A Survey on Document-Level Relation Extraction: Methods and Applications

Yifan Zheng(✉), Yikai Guo, Zhizhao Luo, Zengwen Yu, Kunlong Wang, Hong Zhang, and Hua Zhao

Beijing Institute of Computer Technology and Application, Beijing, China
zhengyifan_ht@163.com

Abstract. Relation extraction is a significant area of research in the field of information extraction, to extract target information accurately and efficiently from vast amounts of data to improve the utilization of information. Relation extraction is widely used in various downstream tasks such as text mining, information retrieval, and question-answering systems. Compared to sentence-level relation extraction, document-level relation extraction is more complex and challenging, yet there is a lack of a comprehensive overview of document-level relation extraction. This paper presents a survey on document-level relation extraction, first categorizing existing techniques into three categories and introducing the most representative models. Then, we describe the primary application domains and commonly used datasets for relation extraction. Finally, we analyse the research challenges and future trends in document-level relation extraction.

Keywords: information extraction · document-level relation extraction · application

1 Introduction

In the current era of knowledge explosion, various types of information are being constantly created and accumulated on the internet, raising the need for extracting valuable relationships from such information. This has become a major issue in the field of natural language processing (NLP). Relation extraction (RE), as a fundamental task in information extraction (IE), endeavors to automatically identify and extract relationships between entities from text, images, video, and other media. RE has been widely employed in various fields, such as text mining, information retrieval, and question-answering systems.

Document-level relation extraction (DocRE) is a subfield of relation extraction that endeavors to identify and extract relationships between various entities in a document. In contrast to sentence-level relation extraction, DocRE involves analyzing longer and more complex texts, and conducting a more comprehensive analysis of contextual information. Therefore, DocRE is a highly challenging task. In recent years, research on DocRE has gained increasing attention due to the advancements in deep learning and big data technology. Various methods have been proposed, that employ these techniques and

D. Kumar et al. (Eds.): IEIT 2023, AHSSEH 10, pp. 1061–1071, 2023.
https://doi.org/10.2991/978-94-6463-230-9_128

show promising results in experimental settings. However, the research direction in this field remains undefined, and further exploration is necessary.

This paper offers an extensive survey of the current state of DocRE research. Section 2 outlines relation extraction as a main task, along with the pertinent techniques employed in DocRE. In Sect. 3, we present a description of the applications of relation extraction in diverse domains, including political and news, law, business finance, and medical fields. Following this, in Sect. 4, we provide an overview of relevant datasets. Finally, Sects. 5 and 6 expound on the primary challenges associated with this task and present prognostications for future research directions.

2 Relation Extraction Techniques

The fundamental objective of RE is to automatically identify and extract factual relationships between pairs of entities from unstructured input data, which typically consists of mentions of these entities and the relationships among them. The mentions of entities often include their names, aliases, or other non-standard descriptions. Therefore, relation extraction is typically divided into two subtasks: named entity recognition (NER) and RE.

Given a document consisting of N sentences $D = \{s_1, s_2, \cdots, s_N\}$, the NER method is first employed to identify the existing entities $E = \{e_1, e_2, \cdots, e_i\}$ and their mentions $M_i = \{m_{i,1}, m_{i,2}, \cdots, m_{i,k_i}\}$ in the document. Then, by combining the entity pairs (e_i, e_j) with their corresponding mention representations and document-level semantic information, relation triplets $< e_i, r, e_j >$ are obtained, where r belongs to the target relation set $R = \{r_1, r_2, \cdots, r_i\}$.

Based on the approach used to perform DocRE, current research can be roughly classified into three classes: neural network-based relation extraction, graph-based relation extraction, and pre-trained language model-based relation extraction. The problem definition and technical classification of DocRE are shown in Fig. 1.

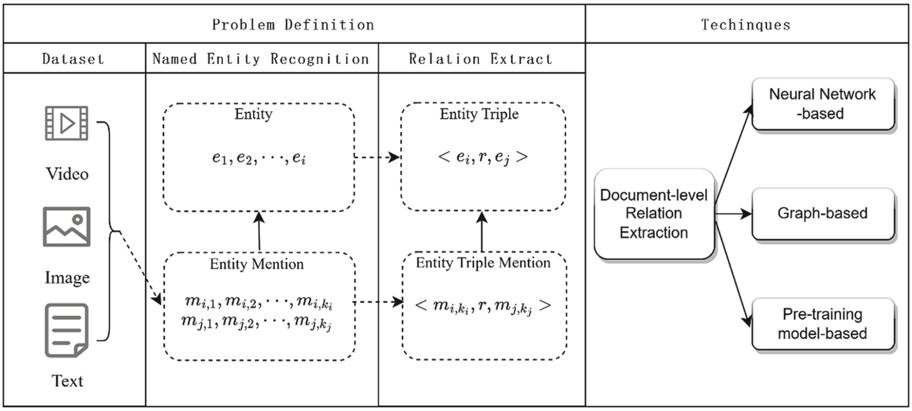

Fig. 1. Taxonomy of document-level relation extraction problems and techniques

2.1 Neural Network-Based Models

The neural network-based approach for DocRE involves the acquisition of entity representations in documents through the application of neural networks. This approach is then followed by the classification of all candidate entity pairs. Early endeavors in information extraction emphasized the utilization of shallow information and pertinent statistical features for information extraction. However, these pre-deep learning techniques were primarily based on statistics or rules, thereby limiting their capability to effectively handle the burgeoning number of data samples on the internet.

Giorgi et al. [1] present a novel sequence-to-sequence approach, named seq2rel, which facilitates end-to-end learning of document-level relation extraction sub-tasks, such as entity extraction, co-reference resolution, and relation extraction. The proposed model uses a simple yet effective technique called entity implication and reports encouraging results on various widely used biomedical datasets. Ru et al. [2] present LogiRE, a probabilistic model for DocRE that leverages logical rules. It treats logical rule as a latent variable and comprises two modules. The first module produces logical rules that may impact the final prediction, and the second module provides the final prediction based on the generated logical rules. The incorporation of logical rules into the neural network enables LogiRE to capture long-range dependency relationships explicitly, leading to improved performance. The work described in [3] presents the Hierarchical Inference Network (HIN) that aims to leverage multi-granularity information at a different level. It treats entity, sentence, and document as three levels. It employs translational constraints and bilinear transformation to extract first-level inference information for entity pairs across multiple subspaces. It then models the relationship between first-level information and sentence representations to obtain second-level inference information. Finally, HIN adopts a hierarchical aggregation technique to acquire third-level inference information. By integrating these three distinct levels of granularity, the HIN model can effectively aggregate reasoning information.

2.2 Graph-Based Models

The graph-based method to RE involves the construction of a document graph, where nodes represent words, mentions, entities, or sentences, and edges are formed using heuristic rules, such as syntactic knowledge, coreference, adjacency, and co-occurrence. This approach enables the explicit learning of associations between entities through graph propagation.

Makino et al. [4] present an edge-editing method for RE, which aims to take advantage of the interdependence of relations within a document. The method treats relations as a graph, with entities serving as nodes and relationships as edges. The method incrementally constructs it through an iterative process of editing the edges, which involves classifying the edges using document and graph information. To address the issue of the encoder-classifier model for DocRE potentially dispersing attention on entity pairs that lack relations, a reconstruction model [5] was proposed. The reconstruction model endeavors to restore the ground-truth path from the graph by generating a sequence of node representation on the path that connects two entity nodes. The aim is to maximize the probability of the path in case of the existence of ground-truth relationships between the entity pair. This emphasis during training enables the proposed DocRE model to concentrate on the learning of entity pairs that exhibit relationships, thus facilitating effective learning of graph representations for subsequent relationship classification. To exploit the anaphoric information of pronouns and the lexical information of entity relations that are directly expressed in a sentence, Park et al. [6] propose two graph structures. The first structure is the anaphoric graph, which is a heterogeneous graph that leverages the anaphoric information by using three edge types: from entity to pronoun, from pronoun to entity, and between pronouns. The second structure is the local context graph, which captures the salient information around sentence entities by connecting all the words between the current mention and the nearest mention.

2.3 Pre-training Model-Based Models

Methods based on PLMs primarily encode documents using pre-trained language models (Transformer and BERT).

Wang et al. [7] employ BERT to encode documents and fine-tune the entire model using annotated data from the DocRED dataset [8]. The proposed model further decomposes DocRE into a two-step process. The first entails predicting the existence of relationships between entity pairs, while the second involves predicting the specific type of relations present between the given entity pair. Xie et al. [9] present an innovative evidence-enhanced framework for DocRE, named EIDER. The objective of this framework is to enhance the performance of DocRE by extracting evidence efficiently and fusing it during the inference stage. The EIDER model involves joint training of a RE model and a lightweight evidence extraction model. The RE model utilizes the extracted evidence and the complete document to predict RE and subsequently merges the predictions through a mixing layer. This approach enables the model to prioritize crucial sentences accessing complete information in the document. Qin et al. [10] propose a framework called ERICA to enhance pre-trained language models' understanding of document semantics using contrastive learning. ERICA utilizes a PLM to learn entity vector representations and their contextual information. The framework then introduces two customized tasks to refine the understanding of entities and relations. The first task involves inferring tail entities given head entities and relations, while the second task consists in constructing entity pairs through distant supervision and using contrastive learning to identify positive and negative relationship instances, recognizing the similarity between entity relationships.

3 Application

In this section, we introduce the application of relation extraction in four areas: political news, law, business finance, and medicine.

With the development of the internet, traditional media and internet platforms have become significant sources of news content. Relation extraction of these events can reduce reading time and improve reading efficiency, which can help decision-making. Managers can obtain core information about hot events through relation extraction and keep abreast of social dynamics and political situations at home and abroad. For example, in the case of the Russia-Ukraine conflict, relation extraction can help us better understand the relationship between various entities such as countries like Russia, Ukraine, and the United States, as well as the relationship between events such as war and ceasefire agreements. By extracting and analysing these relations, more comprehensive and accurate information can be provided, and we can better understand the background and the conflict process of the Russian-Ukrainian conflict. For example, Sheng et al. [11] design a multi-document semantic relation extraction system called MuReX for multisource data in news. The system aims to extract important relations from news articles, analyse the correlations between facts, and eventually visualize the results.

In the legal field, relation extraction can be used to extract relations between entities in legal texts. For instance, it can be used to identify the relation between the plaintiff and the defendant in a case or various relations in legal provisions, helping lawyers find relevant information faster. Relation extraction can also be used in legal risk control and intelligent contracts. For example, various clauses in a contract can be automatically identified, classified, and analysed by relation extraction, thus improving the accuracy of contract management. For instance, Song et al. [12] use LSTM to inject syntactic information for the cascading relations in legal texts and introduce a multi-headed attention mechanism to decompose the overlapping relations, which improves the accuracy of nested relation extraction in legal texts.

Business finance refers to the intersection of business and finance fields and is a complex system. Through relation extraction, investment relations and equity relations between companies can be extracted from news reports. They can also be applied to risk control and credit evaluation to help companies better understand market dynamics and risk situations. For example, Reyes et al. [13] describe a method for obtaining competitive business intelligence from articles by using a bidirectional Transformer Encoding Representation.

The application of relation extraction in the medical field has great prospects. With the development of medical informatization, more and more medical text data are digitized, and how to extract useful information from these data has become an important issue. Relation extraction technology can automatically extract the semantic relations between entities from a substantial quantity of medical resources and build medical knowledge maps. These knowledge graphs can help doctors better understand patients' conditions and improve diagnosis and treatment outcomes. In addition, relation extraction techniques can be applied to drug side effect prediction, drug interaction prediction, disease risk prediction, etc. For instance, Zhang et al. [14] propose a multi-hop approach to get relations from medical resources and retain their multifaceted semantic information and obtain multiple weight vectors, further improving the medical relation extraction accuracy.

4 Corpus

In this section, we present the mainstream evaluation metrics for DocRE, as well as the widely used datasets, which are summarized in the Table 1.

Table 1. Statistics of Document-level Relation Extraction Datasets.

S.NO.	Dataset	Year	Domain	Modality	Construction Method	#Relation Type	Documents	Language
1	DocRED	2019	General	Text	Manually Annotated	96	5053	English
2	SCIREX	2020	Science	Text	Manually Annotated & Distant Supervision	4	438	English
3	CDR	2016	Medicine	Text	Manually Annotated	3116	1500	English
4	GDA	2019	Medicine	Text	Manually Annotated	1	30192	English
5	DWIE	2019	News	Text	Manually Annotated	65	802	English
6	HacRED	2021	General	Text	Manually Annotated & Distant Supervision	26	9231	Chinese

4.1 Evaluation Metrics

The task of DocRE is often framed as multi-label classification where the input consists of a document and the entity pairs it contains, and the output is the predicted relationship between each entity pair. The performance of a relation extraction model is typically evaluated using precision (P_{Macro}), recall (R_{Macro}), and F1 score ($F1_{Macro}$). Moreover, to mitigate the impact of identical entity pairs in the validation/test sets and training set on the final evaluation metrics, researchers have proposed $IgnF1$. The definitions of these metrics are as follows:

$$P_{Macro} = \frac{1}{N}\sum\nolimits_{i=1}^{N} P_i = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{TP_i}{TP_i + FP_i} \tag{1}$$

$$R_{Macro} = \frac{1}{N}\sum\nolimits_{i=1}^{N} R_i = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \tag{2}$$

$$F1_{Macro} = \frac{1}{N}\sum\nolimits_{i=1}^{N} F1_i = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{2P_iR_i}{P_i + R_i} \tag{3}$$

where the true positives (TP) are the number of samples correctly predicted as positive by the model. True negatives (TN) refer to the number of samples correctly predicted as negative by the model. False positives (FP) denote the number of samples incorrectly predicted as positive by the model, while false negatives (FN) represent the number of samples incorrectly predicted as negative. In addition, N denotes the total number of samples with a relation label.

$$IgnF1 = \frac{2 \cdot P_{Ign} \cdot R_{Ign}}{P_{Ign} + R_{Ign}} \tag{4}$$

where, P_{Ign} and R_{Ign} represent the precision and recall of removing overlapping relationship facts in training and verification/test, respectively.

4.2 DocRED Corpus

The DocRED Corpus [8] is a large-scale and manually annotated dataset. The corpus was made from Wikipedia and Wikidata and includes a broad range of entity types, such as people, locations, organizations, time, quantity, and others. It contains a total of 5,053 human-annotated documents and 101,873 distantly supervised documents and covers 96 relation types that are relevant to domains such as science, art, and personal life. The majority of the relations in this corpus require complex reasoning, with 46.4% of them being cross-sentence relations, and 40.7% requiring multiple sentences to be inferred. Furthermore, 61.1% of the relations necessitate different reasoning mechanisms such as logical inference to be extracted.

4.3 SCIREX Corpus

The SCIREX Corpus [14] is a corpus designed for comprehensive information extraction from scientific documents. The dataset was created through a combination of distant supervision and manual annotation and consists of a total of 438 documents, with an average length of 5,735 words per document. Notably, the document length in SCIREX is longer than other corpora, such as DocRED.

4.4 CDR Corpus

The CDR Corpus [15] is a manually annotated dataset specifically designed for predicting chemical-disease relations in CoeRE within the biomedical domain. The corpus comprises over 1500 documents retrieved from the biomedical literature database PubMed, in which over 90,000 drug and chemical entities are annotated, and various types of relationships between them are identified. Due to its specialized domain and comprehensive annotations, the CDR Corpus has significant implications for advancing biomedical research.

4.5 GDA Corpus

The GDA Corpus [16] is used for binary relation recognition between genes and diseases. The corpus was created using the distant supervision method, where 30,192 MEDLINE abstracts were annotated based on the knowledge base generated from the PubMed abstract dataset via the DisGeNET platform. As with the CDR Corpus, the GDA Corpus is a crucial benchmark dataset for research in the biomedical domain.

4.6 DWIE Corpus

The DWIE Corpus [17] is a multitask dataset that primarily emphasizes entity-centric and document-level annotations, incorporating implicit interactions between entities within a document. The corpus consists of 802 news articles in the English language, comprising 23,130 entities of 311 multi-labelled entity types, 21,749 relationships, and 65 multi-labeled relationship types.

4.7 HacRED Corpus

The HacRED Corpus [18] is a Chinese DocRE dataset that has been constructed using a combination of manual annotation and distant supervision. It consists of 9,231 documents with 9 types of entities and 26 predefined relations, resulting in 65,225 relational facts. The dataset is notable for its low fact redundancy, even distribution of entity relations, and diverse cases. As a result, it is a considerable resource for researchers investigating DocRE.

5 Challenges

In practical settings, numerous relational facts are implicitly expressed among entity pairs, which are distributed across different sentences in a document. Additionally, multiple entities within the document often have intricate relationships with each other. Consequently, DocRE poses a demanding task because it involves processing much longer texts with a larger number of entities.

5.1 Effectively Modelling Useful Information Within Documents

In DocRE, entities can appear in multiple sentences, and relationships can span across multiple sentences. Thus, models require the use of multiple sentences to infer relationships. Additionally, in natural language text, entity coreference can pose a problem where the same entity can have multiple references. This problem becomes more complex in documents, where multiple references to the same entity can exist, and different references need to be aggregated to learn entity representations. The complex semantic information within a document involves reasoning across multiple aspects, such as logical reasoning, coreference resolution, and commonsense knowledge reasoning. Logical reasoning is crucial for relations between different entities, and as a document contains multiple entity-relation triples, logical connections between different entity-relation triples are inevitable. Hence, DocRE models must possess some degree of logical reasoning ability. Therefore, effectively modelling the complex interactions and semantic dependencies in a document to improve the performance of DocRE models remains a challenging research problem.

5.2 Construction of Annotated Corpus

At present, there exists a scarcity of high-quality benchmark datasets suitable for DcoRE in several domains, such as finance and security. Furthermore, developing relevant datasets requires significant domain knowledge and substantial resources, making it challenging to make full use of unsupervised data. Consequently, the challenge of promptly constructing large-scale, high-quality benchmark datasets in a cost-effective and less labor-intensive way remains a crucial issue in the field of DocRE.

6 Future Directions

6.1 Multi-document Relation Extraction

In real-world scenarios, a relation between two entities in an entity pair may span across multiple documents. Cross-document relation extraction is, therefore, crucial for modelling and identifying relationships that exist beyond a single document. Multi-document datasets are often more complex, with a larger number of entities and relationships, which can be challenging for graph model training. Furthermore, pre-training language models may struggle to effectively capture long-distance contextual semantic information between documents, particularly for lengthy documents. As a result, novel methods are required to address multi-document relation extraction, particularly in the context of large and complex information domains.

6.2 Multimodal Relation Extraction

In the future, relation extraction is poised to expand beyond textual information and extend to other modalities such as video and images, enabling the extraction of more effective information. However, compared to text, these modalities present distinct challenges related to methodology, terminology, task definition, and annotation data. Although some research has explored multimodal relation extraction, the field is still nascent, and multimodal relation extraction remains a promising research direction that warrants further exploration.

References

1. John Giorgi, Gary Bader, and Bo Wang. 2022. A sequence-to-sequence approach for document-level relation extraction. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 10–25.
2. Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. 2021. Learning Logic Rules for Document-Level Relation Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1239–1250.
3. Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. HIN: Hierarchical Inference Network for Document-Level Relation Extraction. In Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, 197–209.

4. Kohei Makino, Makoto Miwa, and Yutaka Sasaki. 2021. A Neural Edge-Editing Approach for Document-Level Relation Graph Extraction. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2653–2662.
5. Xu, W., Chen, K. and Zhao, T. 2021. Document-Level Relation Extraction with Reconstruction. Proceedings of the AAAI Conference on Artificial Intelligence, 14167–14175.
6. Seongsik Park, Dongkeun Yoon, and Harksoo Kim. 2022. Improving Graph-based Document-Level Relation Extraction Model with Novel Graph Structure. In Proceedings of the 31st ACM International Conference on Information &amp; Knowledge Management (CIKM '22). Association for Computing Machinery, New York, NY, USA, 4379–4383.
7. Wang, Hong, Focke Christfried, Sylvester Rob, Mishra Nilesh, Wang William. 2019. Fine-tune Bert for DocRED with Two-step Process.
8. Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 764–777.
9. Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering Document-level Relation Extraction with Efficient Evidence Extraction and Inference-stage Fusion. In Findings of the Association for Computational Linguistics: ACL 2022, pages 257–268.
10. Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3350–3363.
11. Yongpan Sheng, Zenglin Xu, Yafang Wang, and Gerard de Melo. 2020. Multi-document semantic relation extraction for news analytics. World Wide Web 23, 3 (May 2020), 2043–2077.
12. Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint Entity and Relation Extraction for Legal Documents with Legal Feature Enhancement. In Proceedings of the 28th International Conference on Computational Linguistics, pages 1561–1571.
13. Daniel De Los Reyes, Douglas Trajano, Isabel Harb Manssour, Renata Vieira, and Rafael H. Bordini. 2021. Entity Relation Extraction from News Articles in Portuguese for Competitive Intelligence Based on BERT. In Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29 – December 3, 2021, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 449–464.
14. Zhang, T., Lin, H., Tadesse, M.M. et al. Chinese medical relation extraction based on multi-hop self-attention mechanism. Int. J. Mach. Learn. & Cyber. 12, 355–363 (2021).
15. Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7506–7516.
16. Li J, Sun Y, Johnson RJ, Sciaky D, Wei CH, Leaman R, Davis AP, Mattingly CJ, Wiegers TC, Lu Z. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford).
17. Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: An entity-centric dataset for multi-task document-level information extraction. Inf. Process. Manage. 58, 4 (Jul 2021).
18. Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 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, pages 2819–2831.