \tag{5}$$

$$\hat{r} = \arg\ \max_r P(r|S) \tag{6}$$

Given a document $D = \{S_1, S_2, \cdots, S_m\}$, the representations of entities, relations, and sentences are first obtained by the sum or average functions based on equations (1)–(4). Then, the other embedding models like GCN are adopted to capture the global information in documents and other inference methods are adopted to predict the final relations among entities.

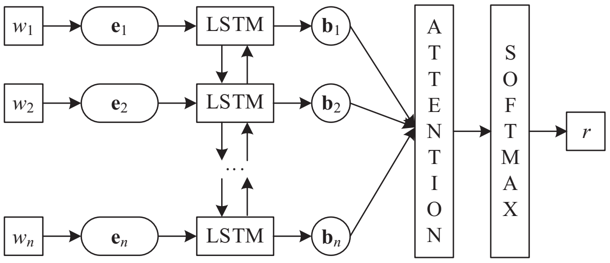
Figure 3
Framework of supervised sentence/document-level RE



According to different models used in the domain-specific RE tasks, supervised methods could be categorized into four classes and each of them could be extended by introducing external knowledge.

(1) **The CNN-based model** is divided into a feature extraction layer and a feature mapping layer. The input of each neuron in the feature extraction layer is connected to the local acceptance domain of the previous layer, which extracts local features. The feature mapping layer is composed of multiple feature mappings, in which all neurons have the same weight, reducing the number of free parameters in the network and enabling parallel learning. Li & Mao (2019) propose a novel method for causal RE from natural language texts and leverage a knowledge-oriented CNN to capture the semantic and syntactic features. This method leverages external knowledge sources, such as WordNet and ConceptNet, to enrich the representation of words and phrases in the text and introduces a causal attention mechanism to weight the importance of different parts.

Figure 4
BiLSTM-based RE



(2) **The RNN-based model** contains both internal feedback and feedforward connections, which could process the sequence information of any time sequence and learn the combination vector representation of various phrases and sentences of any length. However, this method is susceptible to the problems of gradient disappearance and gradient explosion and has long training time. Based on the RNN model, the long short-term memory (LSTM)-based model proposes the gating mechanism and cell state to solve the long-term dependence between characters. Figure 4 shows an example of the bidirectional long short-term memory (BiLSTM)-based RE.

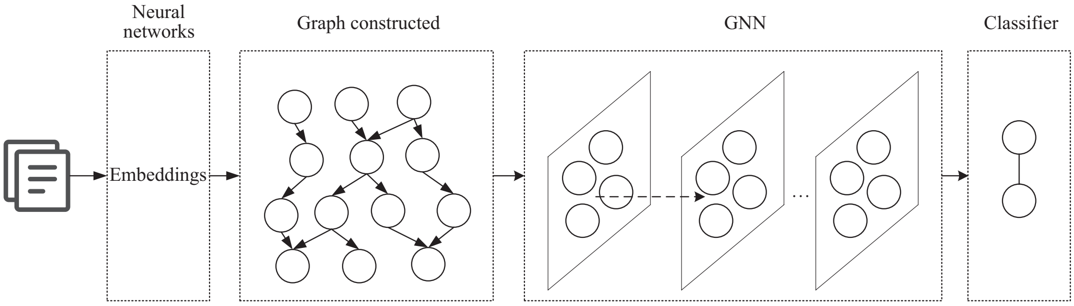
Yuan et al. (2020) propose a novel method for joint entity and RE that leverages a relation-specific attention network to capture the relation-specific information in the text and improve the performance of the model of joint extraction of entities and relations. This method involves using a multi-task learning framework that jointly learns the entity and RE tasks while leveraging the relation-specific attention mechanism to evaluate the importance of different parts of the input.

(3) **The graph-based model** represents the text as a graph, with entities as nodes and relations as edges connecting the nodes. The graph is then analyzed to identify the relations between the entities. One way to implement the graph-based RE is to make use of dependency parsing, which involves identifying the syntactic dependencies between words in sentences or documents. The dependencies are then used to construct a graph, with the words represented as nodes and the dependencies between them represented as edges. The entities in the sentence could be identified by selecting the nodes corresponding to nouns or named entities. Once the graph is constructed, graph-based algorithms could be used to identify the relations between entities. Alternatively, the algorithms including community detection, nodes clustering, and so on could be used to identify the groups of entities that are strongly connected, which may indicate a relation between them. Finally, the classifiers or links cloud be adopted to predict the real relations and acquire the triplets. Figure 5 shows the graph-based RE.

(4) **The transformer-based model** uses a self-attention mechanism to capture the dependencies between words in a sentence and encode the sentence into a fixed-length vector representation. In this class of methods, the input text is first tokenized and then passed through the transformer-based model, which then uses the encoded representation of the text to predict the relation between entities in the text. Cabot and Navigli (2021) propose the method to generate natural language descriptions of the relations between entities in the text. This method involves training a NN to generate a sequence of words that describe the relations between entity pairs, given their context in the input text.

Apart from algorithmic methods, external resources, such as knowledge graphs and ontologies, have been investigated to enhance the accuracy and interpretability of RE models. These resources could assist in guiding the model's predictions and providing supplementary context for RE.

Figure 5
Graph-based RE



3.1.2. Sentence level

In recent years, there has been an increasing interest in developing the methods for automatically extracting relations from unstructured sentences in various domains including biomedical, agriculture, legal, military, and finance. These methods are expected to enhance decision making, knowledge discovery, and data analysis in these fields. We first review the typical methods proposed for extracting biomedical entity relations at the sentence level. These methods use a variety of techniques, including traditional ML, CNN, RNN, and transformer-based models. Then, we describe the key features of each method, including the techniques used for entity recognition, feature extraction, and relation classification. We also highlight the strengths and limitations of each method, as well as the potential areas for future research. Overall, these methods demonstrate the versatility and applicability of entity and RE across diverse domains.

3.1.2.1. Biomedical

Traditional ML-based model. Chen et al. (2021) present the method for extracting chemical protein relations from biomedical literatures. This method first adopts a dependency parser to parse the sentence into a dependency tree and then extracts the shortest dependency path between the chemical and protein entities in the tree. A set of features are extracted from the shortest dependency path, including lexical, syntactic, and semantic features. Then, an ensemble learning method is used to improve the performance of RE.

CNN- and RNN-based model. Zhang et al. (2018) present a hybrid model for extracting biomedical relations from unstructured texts, which consists of three components, a CNN for sentence-level feature extraction, a BiLSTM for capturing contextual information, and a SVM classifier for predicting relation types. The CNN model is used to extract features from the input sentences, while the BiLSTM is used to capture the important contextual information for the RE task. The SVM classifier is trained on the output of the BiLSTM to predict the relation types.

Transformer-based model. Zuo and Zhang (2022) present the method for jointly extracting entities and relations of bacteria biotopes by using a pre-trained language model to encode the input text and extract features and predict the start and end positions of the entity spans, as well as their relation labels. The method also incorporates external knowledge from a domain-specific ontology to improve the performance of the entity and RE tasks. Ding et al. (2021) propose a system for extracting biomedical entity relations from Chinese language texts by using a pre-trained language model to encode the input sentences and extract features. Then, the BiLSTM is used to classify the relations between biomedical entities.

3.1.2.2. Agriculture

Traditional ML-based model. Liu et al. (2018) propose the method for extracting relations in the agriculture domain with both syntactic and semantic features by using a combination of rule-based and ML methods. First, a set of domain-specific rules are defined to identify relevant sentences for RE. Then, a feature selection technique is used to select the most relevant syntactic and semantic features for RE. Finally, a SVM classifier is used to classify the relations between entities in the identified sentences. Yang et al. (2021) propose the method for RE in the fishery domain using a dual attention mechanism that incorporates both contextual and semantic information by using a combination of rule-based and ML methods for entity recognition and RE.

Transformer-based model. Yuan et al. (2021) propose the method for RE from a rice phenotype knowledge graph using BERT. The method adopts a pre-trained BERT model to encode the input sentences and generate contextualized word representations. The fine-tuned BERT model is then used to classify the relations between entities in the input sentences. Besides, domain-specific knowledge from a rice phenotype knowledge graph is incorporated to improve the performance of RE.

3.1.2.3. Legal

Transformer-based model. Hong et al. (2020) propose the method of knowledge graph construction for judicial case facts by using a combination of rule and ML-based NER and RE. A set of domain-specific rules are defined to identify relevant entities and their relations based on their embeddings in judicial case facts. Then, the ML model is used to classify the relations between the identified entities.

3.1.2.4. Military

Traditional ML-based model. Liang et al. (2018) propose the method for extracting entities and relations in the military field by using a combination of rule and ML-based methods for NER and RE. A set of domain-specific rules is defined to identify entities and their relations, and then a SVM classifier is used to classify the relations between the identified entities.

RNN-based model. Wang et al. (2018a) propose the method for extracting relations in the military domain by using a BiLSTM to classify the relations between embeddings of entities in the input sentences and incorporating external knowledge from a military knowledge graph to improve the performance of the RE task.

Transformer-based model. Lu et al. (2021a) propose the method for detecting military events using a combination of BERT, BiGRU, and attention mechanisms. The method adopts a pre-trained BERT model to encode the input sentences and extract features. Then, a bidirectional gated recurrent unit network is used to capture the contextual information of the input sentences, and the attention mechanism is then used to focus on the most relevant parts of the input sentences for event detection. Lu et al. (2021b) propose a military RE model using a combination of BiGRU and multi-head attention mechanisms. The method adopts a BiGRU-based NN to encode the input sentences and generate contextualized word representations. The multi-head attention mechanism is then used to capture the relations between different entities in the input sentences.

3.1.2.5. Finance

CNN-based model. Sun et al. (2018) present the method for extracting relations from Chinese financial texts to construct an enterprise knowledge graph. First, entities are recognized by using a rule-based method upon a dictionary and a set of rules. Then, a CNN that is trained on a large-scale corpus with manually annotated data is used to classify the relations between the identified entities based on the maximum entropy.

RNN-based model. Yan et al. (2019) propose a NN-based method for extracting relations between enterprises in credit risk management by using a BiLSTM to encode the input sentences and extract features. The method also incorporates external knowledge from a domain-specific ontology to improve the performance of the RE task.

Graph-based model. Fu et al. (2021) present a method for jointly extracting entities and relations to construct a domain knowledge graph. The method first adopts a representation model to encode the input sentences and then uses a multi-head attention mechanism to capture the interactions between entities and relations. Then, a GCN is used to encode the graph structure and entities, and a multi-task learning method is adopted to jointly train the NER and RE tasks.

Transformer-based model. Lai et al. (2022) propose a multi-modal RE model for the internet security of finance in Chinese. The method uses a combination of text, image, and audio modalities to extract relations between financial entities. A pre-trained language model is adopted to encode the input text, and a CNN is adopted to extract features from the input images. A RNN is used to extract features from the input audio, and a multi-task learning method is adopted to jointly train the RE tasks for each modality.

3.1.3. Document level

There are various methods proposed for document-level RE that leverage different techniques to capture the complex dependencies between entities in a document (Huang et al., 2021). Most of these methods are established on the basis of the graph models and their extension and highlight the importance by considering the complex relations between entities in a document to improve the performance of RE.

Christopoulou et al. (2019) propose a novel method for document-level RE that leverages edge-oriented graphs to capture the complex relations between entities in a document. Tang et al. (2020) propose a novel method for document-level RE that leverages a hierarchical inference network model for the global dependencies between entities in a document. Nan et al. (2020) propose the method for document-level RE that leverages a latent structure refinement mechanism by using a GCN to predict the existence of relations between entities in a document and then refines the predicted relations by iteratively reasoning over the latent structure of the document. Xu et al. (2022) propose a multi-level attention mechanism that attends to different levels of evidence, such as sentences, words, and entities in the input text. The method involves using a hierarchical attention network to attend to the relevant evidence and predict the relation between entities.

3.1.3.1. Biomedical

CNN-based model. Bai et al. (2022) propose the method for extracting entity relations in traditional Chinese medicine using a CNN with segment attention. The method first splits the input text into segments and then uses a pre-trained language model to encode each segment and extract features. The CNN with segment attention is used to classify the relations between entities in each segment. The segment attention mechanism is used to focus on the important segments for RE.

RNN-based model. Peng et al. (2017) propose a general framework for RE that could be extended to extract the *n*-ary relations spanning multiple sentences. The proposed framework is based on graph LSTM networks and provides a unified way of exploring different LSTM methods and incorporating various intra-sentential and inter-sentential dependencies, such as sequential, syntactic, and discourse relations.

Graph-based model. Sahu et al. (2019) propose a novel method for inter-sentence RE. In this method, a document is represented as a graph, where each sentence is represented as a node and the relations between sentences are represented as edges. The model consists of two main components, a graph encoder that encodes the graph structure and sentence embeddings, and a relation classifier that predicts relation types between pairs of entities. Li et al. (2019b) propose a novel method for RE in clinical narratives by using segment graph CNN and RNN. This method involves representing each sentence in the clinical narrative as a segment graph, where medical concepts are represented as nodes and relations are represented as edges. Wan et al. (2016) construct heterogeneous entity networks from the traditional Chinsese medical literature, in which each edge is a candidate relation, and then use a heterogeneous factor graph model to simultaneously infer the existence of all the edges.

Transformer-based model. Verga et al. (2018) propose a novel NN-based method for extracting biological relations from unstructured texts. The authors argue that previous methods of RE have focused on identifying relations between specific entities, while ignoring the larger context in which the entities are mentioned. The proposed self-attentive mention-pair model is based on self-attention mechanisms that allow the model to simultaneously attend to all mentions in a document. Li et al. (2022) propose the method for extracting chemical-induced disease relations at the document level using cross self-attention by using a pre-trained language model to encode the input sentences and extract features. In this method, a cross self-attention mechanism is used to capture the inter-sentence dependencies and relations, and external knowledge is incorporated from a biomedical knowledge graph to improve the performance of RE.

3.1.3.2. Agriculture

Traditional ML-based model. Kaushik and Chatterjee (2016) propose a practical method for term and RE for automatic ontology creation from agricultural texts. The method uses a combination of rule-based and statistical methods for term and RE. Domain-specific rules are first defined to identify relevant terms and relations, and a statistical model is used to extract additional terms and relations from the identified texts.

3.1.3.3. Literature

Traditional ML-based model. Zhang et al. (2019) propose the method to identify the relations between protagonists based on their interactions and emotions. The method consists of four main steps, data preprocessing, entity extraction, relation identification, and relation analysis. In the data preprocessing step, the fictions are cleaned and segmented into chapters. In the entity extraction step, the protagonists and their love interests are identified using a NER tool. In the relation identification step, the relations between the protagonists are identified upon their interactions and emotions using a rule-based method. In the relation analysis step, the love relations between protagonists are analyzed and visualized using network analysis techniques.

3.1.3.4. Legal

Traditional ML-based model. Chen et al. (2020) propose the method for joint entity and RE in legal documents with legal feature enhancement by using a combination of rule and ML-based methods for entity recognition and RE. Domain-specific rules are first defined to identify relevant entities and their relations in legal documents.

Then, a feature enhancement technique is used to incorporate legal features, such as legal keywords and legal context, to improve the performance of RE. Thomas and Sivanesan (2022) propose an adaptable and high-performance system for RE in complex sentences by using a hybrid method. A set of domain-specific rules are defined to identify relevant entities and their relations in complex sentences. Then, a ML model is used to classify the relations between the identified entities, and the novel method is introduced for selecting the most informative words from the input sentences to improve the accuracy of RE.

CNN-based model. Gao et al. (2018) propose the method for RE in large-scale legal text data by using the CNN to extract features from the input sentences. In this method, the external knowledge from an ontology is also incorporated to improve the performance.

3.1.3.5. Finance

Graph-based model. Wan et al. (2023) present the method for extracting multi-type financial events and relations from Chinese text. First, a set of financial event trigger words and patterns are defined based on domain knowledge, and then a representation model is used to encode the input sentences and extract features. A GCN is used to encode the dependency graph structure of the sentences, and a multi-task learning method is adopted to jointly train the event trigger and RE.

Transformer-based model. Zhou et al. (2022) propose a multi-RE method for unstructured financial reports based on multi-task learning and pre-trained language models to address the difficulty of multi-RE in financial reports. The method first uses the pre-trained language model BERT for representation learning of financial reports and then adopts a multi-task learning method to jointly train multiple RE tasks to improve the accuracy and generalization. At the same time, the method also uses an attention mechanism based on sentence level and document level to improve the model's ability to capture the correlation between different sentences in RE.

In summary, traditional ML-based RE methods are flexible, scalable, and easy to generalize but may be prone to overfitting and have lower accuracy. CNN-based RE has the advantages of capturing local features, being computationally efficient and scalable for large dataset-based real-time applications, but may have limited context information and difficulty with variable-length inputs. RNN-based RE has the advantages of capturing contextual information and flexibility with variable-length inputs but may have difficulty with vanishing and exploding gradients and be computationally intensive. Graph-based RE has the advantages of capturing global dependencies, having flexible inputs, and being intuitive and interpretable but may have high computational complexity and be dependent on graph construction. Transformer-based RE has the advantages of capturing language features but may have a large number of parameters, making it computationally intensive.

3.2. Unsupervised methods

When working with large-scale corpora, it becomes impractical to predict all possible relations between entities due to the vast amount of annotated corpus and preprocessing work required, which may be time-consuming. To address the limitations of supervised and semi-supervised RE methods that rely on deep learning, some researchers have started exploring unsupervised ML techniques for extracting entities and relations (Tran et al., 2020).

Unsupervised methods for RE could be divided into two main categories, rule/pattern-based and clustering-based methods. In the rule-based method (Qin et al., 2015), artificial rules are designed to match text, which could be achieved using either a pattern-based or a dependency tree-based method. The pattern-based method (Fang & Chang, 2011) requires manual creation of rule templates, which are then used to identify instances in the texts that satisfy the rule and form a triplet consisting of the entities and corresponding relations. The dependency tree-based method leverages semantic dependency and dependency syntactic parsing to analyze the sentence structure and capture deep semantic information between different parts of speech words, such as nouns and verbs. This method identifies the structure within the sentence that meets specific rules, such as subject-verb-object, verb complement structure, verb structure, inter-object relation, and object preposition and then extracts corresponding triples. The clustering-based method (Mesquita, 2012) involves grouping similar entities or entity pairs together based on their co-occurrence patterns in a large corpus of texts. This method relies on the assumption that frequently mentioned entities in context are likely to have a relation.

Although the rule-based methods may be accurate, particularly for domain-specific data, and easy to implement on small datasets, they suffer from several shortcomings, such as low recall rates, lack of robustness, reliance on domain expert knowledge, high time and energy consumption, and difficulties in maintaining continuous operations. The clustering-based methods may fail to capture the intricacies of complex relations, and the quality of extracted relations may be influenced by the quality of the clustering algorithm and the concerned features. Moreover, the clustering-based methods could be computationally demanding, particularly when working with large datasets.

The introduction of document-level RE tasks has shifted the focus away from analyzing sentence structures and features, toward understanding the associations and propagation between sentences within a given document. Unsupervised methods are not constrained by the length or format of input texts, making them suitable for a range of RE tasks, including those at the sentence and document levels, as well as across different data formats. Potential applications in document-level RE tasks are worth exploring.

Rule-based model. Qian et al. (2008) propose the method of using constituent dependencies between the sub-constituents of a sentence, to extract semantic relations between entities. The method uses a tree kernel-based method to represent the dependencies between sub-constituents and calculate the similarity between sentences. A variety of features including word features and syntactic features are also used to improve the performance. Alicante et al. (2016) present an unsupervised method for extracting entities and relations in Italian clinical records by combining distributional semantics and pattern-based methods. The method involves multiple steps, starting with the use of a set of rules to segment the text into sentences and identify the entities in each sentence. Then, distributional semantics are employed to identify the entities and their types based on the co-occurrence patterns of words in the text. Finally, pattern-based methods are adopted to identify the relations between entities based on the syntactic and semantic patterns.

Clustering-based model. Kawashima et al. (2017) propose the method for RE of breast cancer and related genes using text mining and pattern clustering and propose a two-step method that involves text mining and pattern clustering. Text mining is first used to identify the relevant sentences from the literature, and the clustering algorithm is then used to group the sentences based on their syntactic and semantic similarities. Gonzalez and Turmo (2009) propose an unsupervised method for RE that uses massive clustering to group similar entities and identify their relations. A named entity recognizer is used to extract entities from the text and represent the documents using a bag-of-words model, and a massive clustering algorithm is used to group similar entities and identify the relations based on their co-occurrence patterns in the clusters. Plank and Moschitti (2013) propose the method that leverages semantic similarity between the source and target domains to adapt the RE model. The method involves embedding semantic similarity into tree kernels, which are then used to compare the syntactic structures of sentences in the source and target domains. Eichler et al. (2008) present a method for identifying entity relations by clustering documents based on their content and co-occurrence patterns. The method involves multiple steps, starting with the use of a set of rules to clean and filter web documents. Then, a part-of-speech tagger and a named entity recognizer are employed to identify the entities in the documents. Finally, a clustering algorithm is adopted to group the documents based on their contents, and then the relations are identified from their co-occurrence patterns in the clusters.

Dependency tree-based model. Miwa et al. (2010) compare and evaluate different dependency representations for relation and event extraction, including basic dependencies, collapsed dependencies, and enhanced dependencies. The method is proposed for integrating event-specific information into the dependency representations to improve the performance of relation and event extraction. Panyam et al. (2018) propose the method by using a graph-based representation of sentences and graph kernels to compare the structures of different sentences and identify the relations. The method involves extracting entities and their features from the sentences and constructing a graph representation, where each node corresponds to an entity and the edges correspond to syntactic and semantic relations.

The rule-based RE has the advantages of being specific, explainable, accurate, and requiring low resources but may be hard to scale and require manual maintenance. Thus, this class of methods has limited generalization and is sensitive to noise. Clustering-based methods could be useful for unsupervised RE by grouping similar entities and identifying their relations based on co-occurrence patterns. However, the quality of the results may depend on the quality of the entity extraction and the adopted clustering algorithms, and the method may not be effective for extracting more complex relations. Dependency tree-based methods can capture the syntactic and semantic dependencies between entities in a sentence, which will be useful for identifying complex relations. However, selecting the appropriate dependency representation and designing effective models are still challenging, and the models may be difficult to interpret.

3.3. Semi-supervised methods

3.3.1. Overview

Semi-supervised RE methods were proposed to address the challenge of limited annotated data (Lin et al., 2017), which are expensive and time-consuming to create and may not always be available for a specific domain or language. Supervised RE methods require a large amount of annotated data to train the accurate models, which may not be feasible in many scenarios. On the other hand, unsupervised methods do not require annotated data but may not be as accurate as the supervised methods. Semi-supervised methods aim to leverage a combination of annotated and unannotated data to train more accurate models with fewer annotated examples. By using the available unannotated data, semi-supervised methods can improve the generalization and robustness of the models, while reducing the reliance on large amounts of annotated data. Therefore, the semi-supervised RE methods are proposed as a compromise between the accuracy of the supervised methods and the data requirement of the unsupervised methods.

The basic idea behind the semi-supervised RE is to use the annotated data to train a base model and then use the unannotated data to refine the model. One common method is to use a combination of the unsupervised learning and transfer learning techniques to improve the performance of the base model. Transfer learning-based semi-supervised learning involves using a pre-trained model on a source domain to extract relations from the target domain. The pre-trained model is then fine-tuned on a small amount of annotated data from the target domain. This method is particularly useful when the annotated data in the target domain are scarce or non-existent. Another method is to use distant supervision by co-training, self-training, or active learning, which involves iteratively selecting the most informative examples from the unannotated data and adding them to the annotated data for retraining the model. This can facilitate to reduce the annotation cost and improve the overall performance of RE.

Transfer learning-based method. Jiang (2009) proposes the method by using transfer learning to transfer knowledge learned from related tasks to improve the performance RE. The method involves training a multi-task learning model that simultaneously performs multiple related tasks, such as NER or part-of-speech tagging, and then transferring the learned knowledge to the RE task. Di et al. (2019) propose the method by using transfer learning to transfer knowledge learned from a source domain to a target domain. The method involves training a model on the source domain, and then fine-tuning the model on a small amount of annotated data from the target domain. Then, a domain-aware regularization term is introduced to make the model learn domain-specific features.

Distant supervision-based method. Abad et al. (2017) propose the method by using the self-crowdsourcing training to automatically generate training data from unannotated text. The method involves iteratively selecting the most informative examples from the unannotated data that are further used to generate the training data. The self-crowdsourcing module that uses syntactic and semantic features is introduced to generate candidate relations. He et al. (2020) propose the method by using the multi-level distant supervision to improve the quality of the training data. The method involves using multiple knowledge bases at different levels of abstraction to generate training data and then bootstrapping the model to improve the quality of the training data. Zhang et al. (2022) propose the method by using the co-training with validation to iteratively select the most informative examples from the unannotated data and add them to the annotated data for retraining the model. The method involves training two independent models on different feature sets and validating the predicted labels on the unannotated data to select the most confident examples for label propagation.

3.3.2. Sentence level

3.3.1.1. Biomedical

Transfer learning-based model. Sun et al. (2021) propose the method by using the transfer learning to transfer knowledge learned from a source domain to a target domain. The method involves training a model on the source domain and then fine-tuning the model on a small amount of annotated data from the target domain. A multi-task learning framework is introduced to jointly learn multiple types of microbial relations. Fabregat et al. (2023) propose the method that incorporates negation cues into the transfer learning process by using a negation-aware representation learning algorithm. The method involves training a model on a source domain and then fine-tuning the model on a small amount of annotated data from the target domain, while incorporating negation cues into the learning process.

Distant supervision-based model. Xiao et al. (2022) propose a hybrid attention-based transformer block model for distant supervision RE. The method uses a pre-trained language model to encode the input sentences and extract features. A hybrid attention mechanism is used to combine both positional and attention-based information in the model. The transformer block model is used to capture the contextual information of the input sentences and predict the relations between the entities. The method also uses a distant supervision method, where entity pairs are automatically annotated based on their co-occurrence in a knowledge base.

3.3.1.2. Literature

Distant supervision-based model. Huang et al. (2018) propose a multi-factor model for extracting person entities and relations based on the distant supervision by using a combination of rule-based and ML methods for RE. Domain-specific rules are first defined to identify relevant sentences, and then a ML model is used to classify the relations between person entities in the identified sentences. Multiple factors, such as syntax, semantics, and domain knowledge, are also incorporated to improve the performance of RE.

3.3.2. Document level

3.3.2.1. Biomedical

Distant supervision-based model. Li et al. (2018) propose the method for extracting chemical-induced disease relations from biomedical literature. First, a set of trigger words and patterns are defined based on domain knowledge to identify relevant sentences. Then, a CNN is used to classify the relations between chemical and disease entities in the identified sentences. The method also incorporates external knowledge from a biomedical knowledge graph to improve the performance of RE.

3.3.2.2. Legal

Distant supervision-based model. Okamoto et al. (2017) propose the method by using information extraction to automatically identify and extract key elements of patent structures, such as title, abstract, claims, and references. The method involves parsing the text and identifying relevant entities and relations.

3.3.2.3. Agriculture

Distant supervision-based model. Veena et al. (2022) propose a semi-supervised method for RE in agriculture documents by using a combination of supervised and unsupervised learning methods. First, a set of seed relations are manually annotated to train a supervised learning model. Then, an unsupervised learning model is used to generate additional candidate relations from unannotated data. A filtering mechanism is used to select the most relevant candidate relations for RE, and a joint learning model is finally used to refine the extracted relations.

3.3.2.4. Social

Distant supervision-based model. Hou et al. (2023) propose the method by using bootstrapping labeling rule discovery to automatically discover labeling rules from the unannotated data to generate annotated data for training the RE model. The method involves iteratively discovering new labeling rules based on the current model's predictions to generate new annotated data for retraining the model.

3.4. Summary

In summary, we conclude all the pros and cons of the aforementioned RE methods and give the corresponding task level, learning category and interpretabilities, shown in Table 2. "X" represents the task that includes traditional ML, CNN, RNN, graph, and transformer-based RE methods.

The rule-based model is advantageous, since it is easily explainable and accurate and requires low resources. However, the scalability and generalization to other domains are still limited, and the model's manual maintenance is required. Traditional ML models have been proposed to address the efficiency problem, as they can be easily defined and applied.

Since the above models suffer from overfitting and low accuracy, the CNN-based models are proposed for efficient and scalable application on large datasets, and the RNN based models are proposed to handle variable-length inputs and capture contextual information. Besides, the graph-based models are proposed to capture global dependencies, especially for document-level RE.

With the great success of large pre-trained language models on various NLP tasks, the transformer-based models have been emerged as a powerful approach for RE due to their ability to handle large amounts of text data and capture complex linguistic patterns. Particularly, the transformer architecture is suited for sequence-to-sequence tasks, and the self-attention mechanism allows the model to focus on relevant parts of the input sequence, considering the dependencies between words and the context. This makes the model more robust to variations in sentence structure and wording and useful to identify the relations that might not be apparent based on the surface-level features.

To facilitate the interpretability of the model, hybrid models like "X+Rule" and "X+External knowledge" are proposed. These models leverage external knowledge sources such as ontologies, databases, or domain-specific knowledge bases to improve their accuracy and coverage. They are particularly useful when dealing with rare or specialized entities or when the domain knowledge is essential for accurate RE. By incorporating external knowledge, these models can improve the accuracy of RE tasks and achieve better generalization performance.

In conclusion, the selection of an appropriate ML model for domain RE depends on the specific requirements of the task, available resources, and desired level of interpretability. Each model has its own set of advantages and disadvantages, and it is essential to carefully evaluate them before choosing the most suitable one. Overall, the success of domain RE tasks heavily relies on selecting the right ML model, and careful consideration of the advantages and disadvantages of each model is necessary for achieving accurate and reliable results.

Table 2
Summary of models

Model	Advantages	Disadvantages	Task level	Learning category	Interpretability
Rule	Specific features Explainable Accurate Low resource required	Hard scalable Manual maintainable Limited generalization	Sentence Document	Unsupervised	Excellent
Traditional ML	Specific features Flexible scalability Easily generalized	Overfitting Low accuracy	Sentence	Supervised	Hard
CNN	Local features Efficient computation Scalable for large datasets	Limited context information Limited input length	Sentence Document	Supervised Semi-supervised	No
RNN	Contextual information Variable-length inputs	Gradient explosion Computationally intensive	Sentence Document	Supervised Semi-supervised	No
Graph	Global dependencies Flexible inputs Interpretable representations	High computational complexity Various graph construction	Sentence Document	Supervised Semi-supervised	Hard
Transformer	Language features	Large number of parameters	Sentence	Supervised Semi-supervised	No
X+Rule	Less data required Auto complex patterns	Hard with complex sentences	Sentence Document	Supervised Semi-supervised	Good
X+External knowledge	Rich information Expert knowledge	Data sparsity Integration complexity	Sentence Document	Supervised Semi-supervised	Good

4. Corpus and Metrics

In this section, we summarize the corpus and metrics of domain RE.

4.1. Corpus

First, the commonly used RE datasets are summarized in Table 3, which present the information on various datasets for the RE tasks, including the number of entities, relations and triplets, the domain, level, language, and links. The datasets cover a wide range of domains, including general, biomedical, literature, finance, and social. They also vary in terms of the number of entities and relations, ranging from a few to millions, as well as in the level of granularity, from sentence level to document level. The datasets are available in English and Chinese languages. Links to the datasets are provided for further access and use. "-" in Table 3 means that we could find the number in their papers or websites.

4.2. Metrics

(1) **Precision** measures the proportion of the identified relations that are correct. It is calculated as the ratio of true positives to the total of true positives and false positives.
(2) **Recall** measures the proportion of actual relations that are correctly identified by the model. It is calculated as the ratio of true positives to the total of true positives and false negatives.
(3) **F1-score** is the harmonic mean of precision and recall. It provides a balanced measure of both precision and recall and is often used as the primary evaluation metric for RE models.

Table 3
Datasets of RE

Datasets	# Entities	# Relations	# Triplets	Domain	Level	Language	Link
ACE 04 (Hachey et al., 2012)	46,108	24	16,771	General	Sentence	English/ Chinese/ Arabic	https://catalog.ldc.upenn.edu/ LDC2005T09
ACE 05 (Hachey et al., 2012)	34,426	6	7105	General	Sentence	English/ Chinese/ Arabic	https://catalog.ldc.upenn.edu/ LDC2006T06
DuIE (Li et al., 2019a)	239,663	49	458,184	General	Sentence	Chinese	https://github.com/PaddlePaddle/ PaddleNLP/tree/develop/ examples/information_ extraction/DuIE
SemEval-2010 Task 8	21,434	9	8853	General	Sentence	English	https://semeval.github.io/ SemEval2021/
TACRED	152,527	41	21,773	General	Sentence	English	https://catalog.ldc.upenn.edu/ LDC2018T24
FewRel	72,124	100	70,000	General	Sentence	English	https://github.com/thunlp/ FewRel

(*Continued*)

Table 3
(*Continued***)**

Datasets	# Entities	# Relations	# Triplets	Domain	Level	Language	Link
NYT	66,194	24	104,339	General	Sentence	English	https://github.com/thunlp/OpenNRE/tree/master/benchmark
WebNLG (Gardent et al., 2017)	6222	171	14,485	General	Sentence	English	https://gitlab.com/shimorina/webnlg-dataset/-/tree/master/
HacRED (Cheng et al., 2021)	9231	26	67,047	General	Document	English	https://github.com/qiaojiim/HacRED.
DocRED (Yao et al., 2019)	2,690,725	96	157,1747	General	Document	English	https://github.com/thunlp/DocRED
TBGA (Marchesin and Silvello, 2022)	20,983	4	218,973	Biomedical	Sentence	English	https://zenodo.org/record/5911097
BioRel (Xing et al., 2020)	69,513	125	534,406	Biomedical	Sentence	English	https://bit.ly/biorel_dataset
CMeIE (Guan et al., 2020)	11	44	85,282	Biomedical	Sentence	Chinese	https://tianchi.aliyun.com/cblue?spm=5176.12282016.0.0.140e7474IE2ln0
Chinese Clinical (He et al., 2017)	34,268	15	7691	Biomedical	Sentence	Chinese	http://github.com/WILAB-HIT/Resources
Drug Combinations (Tiktinsky et al., 2022)	2411	1248	1634	Biomedical	Document	English	https://huggingface.co/datasets/allenai/drug-combo-extraction
BioRED (Luo et al., 2022)	20,419	6503	–	Biomedical	Document	English	https://ftp.ncbi.nlm.nih.gov/pub/lu/BioRED/
BC5CDR	29,271	3106	–	Biomedical	Document	English	https://huggingface.co/datasets/tner/bc5cdr
CodRED (Yao et al., 2021)	–	–	30,504	Biomedical	Document	English	https://github.com/thunlp/CodRED
Biographical (Plum et al., 2022)	–	10	902,103	Literature	Document	English	https://plumaj.github.io/biographical/
Chinese Literature (Xu et al., 2017)	–	9	–	Literature	Document	Chinese	https://github.com/lancopku/Chinese-Literature-NER-RE-Dataset
FinRED (Sharma et al., 2022)	–	29	7775	Finance	Document	English	https://github.com/soummyaah/FinRED
MNRE (Zheng et al., 2021)	20,278	31	10,089	Social	Sentence	English	https://github.com/thecharm/MNRE

(4) **Hit@*k*** measures the percentage of times that the correct relation is among the top *k* predictions made by the model.

(5) **The area under the receiver operating characteristic curve** measures the ability of a model to distinguish between positive and negative relation instances by varying the classification threshold.

(6) **Mean average precision** measures the average precision across all possible thresholds for a given set of relations. It is commonly used for evaluating models that return ranked lists of relations.

5. Conclusions and Discussions

In domain knowledge, RE is a preliminary task to extract structured information from unstructured texts and facilitate various tasks of knowledge discovery or information services pervasively. In this paper, we start from the key challenges and ideas associated with domain RE and highlight the commonly used models. According to the specialties of the domain-specific data and the performance requirements of RE tasks, RE models and techniques could be chosen. We discuss different RE tasks in specific domains and the difference between sentence-level and document-level RE methods, which we categorize into three groups, unsupervised, semi-supervised, and supervised. For each group, we survey the models and techniques of both sentence level and document level detailly. Finally, we summarize the corpus in various domains and the commonly used evaluation metrics.

With the continuous increase of data volume, intelligent analysis applications, as well as the rapid development of artificial intelligence technologies, domain RE plays an increasingly important role, especially in NLP. On the other hand, with the powerful support of rapidly developing artificial intelligence models and big data processing technologies, the research on domain RE has also emerged many new ideas, paradigms, concepts, and frameworks. The continuous evolution of the models and underlying techniques reflects the large-scale and multimodal nature of domain-specific data, as well as the comprehensive and diverse needs of domain RE. The methods of domain RE are developed toward the end-to-end "perception and inference" direction.

From the perspective of systematic methods of domain RE, domain-specific data, domain-oriented model/algorithm, and domain-extended knowledge are centered. Surrounding with these three aspects, more in-depth research directions are worthwhile with theoretical significance and practical values in the future.

(1) **Domain-specific data** provide the fundamental context and information incorporated to accurately identify and extract relationships between entities within a specific domain. To improve the performance upon the domain-specific data, the following issues are valuable.

High-quality corpus. Creating a high-quality corpus is a time-consuming process, but the well-annotated corpus could serve as a fundamental resource for all the models of domain RE. For this purpose, it is necessary to define and collect domain-specific data, identify relevant entities and relations, as well as annotate, evaluate, and refine the corpus. Actually, domain RE results could enrich the creation of domain-specific corpus, and the gradually enriched corpus could strengthen the RE results simultaneously.

RE methods under no/less annotation data. Few-shot RE aims to extract relations between entities using only a small amount of annotated data. Unsupervised methods, which do not require any annotated data, have the potential to automatically discover new types of relations and extract information from large amounts of unstructured data. By incorporating the techniques such as semantic dependency tree with supervised methods, novel information could be extracted. On the other hand, to improve the accuracy of RE models, novel semi-supervised methods should be established upon more useful data by integrating both annotated and unannotated data. These techniques are particularly useful in the domains where annotated data are scarce or expensive to obtain.

(2) **Domain-oriented model/algorithm** provides more accurate and reliable results for domain RE tasks. Focusing on similar tasks in a specific domain, a domain-oriented model/algorithm could leverage the specific characteristics, entities, and relations to improve the accuracy of the prediction results. This could make RE more feasible and effective in a wide range of domains and applications. By focusing on the common problems of RE in different domains, a domain-oriented model can share and learn the associations of similar entities and relations across multiple domains. This helps address the challenge of developing accurate and reliable RE models and consequently improve the performance across multiple domains, which involves the following two aspects:

Multi-task learning involves training a single model on multiple related tasks simultaneously. In the context of RE, this could involve training a model to perform multiple RE tasks within a single domain or across multiple domains. Novel multi-task learning methods are desirable to improve the performance of RE models and reduce the need for domain-specific models.

Domain adaptation and transfer learning aim to improve the generalization ability of RE models across different domains. This involves adapting or transferring the knowledge learned from one domain to another, which could help improve the performance of domain RE models with limited annotated data.

(3) **Domain-extended knowledge** includes domain-specific ontologies, taxonomies, and knowledge graphs, by which additional context and information could be provided to improve the accuracy and effectiveness of RE tasks. To this end, the following aspects are worth exploring:

Integration with knowledge graphs. Knowledge graph could be used to represent heterogenous structured knowledge with entities and their relations. Integrating RE models with knowledge graphs can help improve the accuracy and interpretability of RE models, as well as enable more advanced knowledge inferences.

Explainable RE aims to make RE models more transparent and interpretable by providing explanations for the model's predictions, by incorporating the logic rules and inferences. This facilitates to increase the trust and reliability of RE models for the applications like decision support, intelligent question and answer, and so on.

To sum up, future research of domain RE will focus on improving the quality of or reducing the dependence on the domain-specific data, enhancing the generalization ability and scalability of domain-oriented model, and extending the external knowledge and interpretability.

Funding Support

This work was supported by the National Natural Science Foundation of China (62002311), Program of Key Laboratory of Yunnan Province (202205AG070003), and Major Science and Technology Special Foundation of Yunnan Province (202202AD080001).

Acknowledgements

The authors would like to appreciate Chengjie Xue and Shiyao Li for helping collect materials, and we would like to appreciate the anonymous reviewers for the constructive comments, which helps improve this paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

Abad, A., Nabi, M., & Moschitti, A. (2017). Self-crowdsourcing training for relation extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2*, 518–523.

Akkasi, A., & Moens, M. F. (2021). Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey. *Journal of Biomedical Informatics*, *119*, 103820.

Alexander, P. A. (1992). Domain knowledge: Evolving themes and emerging concerns. *Educational Psychologist*, *27*(1), 33–51.

Alicante, A., Corazza, A., Isgro, F., & Silvestri, S. (2016). Unsupervised entity and relation extraction from clinical records in Italian. *Computers in Biology and Medicine*, *72*, 263–275.

Bai, T., Guan, H., Wang, S., Wang, Y., & Huang, L. (2022). Traditional Chinese medicine entity relation extraction based on CNN with segment attention. *Neural Computing and Applications*, 1–10.

Cabot, P. L. H., & Navigli, R. (2021). REBEL: Relation extraction by end-to-end language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2370–2381.

Cai, R., Zhang, X. D., & Wang, H. F. (2016). Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1*, 756–765.

Chen, W., Shao, Y. F., Qian, L. H. & Zhou, G. D. (2021). Chemical protein relation extraction based on shortest dependency path and ensemble learning. *Journal of Chinese Information Processing*, *35*(4), 58–65.

Chen, Y. G., Sun, Y. Y., Yang, Z. H., & Lin, H. F. (2020). Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1561–1571.

Cheng, Q., Liu, J. T., Qu, X. Y., Zhao, J., Liang, J. Q., Wang, Z. F., . . . , & Xiao, Y. H. (2021). HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2819–2831.

Christopoulou, F., Miwa, M., & Ananiadou, S. (2019). Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 4925–4936.

Di, S. M., Shen, Y. Y., & Chen, L. (2019). Relation extraction via domain-aware transfer learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1348–1357.

Ding, Z. Y., Yang, Z. H., Luo, L., Wang, L., Zhang, Y., Lin, H. F. & Wang, J. (2021). Chinese biomedical entity relation extraction system based on deep learning. *Journal of Chinese Information Processing*, *35*(5), 70–76.

E, H. H., Zhang, W. J., Xiao, S. Q., Cheng, Q., Hu, Y. X., Zhou, X. S., & Niu, P. Q. (2019). Survey of entity relationship extraction based on deep learning. *Journal of Software*, *30*(6), 1793–1818.

Eichler, K., Hemsen, H., & Neumann, G. (2008). Unsupervised relation extraction from web documents. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 1674–1679.

Fabregat, H., Duque, A., Martinez-Romo, J., & Araujo, L. (2023). Negation-based transfer learning for improving biomedical named entity recognition and relation extraction. *Journal of Biomedical Informatics*, 104279.

Fang, Y., & Chang, K. C. C. (2011). Searching patterns for relation extraction over the web: Rediscovering the pattern-relation duality. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, 825–834.

Fu, R., Li, J. Y., Wang, J. H., Yue, K., & Hu, K. (2021). Joint extraction of entities and relations for domain knowledge graph. *Journal of East China Normal University (Natural Science)*, *2021*(5), 24.

Gao, D., Peng, D. L., & Liu, C. (2018). Entity relation extraction based on CNN in large-scale text data. *Journal of Chinese Computer Systems*, *39*(5), 1021–1026.

Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. (2017). Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 1*, 179–188.

Geng, Z. Q., Zhang, Y. H., & Han, Y. M. (2021). Joint entity and relation extraction model based on rich semantics. *Neurocomputing*, *429*, 132–140.

Gonzalez, E., & Turmo, J. (2009). Unsupervised relation extraction by massive clustering. In *Proceedings of the 9th IEEE International Conference on Data Mining*, 782–787.

Guan, T., Zan, H., Zhou, X., Xu, H., & Zhang, K. (2020). CMeIE: Construction and evaluation of Chinese medical information extraction dataset. In *Proceedings of the 9th Natural Language Processing and Chinese Computing*, 270–282.

Gutiérrez, B. J., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., & Su, Y. (2022). Thinking about gpt-3 in-context learning for biomedical IE? Think again. In *Findings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4497–4512.

Hachey, B., Grover, C., & Tobin, R. (2012). Datasets for generic relation extraction. *Natural Language Engineering*, *18*(1), 21–59.

Han, X., Gao, T. Y., Lin, Y. K., Peng, H., Yang, Y. L., Xiao, C. J., . . . , & Zhou, J. (2020). More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 745–758.

He, B., Dong, B., Guan, Y., Yang, J. F., Jiang, Z. P., Yu, Q. B., . . . , & Qu, C. Y. (2017). Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. *Journal of Biomedical Informatics*, *69*, 203–217.

He, Y., Li, Z. X., Yang, Q., Chen, Z. G., Liu, A., Zhao, L., & Zhou, X. F. (2020). End-to-end relation extraction based on bootstrapped multi-level distant supervision. *World Wide Web*, *23*, 2933–2956.

Hong, W. X., Hu, Z. Q., Weng, Y., Zhang, H., Wang, Z., & Guo, Z. X. (2020). Automated knowledge graph construction for judicial case facts. *Journal of Chinese Information Processing*, *34*(1), 34–44.

Hou, W. J., Hong, L., Xu, H. S., & Yin, W. (2023). RoRED: Bootstrapping labeling rule discovery for robust relation extraction. *Information Sciences*, *629*, 62–76.

Huang, Q. Z., Zhu, S. Q., Feng, Y. S., Ye, Y., Lai, Y. X. & Zhao, D. Y. (2021). Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2*, 998–1004.

Huang, Y., Jia, Y., Gan, L., Xu, J., & He, Z. (2018). Multi-factor person entity relation extraction model based on distant supervision. *Journal on Communications*, *39*(7), 103–112.

Jiang, J. (2009). Multi-task transfer learning for weakly-supervised relation extraction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*, 1012–1020.

Kaushik, N., & Chatterjee, N. (2016). A practical approach for term and relationship extraction for automatic ontology creation from agricultural text. In *Proceedings of the 2016 International Conference on Information Technology*, 241–247.

Kawashima, K., Bai, W., & Quan, C. (2017). Text mining and pattern clustering for relation extraction of breast cancer and related genes. In *Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, 59–63.

Krallinger, M., Rabal, O., Lourenco, A., Oyarzabal, J., & Valencia, A. (2017). Information retrieval and text mining technologies for chemistry. *Chemical Reviews*, *117*(12), 7673–7761.

Lai, Q. H., Ding, S., Gong, J. H., Cui, J. A. & Liu, S. (2022). A Chinese Multi-modal Relation Extraction Model for Internet Security of Finance. In *Proceedings of the 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops*, 123–128.

Li, A., Wang, X. M., Wang, W. H., Zhang, A. M., & Li, B. H. (2019c). A survey of relation extraction of knowledge graphs. In *Proceedings of the Asia-Pacific Web and Web-Age Information Management Joint International Conference on Web and Big Data Workshops*, 52–66.

Li, F., Zhang, M. S., Fu, G. H., & Ji, D. H. (2017). A neural joint model for extracting bacteria and their locations. In *Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 15–26.

Li, P. F., & Mao, K. Z. (2019). Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, *115*, 512–523.

Li, S. J., He, W., Shi, Y. B., Jiang, W. B., Liang, H. J., Jiang, Y., . . . , & Zhu, Y. (2019a). DuIE: A large-scale Chinese dataset for information extraction. In *Proceedings of the 8th Natural Language Processing and Chinese Computing*, 791–800.

Li, Y. F., Jin, R., & Luo, Y. (2019b). Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks. *Journal of the American Medical Informatics Association*, *26*(3), 262–268.

Li, Z. G., Lin, H. F., Shen, C., Xu, B. & Zheng W. (2022). Document-level chemical-induced disease relation extraction via cross self-attention. *Journal of Chinese Information Processing*, *36*(7), 98–105.

Li, Z. H., Gui, Y. Y., Yang, Z. H., Lin, H. F. & Wang, J. (2018). Chemical-induced disease relation extraction based on biomedical literature. *Journal of Computer Research and Development*, *55*(1), 198–206.

Liang, C., Zan, H. Y., Liu, Y. J., & Wu, Y. F. (2018). Research on entity relation extraction for military field. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, 81–88.

Lin, C., Miller, T., Dligach, D., Bethard, S., & Savova, G. (2017). Representations of time expressions for temporal relation extraction with convolutional neural networks. In *Proceedings of the BioNLP 2017*, 322–327.

Liu, Z. H., Chen, Y. Y., Dai, Y. F., Guo, C. H., Zhang, Z. W., & Chen, X. (2018). Syntactic and semantic features based relation extraction in agriculture domain. In *Proceedings of the 15th International Conference Web Information Systems and Applications*, 252–258.

Lu, Y. W., Yang, R. P., Jiang, X. P., Yin, C. S., Song, X. Y., & Liu, B. (2021a). Research on Military Event Detection Method Based on BERT-BiGRU-Attention. In *Proceedings of the 2021 IEEE International Conference on Consumer Electronics and Computer Engineering*, 1–5.

Lu, Y. W., Yang, R. P., Jiang, X. P., Zhou, D., Yin, C. S., & Li, Z. Z. (2021b). MRE: A military relation extraction model based on BiGRU and multi-head attention. *Symmetry*, *13*(9), 1742.

Luo, L., Lai, P. T., Wei, C. H., Arighi, C. N., & Lu, Z. Y. (2022). BioRED: A rich biomedical relation extraction dataset. *Briefings in Bioinformatics, 23*(5).

Marchesin, S., & Silvello, G. (2022). TBGA: A large-scale gene-disease association dataset for biomedical relation extraction. *BMC Bioinformatics*, *23*(1), 1–16.

Mesquita, F. (2012). Clustering techniques for open relation extraction. In *Proceedings of the 2012 International Conference on Management of Data*, 27–32.

Minard, A. L., Ligozat, A. L., & Grau, B. (2011). Multi-class SVM for relation extraction from clinical reports. In *Proceedings of the 2011 International Conference Recent Advances in Natural Language Processing*, 604–609.

Miwa, M., Pyysalo, S., Hara, T., & Tsujii, J. I. (2010). Evaluating dependency representations for event extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 779–787.

Nan, G. S., Guo, Z. J., Sekulić, I., & Lu, W. (2020). Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1546–1557.

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys*, *54*(1), 1–39.

Nguyen, T. H., Plank, B., & Grishman, R. (2015). Semantic representations for domain adaptation: A case study on the tree kernel-based method for relation extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 1*, 635–644.

Okamoto, M., Shan, Z., & Orihara, R. (2017). Applying information extraction for patent structure analysis. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 989–992.

Panyam, N. C., Verspoor, K., Cohn, T., & Ramamohanarao, K. (2018). Exploiting graph kernels for high performance biomedical relation extraction. *Journal of Biomedical Semantics*, *9*(1), 1–11.

Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). Relation extraction: A survey. *arXiv Preprint:* 1712.05191. https://doi.org/10.48550/arXiv.1712.05191

Peng, N. Y., Poon, H. F., Quirk, C., Toutanova, K., & Yih, W. T. (2017). Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, *5*, 101–115.

Plank, B., & Moschitti, A. (2013). Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 1*, 1498–1507.

Plum, A., Ranasinghe, T., Jones, S., Orasan, C., & Mitkov, R. (2022). Biographical semi-supervised relation extraction dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3121–3130.

Qian, L. H., Zhou, G. D., Kong, F., Zhu, Q. M., & Qian, P. D. (2008). Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics*, 697–704.

Qin, B., Liu, A. & Liu, T. (2015). Unsupervised Chinese open entity relation extraction. *Journal of Computer Research and Development*, *52*(5), 1029–1035.

Ravikumar, K. E., Rastegar-Mojarad, M., & Liu, H. (2017). BELMiner: Adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. *Database*, *2017*.

Sahu, S. K., Christopoulou, F., Miwa, M., & Ananiadou, S. (2019). Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4309–4316.

Sharma, S., Nayak, T., Bose, A., Meena, A. K., Dasgupta, K., Ganguly, N., & Goyal, P. (2022). FinRED: A dataset for relation extraction in financial domain. In *Proceedings of the Web Conference 2022*, 595–597.

Smirnova, A., & Cudré-Mauroux, P. (2018). Relation extraction using distant supervision: A survey. *ACM Computing Surveys*, *51*(5), 1–35.

Song, K. Y., Yue, K., Duan, L., Yang, M. Z., & Li, A. S. (2022). Mutual information based Bayesian graph neural network for few-shot learning. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, 1866–1875.

Sun, C., Fu, Y. N., Cheng, W. L. & Qian, W. N. (2018). Chinese named entity relation extraction for enterprise knowledge graph construction. *Journal of East China Normal University (Natural Science)*, *3*, 55–66.

Sun, X., Fu, C. C., Liu, S. Q., Chen, W. J., Zhong, R., He, T. T., & Jiang, X. P. (2021). Multi-type microbial relation extraction by transfer learning. In *Proceedings of the International Conference on Bioinformatics and Biomedicine 2021*, 266–269.

Sureshkumar, G., & Zayaraz, G. (2015). Automatic relation extraction using naïve Bayes classifier for concept relational ontology development. *International Journal of Computer Aided Engineering and Technology*, *7*(4), 421–435.

Tang, H. Z., Cao, Y. N., Zhang, Z. Y., Cao, J. X., Fang, F., Wang, S., & Yin, P. F. (2020). Hin: Hierarchical inference network for document-level relation extraction. In *Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 197–209.

Thomas, A., & Sivanesan, S. (2022). An adaptable, high-performance relation extraction system for complex sentences. *Knowledge-Based Systems*, *251*, 108956.

Tiktinsky, A., Viswanathan, V., Niezni, D., Azagury, D. M., Shamay, Y., Taub-Tabib, H. & Goldberg, Y. (2022). A Dataset for N-ary Relation Extraction of Drug Combinations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3190–3203.

Tran, T. T., Le, P., & Ananiadou, S. (2020). Revisiting unsupervised relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7498–7505.

Veena, G., Gupta, D., & Kanjirangat, V. (2022). Semi supervised approach for relation extraction in agriculture documents. In *Proceedings of the International Conference on Information Technology 2022*, 199–204.

Vela, M., & Declerck, T. (2009). Concept and relation extraction in the finance domain. In *Proceedings of the 8th International Conference on Computational Semantics*, 346–350.

Verga, P., Strubell, E., & McCallum, A. (2018). Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 872–884.

Wan, H. Y., Moens, M. F., Luyten, W., Zhou, X. Z., Mei, Q. Z., Liu, L., & Tang, J. (2016). Extracting relations from traditional Chinese medicine literature via heterogeneous entity networks. *Journal of the American Medical Informatics Association*, *23*(2), 356–365.

Wan, Q. Z., Wan, C. X., Xiao, K., Hu, R., Liu, D., & Liu, X. P. (2023). CFERE: Multi-type Chinese financial event relation extraction. *Information Sciences*, *630*, 119–134.

Wang, H. L., Qin, K., Zakari, R. Y., Lu, G. M., & Yin, J. (2022a). Deep neural network-based relation extraction: An overview. *Neural Computing and Applications*, 1–21.

Wang, J. H., Yue, K., Duan, L., Qi, Z. W., & Qiao, S. J. (2022b). An efficient approach for multiple probabilistic inferences with Deepwalk based Bayesian network embedding. *Knowledge-Based Systems*, *239*, 107996.

Wang, Q., Mao, Z. D., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, *29*(12), 2724–2743.

Wang, X. F., Yang, R. P., Feng, Y. L., Li, D. S., & Hou, J. F. (2018a). A military named entity relation extraction approach based on deep learning. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, 1–6.

Wang, Y. S., Wang, L. W., Rastegar, M., Moon, S., Shen, F. C., Afzal, N., . . . , & Liu, H. F. (2018b). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, *77*, 34–49.

Xiao, Y., Jin, Y. C., Cheng, R., & Hao, K. R. (2022). Hybrid attention-based transformer block model for distant supervision relation extraction. *Neurocomputing*, *470*, 29–39.

Xing, R., Luo, J., & Song, T. W. (2020). BioRel: Towards large-scale biomedical relation extraction. *BMC Bioinformatics*, *21*, 1–13.

Xu, J. J., Wen, J., Sun, X., & Su, Q. (2017). A discourse-level named entity recognition and relation extraction dataset for Chinese literature text. *arXiv preprint:1711.07010*. https://doi.org/10.48550/arXiv.1711.07010

Xu, K., Reddy, S., Feng, Y. S., Huang, S. F., & Zhao, D. Y. (2016). Question answering on freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1*, 2326–2336.

Xu, L., Chia, Y. K., & Bing, L. D. (2021). Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 1*, 4755–4766.

Xu, T. Y., Hua, W., Qu, J. F., Li, Z. X., Xu, J. J., Liu, A., & Zhao, L. (2022). Evidence-aware Document-level Relation Extraction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2311–2320.

Xue, F. Z., Sun, A. X., Zhang, H., & Chng, E. S. (2021). Gdpnet: Refining latent multi-view graph for relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14194–14202.

Yan, C. W., Fu, X. L., Wu, W. Q., Lu, S. L., & Wu, J. (2019). Neural network based relation extraction of enterprises in credit risk management. In *Proceedings of the 2019 IEEE International Conference on Big Data and Smart Computing*, 1–6.

Yang, H., Yu, H., & Sun, Z. (2021). Fishery standard entity relation extraction using dual attention mechanism. *Transactions of the Chinese Society of Agricultural Engineering*, *37*(14), 204–212.

Yao, Y., Du, J. J., Lin, Y. K., Li, P., Liu, Z. Y., Zhou, J., & Sun, M. S. (2021). CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild. In *Proceedings of the 2021 International Conference on Empirical Methods in Natural Language Processing*, 4452–4472.

Yao, Y., Ye, D. M., Li, P., Han, X., Lin, Y. K., Liu, Z. H., . . . , Sun, M. S. (2019). DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 764–777.

Yuan, P. S., Li, R. L., Wang, C., & Xu, H. L. (2021). Entity relationship extraction from rice phenotype knowledge graph based on BERT. *Transactions of the Chinese Society for Agricultural Machinery*, *52*, 151–158.

Yuan, Y., Zhou, X. F., Pan, S. R., Zhu, Q. N., Song, Z. L., & Guo, L. (2020). A relation-specific attention network for joint entity and relation extraction. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 4054–4060.

Zhang, S., Lu, X. K., & Wu, J. (2022). Co-Training with validation: A generic framework for semi-supervised relation extraction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4697–4701.

Zhang, X., Liang, X., Li, Z. Y., Zhang, S. S., & Zhao, X. L. (2019). Identification and analysis of love relationships of protagonists in Jin Yong's fictions. *Journal of Chinese Information Processing*, *33*(4), 109–119.

Zhang, Y. J., Lin, H. F., Yang, Z. H., Wang, J., Zhang, S. W., Sun, Y. Y., & Yang, L. (2018). A hybrid model based on neural networks for biomedical relation extraction. *Journal of Biomedical Informatics*, *81*, 83–92.

Zheng, C. M., Wu, Z. W., Feng, J. H., Fu, Z., & Cai, Y. (2021). MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo*, 1–6.

Zheng, S. C., Wang, F., Bao, H. Y., Hao, Y. X., Zhou, P., & Xu, B. (2017). Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 1*, 1227–1236.

Zhou, D. Y., Zhong, D. Y., & He, Y. L. (2014). Biomedical relation extraction: From binary to complex. *Computational and Mathematical Methods in Medicine*, *2014*, 1–19.

Zhou, G. D., Qian, L. H., & Fan, J. X. (2010). Tree Kernel-based semantic relation extraction with rich syntactic and semantic information. *Information Sciences*, *180*(8), 1313–1325.

Zhou, Y. T., Meng, J., Guo, Y., Liu, Y., He, G. F., Dong, L. & Cheng, X. Q. (2022). Multiple relationship extraction from unstructured financial announcements. *Journal of Chinese Information Processing*, *36*(2), 76–84.

Zuo, M., & Zhang, Y. (2022). A span-based joint model for extracting entities and relations of bacteria biotopes. *Bioinformatics*, *38*(1), 220–227.

How to Cite: Wang, J., Yue, K., & Duan, L. (2023). Models and Techniques for Domain Relation Extraction: A Survey. *Journal of Data Science and Intelligent Systems* 1(2), 65–82, https://doi.org/10.47852/bonviewJDSIS3202973