Published in ACM Transactions on Information Systems (2023) 42 (1), 22.
DOI: 10.1145/3597610

Document-level Relation Extraction via Separate Relation Representation and Logical Reasoning

HEYAN HUANG and CHANGSEN YUAN, Beijing Institute of Technology, China
QIAN LIU, Nanyang Technological University, Singapore
YIXIN CAO, Singapore Management University, Singapore

Document-level relation extraction (RE) extends the identification of entity/mentions' relation from the single sentence to the long document. It is more realistic and poses new challenges to relation representation and reasoning skills. In this article, we propose a novel model, **SRLR**, using **S**eparate Relation **R**epresentation and **L**ogical **R**easoning considering the indirect relation representation and complex reasoning of evidence sentence problems. Specifically, we first expand the judgment of relational facts from the entity-level to the mention-level, highlighting fine-grained information to capture the relation representation for the entity pair. Second, we propose a logical reasoning module to identify evidence sentences and conduct relational reasoning. Extensive experiments on two publicly available benchmark datasets demonstrate the effectiveness of our proposed SRLR as compared to 19 baseline models. Further ablation study also verifies the effects of the key components.

CCS Concepts: • **Computing methodologies** → **Information extraction**; **Natural language processing;**

Additional Key Words and Phrases: Document-level Relation Extraction, Separate Relation Representation, Mention-level, Logical Reasoning

ACM Reference format:
Heyan Huang, Changsen Yuan, Qian Liu, and Yixin Cao. 2023. Document-level Relation Extraction via Separate Relation Representation and Logical Reasoning. *ACM Trans. Inf. Syst.* 42, 1, Article 22 (August 2023), 24 pages.
https://doi.org/10.1145/3597610

1 INTRODUCTION

Document-level **Relation Extraction** (**RE**) is the task of identifying relations among entities within a document. Compared with sentence-level RE [14, 28, 29, 63], it is more realistic that entity/mentions may appear in different sentences, and it also serves as an essential step in real-world applications, such as question answering [7] and large-scale knowledge graph construction [25, 33, 57].

This work was supported by the National Natural Science Foundation of China (No.U21B2009), Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant, and Beijing Institute of Technology Southeast Academy of Information Technology.
Authors' addresses: H. Huang and C. Yuan (corresponding author), Beijing Institute of Technology, Zhongguancun North Street No. 5, Beijing, China; emails: {hhy63, yuanchangsen}@bit.edu.cn; Q. Liu, Nanyang Technological University, 50 Nanyang Street, Singapore; email: liu.qian@ntu.edu.sg; Y. Cao, Singapore Management University, 81 Victoria Street, Singapore; email: caoyixin2011@gmail.com.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Although sentence-level RE has been successful, document-level RE has to face a new challenge: multi-mention multi-label. To solve this problem, there are two main existing works: Building graphs that capture complex connections among mentions and entities, and reasoning skills that bridge the gap in the relation information extracted by themselves. Recent researchers' attention is to use the **Graph Neural Network** (**GNN**) to construct a mention graph and capture complex interactions among mentions across sentences from the document [60]. Ru et al. [34] proposes a probabilistic model to learn logical rules and capture dependencies between entities and output relations. To heuristically select evidence sentences, Huang et al. [16] employs local path enhanced RE, and Xu et al. [49] models the reasoning skills between two entities. Zeng et al. [59] separates intra-sentence and inter-sentence reasoning, and Li et al. [21] utilizes a locally and globally mention-based reasoning to predict relation. Besides, adaptive threshold technique has been proposed for multi-label classification [65]. However, building on top of them, we further highlight the following two challenges based on our observations:

Indirect relation representation. In the document-level RE, the relations are usually embedded in mention pairs. There are two reasons: (1) In the process of labeling data, it is often subconsciously assumed that if the mention pair has a relation, the entity pair will be marked as having such a relation. (2) Different mention pairs of the entity pair may refer to varying relations due to their various context, which in turn leads to diverse relations of entity pair. Therefore, there is a difference between entity pairs and mention pairs when it comes to relation representation, i.e., mention pairs are direct relation representation while entity pairs are indirect relation representation. As shown in Figure 1, entity pair, `West Virginia` and `United States`, has three relations (i.e., `country`, `located in the administrative territorial entity`, and `country of citizen`). `West Virginia` has two mentions, which are at 0th and 2th sentences. Entity pairs (`West Virginia` and `United States`) are labeled as relations because these relations exist in the mention pair consisting of `West Virginia` of the 0th sentence and `United States` of the 0th sentence, while `West Virginia` in the 2th sentence does not play any role. **Therefore, is there any way to efficiently bridge the gap from the indirect entity pairs to the direct mention pairs for relation representation?**

Massive Interaction Support. Probing and inferring evidence sentences are crucial to identifying the correct relation. There is some overlap and interaction among the evidence sentences, which can be used to infer and support the relations of different entity pairs. As shown in Figure 1, the entity pairs, `(Washington Place, United States)`, and `(Hampshire Country, United States)`, have the same relation (`Country`) and evidence sentences, and they can connect and infer relation information. Each entity pair can extract unique relational features from the evidence sentences, and these features with the same relations have a certain connection. However, the 0th sentence can support three relational facts for eight entity pairs. More noise inference will be elicited if the evidence sentences information about eight entity pairs interacts simultaneously. **Therefore, can we extract useful information from evidence sentences and reduce unnecessary interactions to improve the performance of RE?**

To address the issues, we propose a novel **S**eparate Relation **R**epresentation and **L**ogical **R**easoning for document-level RE, namely SRLR. The basic idea is to separate the entity pair to the mention-level that can capture more fine-grained relation representation and elicit logical reasoning based on evidence sentences that can employ potential reasoning information to enhance the performance of RE. To address the first challenge for indirect relation representation, we hierarchically integrate the information of the relational facts to comprehensively understand and capture relation semantics. To address the second challenge of massive interaction support, we further highlight the evidence sentences and infer relation information based on the evidence sentence. Conditioned on the constraints, only relations with overlapped entities and entity types shall be

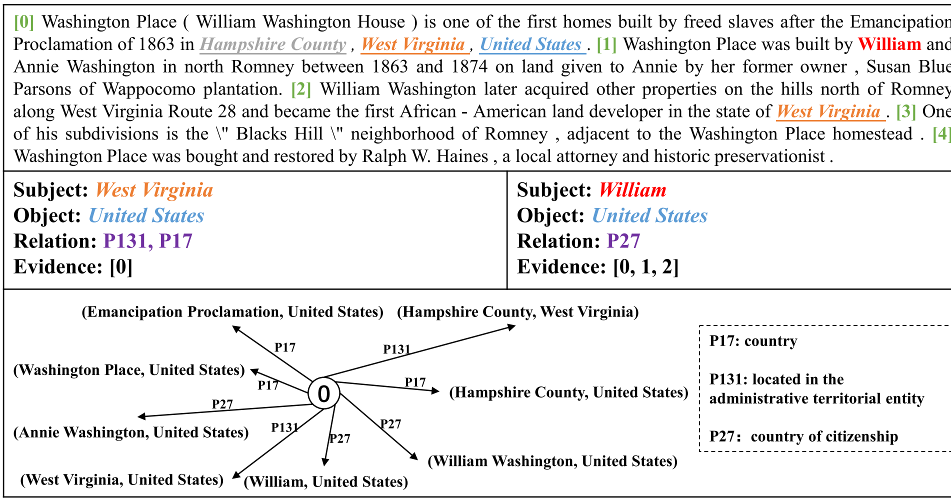

Fig. 1. An example of document-level RE is excerpted from the DocRED dataset. Different entity pairs containing the same entity may indicate the same relation, but the same evidence sentence may support different relations. The sentence 0 supports three relations of eight entity pairs.

visible to each other to mitigate unnecessary interactions. We conduct extensive experiments on two widely-used benchmarks, including DocRED [53] and CDR [20]. The results show that SRLR has achieved better performance.

The contributions of this article can be summarized as follows:

- We propose the Separate Relation Representation to capture the multi-level relation representation, which can solve the indirect representation for the entity pairs.
- We develop the Logical reasoning that can utilize the valid evidence sentences and logical inference to enhance the performance of document-level RE.
- We conduct extensive experiments on two publicly available datasets. Compared with baseline models, the results show a superior performance of our method, and further ablation study demonstrates the effectiveness of the key components.

The rest of this article is organized as follows. Section 2 describes the related work about sentence-level RE and document-level RE. In Section 3, we briefly describe the preliminaries and framework of our model. In Sections 4 and 5, the Separate Relation Representation and Logical reasoning model is proposed and presented its inference process. In Section 6, we report the results of the experiment and conduct the various ablation studies. Finally, the conclusion of this work is made in Section 7.

2 RELATED WORK

2.1 Sentence-level RE

Sentence-level RE [8, 27] is to identify relational facts between entities from intra-sentence, which is divided into supervised RE [3], semi-supervised RE [2, 4, 13, 31], unsupervised RE [5, 51], and distant supervised RE [22, 32, 38, 42]. Currently, the researchers pay more attention on supervised RE and distant supervised RE.

Supervised RE relies on a large number of high-quality datasets with labels to predict the relations. Christopoulou et al. [9] makes up to l–length walks between entity pair to distinguish different relation paths between two entities. To introduce and alleviate the confusion of dependency information, Miwa and Bansal [29] proposes to incorporate the pruned external syntactic information into the lowest common ancestor subtree, and Yan et al. [50] proposes to employ the shortest dependency path to predict relations. However, these strategies are too hard and lose some important contexts or elicit more noise. Zhang et al. [63] adopted **graph convolutional network** (**GCN**) to learn the dependency features and utilize the tradeoff pruning strategy to capture the important information. Guo et al. [14] utilize the multi-head attention to build the soft graph to learn the semantic information. Besides, there are other models with graph that use dynamic pruning [56], sub-graph [58], and attention graph [6] to encode dependency features. However, the limitation of supervised RE is that requires a large amount of high-quality manually labeled data, which exceeds consuming-time and expensive.

Distant supervised RE [28, 62] is an efficient method to automatically construct a large number of relation extraction datasets in a short period of time, but the datasets contain a lot of noisy information. The study of distant supervised RE focuses on how to resolve noisy information in the datasets. Some researchers use the multi-instance multi-label learning to alleviate the noisy information in the datasets, and the term "multi-instance learning" is proposed to predict the drug activity [12]. In the multi-instance learning, the sentence containing the same entity pair can be considered as a bag, and using the label of the bag replaces the label of entity pair [12]. Ji et al. [18] also adopts a sentence-level attention strategy to learn the importance of sentences and assign the different weights to sentences by entity descriptions. Ye and Ling [55] adopts the intra-level and inter-level attentions to deal with the noise at sentence-level and bag-level, respectively. However, these methods of RE are sentence-level RE, but in practice entity pairs exist in the document and the model needs to consider the more complex scenarios.

2.2 Document-level RE

Recently works on document-level RE [10] often relies on explicit co-reference annotations or a single entity in the document [39, 52]. For example, Jia et al. [19] combines representations learned over various text spans throughout the document and across the sub-relation hierarchy. Verga et al. [46] proposes a Transformer-based model for document-level RE with multi-instance learning, merging multiple mention pairs. Yao et al. [53] proposes a new dataset (DocRED) about the document-level RE dataset and some baselines. This dataset is not limited to any specific domain, making it suitable to train and evaluate general-purpose document-level RE systems.

Based on DocRED, researchers exploit how to construct effective graph models to capture information between entities and mentions and how to effectively use reasoning to more accurately identify the relations between the entity pair. Nan et al. [30] proposes **Latent Structure Refinement model** (**LSR**) that enhances the relational reasoning across sentences. Relational reasoning can elicit the potential document-level graphs and utilize the refinement strategy to incrementally aggregate relevant information. Wang et al. [47] proposes a novel model to encode the document, which can integrate capture the contextual relation representation and entity representation from the local and global information. Zhou et al. [65] proposes adaptive thresholding and localized context pooling to learn an adaptive threshold for multi-label classification and transfer attention from pre-trained language models to locate relevant context. Zeng et al. [60] constructs two graphs: **heterogeneous mention-level graph** (**hMG**) and **entity-level graph** (**EG**), which can model complex interaction among different mentions and infer relations between entities, respectively. Makino et al. [26] treats the relation as a relation graph that can predict them by using complex graph information. The graph consists of entities and relations, where entities are nodes and relations are edges. Zhang et al. [61] exploits a novel method by decomposing document-level RE into identifying the local relations and reasoning to enhance the resolution, which can combine the explicit discourse model and self-supervision for each

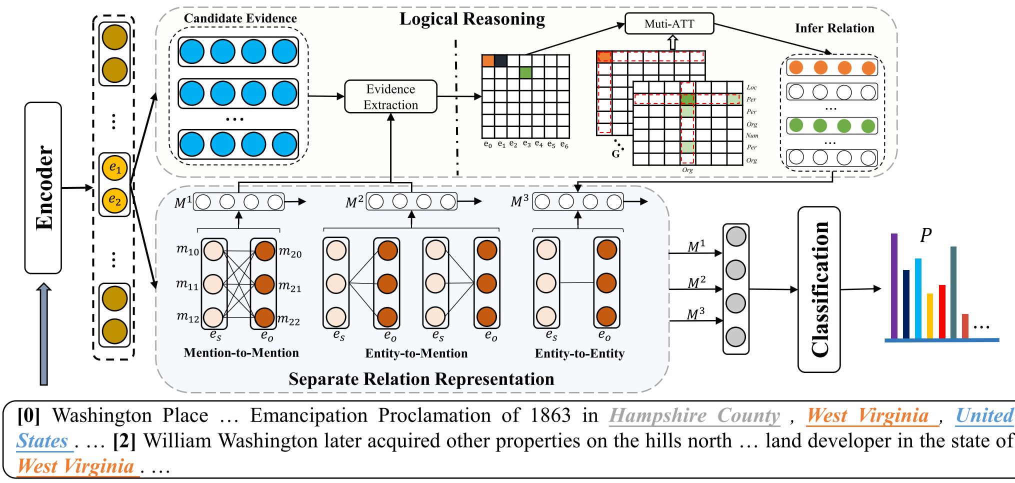

Fig. 2. Architecture of the proposed model. In the infer relation, the green square in the graph is the target relation. The light green squares are the candidate relations that can infer the target relation. The dashed boxes indicate the search range of the candidate relations based on the entity type.

sub-problem. Tran et al. [43] proposes to clearly compute the representations for the nodes to improve the graph-based edge-oriented model for document-level RE, which can deal with the shortages of more focusing on the edge representations and ignoring the node representations. Jia et al. [19] combines representations learned over various text spans throughout the document and across the sub-relation hierarchy to solve the limitation of short text spans in n-ary RE.

However, these works focus more on mention graphs, entity graphs, or semantic graphs, ignoring the representation of relational facts and reasoning of evidence sentences in the document-level RE. In this article, the significant difference between our model and previous models is that we separate the relational facts to the mention-level to capture more accurate and rich relational information. And using the evidence sentences to build a valid graph, which can reason the relations for the entity pairs.

3 PRELIMINARIES AND FRAMEWORK

3.1 Preliminaries

Given a sequence of document $D = [w_1, w_2, \ldots, w_n]$ and entities $[e_1, e_2, \ldots, e_N]$, where n is the length of document, and w_i is the token in the document, the goal of the document-level RE task is to identify the relation (r) between the entity pair (e_s, e_o), namely $P(r|e_i, e_j)$, which may be in the same sentence or scattered across the document. Any two entities in the document can form an entity pair, and the entity (e) has some mentions, $e = [m_1, m_2, \ldots, m_k]$, where k is the number of mentions in the entity. m_s and m_o are from the e_s and e_o respectively, and m_s and m_o can form the mention pair. And the entity pairs have some sentences to support the relation, which are called the evidence sentences.

3.2 Framework

The goal of SRLR is to capture the multi-level relation representation via learning the relational facts of mention pairs and learn the valid information of supportive evidence sentence via logical reasoning to enhance the performance of document-level RE. As shown in Figure 2, our framework

Table 1. Notations in Our Model

Name	Description
n	Document Length
w_i	Token
H	Hidden Matrix
h_i	Hidden Vector of i^{th} token
A	Token-level Attention Heads
m_i	i^{th} mention of entity
$m^{'}$	Entity that denotes Multiple mentions
M^1	Relation representation of Mention-to-Mention
M^2	Relation representation of Entity-to-Mention
M^3	Relation representation of Entity-to-Entity
S	Sentences about Candidate evidence sentences
c	Contextual Relation Matrix
c_e	Inference Contextual Relation Matrix

has two steps: **Separate Relation Representation** and **Logical Reasoning**. The notations and symbols frequently used in this article are displayed in Table 1.

Separate Relation Representation (**SR**) aims at aggregating the multi-level relation representation based on mentions, which mainly contains three components: (1) **Mention-to-Mention** that is the combination of any two mentions in the entity pair. (2) **Entity-to-Mention** which is the combination of an entity and all mentions in another entity. (3) **Entity-to-Entity** that is the combination of two entities. The multi-level relational information can capture the fine-grained and rich relational facts to improve the relation representation.

Logical Reasoning (**LR**) aims at efficiently utilizing complicated evidence sentences and entity pairs to infer the relation. LR has two steps of **Evidence Extraction** (**EE**) Module and **Infer Relation** (**IR**) Module: (1) EE can identify the candidate evidence sentences and capture the sentence supporting relation information. (2) IR utilizes the entities and entity types to build the reason graph, then selects the candidate relations to infer the target relation.

4 SEPARATE RELATION REPRESENTATION

This section introduces SR and how to aggregate inside information to capture multi-level information of the document. First, we adopt the encoder module to model the document and obtain the hidden information, H and A, the document, and attention features. Then, we employ the SR module to capture the significant relation information.

4.1 Encoder Module

In this section, we utilize the BERT [11] to extract crucial information, which can be used to identify the relations. To make use of sub-word information, we tokenize the document by **byte pair encoding** (**BPE**) [36], $D = [w_1, w_2, \ldots, w_n]$, where w_i is the ith token in the document. To solve the excessively long document (Over 512 tokens), we follows the previous work [37] to take the first 512 tokens and the last 512 tokens as the document's information into the BERT model separately, then concatenate these two parts together as the information of the whole document information. And we also mark the positions of mentions by inserting a special symbol "*" at the start and end [37, 64]. For example, the entity is "New York". When the special symbol "*" is added, the entity becomes "* New York *". In special, we briefly represent the process of encoding by BERT, which can be computed as follows:

$$H, A = \text{BERT}([w_1, w_2, \dots, w_n]), \tag{1}$$

where $H = [h_1, h_2, \dots, h_n]$ is considered as the hidden matrix of tokens. A is the average of the attention heads in the last transformer layer.

For mention embedding, we follow the method used in [53]. And for sentence embedding, we have experimented with mean, max, and add to compute it. And we found that these different representation methods had little effect on the experiment results, but max-pooling was the best. Thus, we employed max to compute the sentence embedding. m_j is the jth mention in the entity from the cth token to the dth token, and s_i is the ith sentence in the document from the ath token to the bth token, which can be represented as

$$\begin{cases} m_j = \frac{1}{d-c+1}\sum_{j=c}^{d} h_j \\ s_i = \text{Max}([h_a, \dots, h_b]) \end{cases}. \tag{2}$$

4.2 Hierarchical Aggregation Module

The purpose of the **hierarchical aggregation** (**HA**) module is to explore the rich and effective representation of relational facts for the entity pair. The conventional sentence-level RE using the entity pairs to represent the relational facts does not work well for document-level RE. There are two reasons: **(1) In traditional sentence-level RE, entity pairs and mention pairs are usually the same, and the entity pairs can represent the relations well.** But there are multiple mentions in the document-level RE, and it is difficult to directly determine which mentions can represent the relations. **(2) People subconsciously label relational facts based on mentions in the document.** This subconscious labeling behavior leads to indirect relation representation for the relational fact based on the entity pair, which also can be called a gap between entity pairs to mention pairs. Because if mentions contain some relations with each other, then entity pairs can represent these relations. The relation representation of the entity pair is represented by mention pairs. Therefore, using entity pairs to represent relational facts is too rough and indirect.

To bridge the gap between entity pairs and mention pairs, the hierarchical aggregation module traverses the possible representations of the relational fact based on mentions, which are Mention-to-Mention, Entity-to-Mention, and Entity-to-Entity.[1] Mention denotes the one mention, and Entity denotes the information of multiple mentions in the entity. We highlight the core idea of HA — **Iterate through all possible relational representation features in the entity pair**. Therefore, these three forms can capture the multi-level useful representation of the relational fact for the entity pair (e_s, e_o), which are defined as follows:

- **Mention-to-Mention.** Any Mention in e_s, e_o is combined two by two to form multiple mention pairs to represent the relational fact. Suppose that the relational fact exists in Mention-to-Mention pairs, and their representation is one-to-one. For example, given two entity ($e_1 = [m_1, m_2, m_3]$ and $e_2 = [m_4, m_5, m_6]$) that contain 3 mentions respectively, the mention-to-mention has 9 mention pairs, which can express as follows: $(m_1, m_4), (m_1, m_5), \dots, (m_2, m_6), (m_3, m_6)$.
- **Entity-to-Mention.** Entity and Mention are from e_s, e_o, respectively. Suppose that the relational fact exists in Entity-to-Mention pairs, and it is multiple mentions and single mention to represent the relational fact. For example, given two entity ($e_1 = [m_1, m_2, m_3]$ and $e_2 = [m_4, m_5, m_6]$) that contain 3 mentions respectively, the entity-to-mention has 6 pairs, which can express as follows: (e_1, m_4),(e_1, m_5),(e_1, m_6),(m_1, e_2),(m_2, e_2), (m_3, e_2).

[1]The relation are orientational, and different order of the entity pair indicates different relations.

— **Entity-to-Entity.** Entity and Entity are from e_s, e_o, respectively. Suppose that relational fact exists in Entity-to-Entity pairs, and it is multiple mentions and multiple mentions to represent the relational fact. For example, given two entity (e_1 and e_2), the entity-to-entity has 1 pair, which can express as follows: (e_1, e_2).

4.2.1 Mention-to-Mention. Given an entity pair (e_s, e_o) that contains t mention pairs, we employ the model to automatically assign the different weights (α^1) to mention pairs. The representation of relational fact, M^1, can be computed as follows:

$$\begin{cases} M^1 = \alpha^1 \widehat{M}^1 \\ \alpha^1_{ij} = \frac{exp(W^1 \widehat{M}^1_i)}{\sum_{j=1}^{t} exp(W^1 \widehat{M}^1_j)} \end{cases}, \tag{3}$$

where $\widehat{M}^1$ is the set of relation representations of mention pairs in the **Mention-to-Mention**. W^1 is the trainable parameters. The relation information of ith mention pair can be computed as follows:

$$\begin{cases} \widehat{M}^1_i = \tanh(z^1_s W_b z^1_o) \\ z^1_s = \tanh(W_1 m_s + W_c c^1) \\ z^1_o = \tanh(W_2 m_o + W_c c^1) \end{cases}, \tag{4}$$

where $W_b \in \mathbb{R}^{d\times d}$, $W_1 \in \mathbb{R}^{d\times d}$, $W_2 \in \mathbb{R}^{d\times d}$, and $W_c \in \mathbb{R}^{d\times d}$ are trainable parameters. The m_s and m_o are mentions in the e_s and e_o, respectively, and m_s and m_o are the ith mention pair in the Mention-to-Mention. c is the contextual relation matrix by context pooling [65], which can enhance the embedding of the entity pair with additional contextual information without using external knowledge (e.g., syntactic, discourse, or coreference). We use the token-level attention heads, A, from the last transformer layer (See Equation (1)). The contextual relation matrix, c, is computed as follows:

$$\begin{cases} A^{(s,o)} = A^s A^o \\ q^{(s,o)} = \sum_{i=1}^{L} A_i^{(s,o)} \\ \beta = \mathrm{Mean}(q^{(s,o)}) \\ c^1 = H^T \beta \end{cases}, \tag{5}$$

where β is attention weight for tokens in the document. Mean is the average operation. H is the document embedding. L is the number of heads in the transformer. A^s and A^o is the attention matrices for mention pairs.

4.2.2 Entity-to-Mention. Given an entity pair (e_s, e_o), and suppose that multiple mentions of e_s and certain mention in the e_o can express the relation, which contains t entity-to-mention. Therefore, we first compute the information of multiple mentions, **Entity**. Specially, we use **graph attention networks** (**GAT**) [45] to capture the interaction of mentions in the e_s, and the graph of GAT is a fully connected graph, which connects the target mention node with each mention nodes in the same entity. Then, we use the mentions in the e_o as queries to collect the information of multiple mentions from the e_s, which can be computed as follows:

$$\begin{cases} m_s^{'} = \lambda \widehat{e}_s \\ \lambda_i = \frac{exp(m_i q)}{\sum_{i=1}^{k} exp(m_i q)} \\ \widehat{e}_s = [m_1, m_2, \ldots] \\ \widehat{e}_s = \mathrm{GAT}(e_s) \end{cases}, \tag{6}$$

where $m_s^{'}$ is the **Entity** representation. k is the number of mentions in the e_s. m_i is the mention from the $\widehat{e}_s$, and q is the mention in the e_o. Then, we compute the relation representation of i^{th}

Entity-to-Mention, which can be computed as follows:

$$\begin{cases} \widehat{M}_i^2 = \tanh(z_s^2 W_b z_o^2) \\ z_s^2 = \tanh(W_1 m_s' + W_c c^2) \\ z_o^2 = \tanh(W_2 m_o + W_c c^2) \end{cases}, \tag{7}$$

where Equations (4) and (7) share the same parameters. And we also use the Equation (5) to compute the contextual relation matrix (c^2), but for the attention of **Entity** (A^s), we average the attention for all mentions of the same entity (e_s) to obtain entity attentions.

Finally, the representation of relational fact about entity-to-mention, M^2, can be computed as follows:

$$\begin{cases} M^2 = \alpha^2 \widehat{M}^2 \\ \alpha_{ij}^2 = \frac{exp(W^2 \widehat{M}_i^2)}{\sum_{j=1}^{t} exp(W^2 \widehat{M}_j^2)} \end{cases}, \tag{8}$$

where W^2 is the trainable parameters.

4.2.3 Entity-to-Entity. Entity-to-Entity that contains many mentions can be considered as multi-to-multi format. For the entity with mentions, we use the GAT [45] and logsumexp pooling to get the entity embedding, $\widehat{e} = \log \sum exp(\text{GAT}(e))$. Then, predicting the relational fact, which can be computed as follows:

$$\begin{cases} M^3 = \tanh(z_s^3 W_b z_o^3) \\ z_s^3 = \tanh(W_1 \widehat{e}_s + W_c c^3) \\ z_o^3 = \tanh(W_2 \widehat{e}_o + W_c c^3) \end{cases}, \tag{9}$$

where c^3 is the contextual relation matrix. Equations (9) and (4) share the same parameters. For the attention of A^s and A^o about Entity-to-Entity, we average the attention for all mentions of the same entity (e_s, or e_o) to obtain entity attentions, then c^3 can be computed by Equation (5).

5 LOGICAL REASONING

LR module aims at capturing the potential and critical information of evidence sentences as the reasoning information and infer the relations, which contains two modules, i.e., Evidence Extraction Module and Infer Relation Module.

5.1 Evidence Extraction Module

Evidence extraction module aims at capturing the candidate evidence sentences, then enhance the representation of the relational information of entity pairs, which has two steps: the first step is to select candidate evidence sentences. And suppose that sentences where mentions are located are candidate evidence sentences, which have the high probability to help the model identify the relation. Statistically, the candidate evidence sentences cover 90% of the true evidence sentences in Table 7. Inevitably, there are more numbers candidate evidence sentences than true evidence sentences. Second key step of evidence extraction is to reduce the noisy candidate evidence sentences. For example, the candidate evidence sentences for entity pair (`West Virginia`, `United States`) are sentence 0th and 2th, which contain noisy sentence 2 in Figure 1.

Given an entity pair (e_s, e_o), its candidate evidence sentences are $S = \{s_i\}_{i=1}^{C}$, where C is the number of sentence, and s_i can be computed by Equation (2). Then, we consider hierarchical aggregation information (M^1 and M^2) as the query to reduce the noisy features of candidate evidence sentences, which can be computed as follows:

$$\begin{cases} \widehat{S} = \gamma S \\ \gamma_i = \frac{exp(s_i M^r)}{\sum_{j=1}^{C} exp(s_j M^r)} \\ M^r = W^r [M^1, M^2] \end{cases}, \tag{10}$$

where $W^r \in \mathbb{R}^{2d \times d}$ is the trainable parameters. Therefore, we can obtain the evidence sentences representation $\widehat{S}$ about the entity pair.

5.2 Infer Relation Module

IR Module explores to employ valid and important evidence sentences from non-target entity pairs to enhance the relational representation of the target entity pair. The core idea is that **the evidence sentences of different entity pairs contain different densities of relational information, and the amount of high-density information is used to enhance the amount of low-density information.** The Since many entity pairs have the same relation and similar evidence sentences, the relation can share some relation information, which can be inferred by evidence sentences. Therefore, to capture and learn the high-quality information of evidence sentences, we should solve two problems:

— Which evidence sentences of entity pairs (except target entity pair) can enhance the relation representation of target entity pair?
— How to fuse the information of multiple evidence sentences and improve the performance of the model?

For the first problem, we select candidate evidence sentences based on the following two assumptions: **(1) The evidence sentence relation information about the target entity pair contains the same head entity or tail entity.** Because entity pairs have the strict positional order. Assuming that an entity pair (e_s, e_o) can represent a relation r, it is certain that the entity pair (e_o, e_s) cannot represent the relation r. **(2) The candidate entity pairs have the same entity types.** As shown in Figure 1, there are three entity pairs with the same relation (`country of citizenship`), which are (`Annie Washington`, `United States`), (`William`, `United States`), and (`William Washington`, `United States`). They have the same entity (`United States`) and entity types (`Person` and `Location`). Meanwhile, although these eight entity pairs have the same entity (`United States`), the entity types further filter the noisy evidence sentences of entity pairs to get the relatively accurate evidence sentences, such as (`Emancipation Proclamation`, `United States`) and (`Hampshire County`, `United States`).

For the second problem, we first build a reason graph, G (G is the 0/1 matrix), based on above assumptions for each entity pair, which can infer target relation information, as shown in Figure 3. Second, we consider the entity-to-entity matrix (See relation matrix in Figure 3) as a pixel, inspired by [17]. Finally, we use multi-head attention [44] to infer the relation matrix,[2] c_e, which can be computed as follows:

$$\begin{cases} c_e = AX \\ X = \widehat{S} + c \\ Q = XW_q, K = XW_k, V = XW_v \\ A = softmax(\frac{QK^T}{\sqrt{d_k}} G) \end{cases}, \tag{11}$$

where c_e is the relation matrix. W_q, W_k, and W_v are the trainable parameters. d_k is the dimension of X. c is the contextual relation matrix (See Equation (5)).

[2]In this part, we use two layers of multi-head attention to compute the relation matrix.

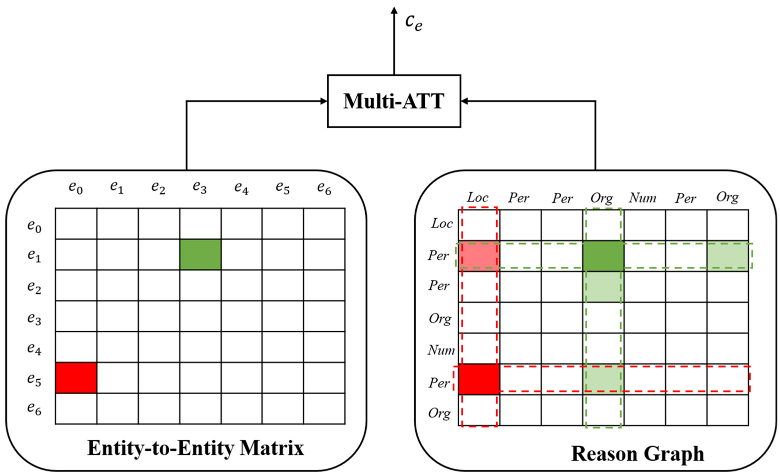

Fig. 3. Infer Relation Module based entities and entity types. The green square in the reason graph is the target relation. The light green squares are the candidate relations that can infer the target relation. The green dashed boxes indicate the search range of the candidate relations based on the entity type.

To infer the target relation through other crucial evidence sentences and entity pairs, the LR is proposed and used in Equation (9) to modify the original relation matrix, c, to the new matrix, c_e. Notes that the logical reasoning is used to enhance the contextual relation matrix, which is only used in the **Entity-to-Entity**. There are two reasons: (1) The contextual information of **Mention-to-Mention** and **Entity-to-Mention** does not contain the complete information of the evidence sentences. (2) The candidate evidence sentences consider all the information of the sentences where mentions are located and conforms to the hypothesis of **Entity-to-Entity**. If the logical reasoning is used in **Mention-to-Mention** and **Entity-to-Mention**, the information of multiple mentions will be applied to them, not conforming to the assumption of single mention information. As shown in Figure 1, the **Mention-to-Mention** (`West Virginia`, `United States`) are in the 2th and 3th sentences, and the candidate evidence sentences are 2th and 3th sentences, ignoring the true evidence sentence 0th. All that the model learns is noisy information from these evidence sentences. But the **Entity-to-Entity** can consider 0th, 2th, and 3th sentences.

Discussion of assumption. Since different entity pairs in the document can express the same relation using different evidence sentences, and the power of the evidence sentences' ability to represent the relations also varies. We design these two hypotheses here to capture as many different evidence sentences of the same relation as possible, enhance the relational representation of the target entity pairs, and improve the prediction ability of the model. Although these two hypotheses do not completely satisfy the filtering of all evidence sentences, they are sufficient for most cases. The core assumption of our design — introducing different evidence sentences of the same relation into the entity representation can enhance the prediction of the model. In the experiments, we also demonstrate that using the evidence sentences is indeed useful.

5.3 Classification Module

For the entity pair (e_i, e_j), we concatenate the following representations in Section 4: M^1, M^3, and M^3, which can be expressed as followed:

$$I_{ij} = [M^1, M^2, M^3]. \tag{12}$$

Table 2. Statistics of DocRED and CDR in Experiments

Statistic	DocRED	CDR
# Train	3,053	500
# Dev	1,000	500
# Test	1,000	500
#Relation Types	97	2
# Avg.# entities per Doc	19.5	7.6
# Avg.# sentences per Doc	8.0	9.7

Finally, we formulate the task as multi-label classification and predict relations between entities:

$$P(r|e_i, e_j) = \text{Sigmoid}(W_b \tanh(W_a I_{ij} + b_a) + b_b), \tag{13}$$

where $W_a \in \mathbb{R}^{3d \times d}$, b_a, W_b, and b_b are the trainable parameters. The model parameters are estimated by Adaptive Thresholding [65], which learns an adaptive threshold for each entity pair and alleviate the unbalanced relation distribution. Specially, [65] design a threshold TH, which can dynamically separate the positive labels and negative labels. Positive labels would have higher probabilities than TH, and negative labels would have lower probabilities than TH. The Adaptive Thresholding loss is formulated as follows:

$$\begin{cases} L = L_1 + L_2 \\ L_1 = -\sum_{r \in P_T} \log\left(\frac{exp(\text{logit}_r)}{\sum_{r' \in P_T \cup \{TH\}} exp(\text{logit}_{r'})}\right) \\ L_2 = -\log\left(\frac{exp(\text{logit}_{TH})}{\sum_{r' \in N_T \cup \{TH\}} exp(\text{logit}_{r'})}\right) \end{cases}, \tag{14}$$

where P_T and N_T are the positive labels set and negative labels set, respectively In this article, we employ dropout to prevent overfitting.

6 EXPERIMENT

In this section, we will first describe the datasets and experimental details, then report the experimental results and analysis. We also conduct a case study.

6.1 Dataset and Metrics

To demonstrate the performance of our model, we conduct the experiments on the DocRED[3] [53] and CDR[4] [20]. The detailed statistic of the datasets is shown in Table 2.

DocRED. DocRED the largest document-level dataset from the Wikipedia, which annotates the entities and relations. An initial dataset is first constructed using the distant supervised method. Then, the distant supervised dataset is manually re-labeled to ensure the correctness of the dataset. For dataset (re-labeled), DocRED contains 132, 375 entities, and 56, 354 relational facts annotated on 5, 053 Wikipedia document. There are more than 40.7% of relational facts in the DocRED, which should employ multiple sentences to identify their relations. And the dataset contains 96 relations, covering most categories, and the relational categories are distributed in various fields such as science, art, people, and so on.

CDR. The Biocreative V **Chemical Disease Relation benchmark** (**CDR**) [20] is a document-level RE dataset for the medical domain, which contains relations between chemicals and diseases, and CDR stems from the **Comparative Toxicogenomics Database** (**CTD**). And CDR contains

[3] https://github.com/thunlp/DocRED
[4] http://www.biocreative.org/

Table 3. Hyper-parameters of SRLR

	Hyperparameter
Batch size	2
Epoch	30
Learning rate for BERT	3e-5
Learning rate for fine-tuning	1e-4
Dropout	0.2
Action Function	Tanh
Embedding size	768

1, 500 documents with 4, 409 annotated chemicals, 5, 818 diseases, and 3, 116 chemical-disease interactions.

Following the previous work [53, 60], we utilize the F1 and Ign F1 as the evaluation metrics to measure the performance of RE models. Ign F1 is calculated excluding relational facts shared by the Train and Dev/Test. We also use the intra-F1 and inter-F1 metrics to evaluate a model's performance on intra-sentence relations and inter-sentence relations on the Dev.

6.2 Parameter Settings

Our model is implemented based on Pytorch[5] and Huggingface's Transformers.[6] We use uncased BERT-base [11] as the encoder on DocRED, and SciBERT-base [1] on CDR. We optimize our model with AdamW [24] using learning rate $3e-5$ for pre-trained and $1e-4$ for fine-tuning with a linear warmup for the first 6% of steps. We conduct grid search to tune hyperparameters: the size of relation representations is 768, which is selected from $\{64; 128; 256; 512; 768\}$. The batch size for pre-training is set to 4 for fine-tuning. The dropout rate is set to 0.2. We train our model with TITAN XP and perform early stopping based on the F1 score on the development set, with a maximum of 30 epochs. All the special tokens are implemented with unused tokens in the BERT vocabulary. We list the hyper-parameters on all datasets in Table 3.

6.3 Baseline

We compare SRLR with the following baselines.

- **Sequence-based Models.** Yao et al. [53] proposed four baseline models, which are CNN, LSTM, Bi-LSTM, and Context-Aware, respectively. Besides, we also adapt other sequence-based models to DocRED, which are HIN-GloVe [41] and LSR-GloVe [30].
- **Graph-based Models.** These models construct mention/entity graphs to capture the long-distance dependence and improve the performance of document-level RE, including GAT [45], BRAN [46], GCNN [35], EoG [10], and AGGCN [15].
- **Transformer-based Models.** These models utilize pre-trained language model to predict relations. BERT-RE$_{base}$, proposed by [48] on DocRED. BERT-Two-Step [48] is similar with BERT-RE$_{base}$, but it first predicts whether two entities have a relation and then identifies the specific relation. HIN-BERTb, proposed by [41]. **Hierarchical Inference Network** (**HIN**) captures the multi-level (entity level, sentence level, and document level) abundant information to predict the relations. LSR-BERT, proposed by [30]. LSR constructs a document-level graph for inference in an end-to-end fashion without relying on co-references or rules. Other pre-trained models like RoBERTa [23], ATLOP [65], GAIN [60], KD-DocRED-BERT [40], and CorefBERT [54] are also used as encoder [48] to document-level RE task.

[5]https://pytorch.org/
[6]https://github.com/huggingface/transformers

Table 4. Main Results on the Test of CDR

Model (%)	F1
CNN	52.60
BRAN	62.10
EoG	63.60
LSR	64.80
ATLOP-BERT$_{Base}$	69.40
SRLR-BERT$_{Base}$	**71.10**

Note that, our method (SRLR) is based on ATLOP [65] to solve the problems of Indirect relation representation and Massive Interaction Support. Therefore, the most important baselines is ATLOP.

6.4 Main Results

We show SRLR's performance on the DocRED and CDR[7] in Tables 4 and 5, in comparison with other baselines. We can see that:

— SRLR consistently outperforms all baselines on both datasets. This is mainly because our method can solve the issue of confusing relational facts and empower reasoning ability.
— In Table 5, all models perform better on Intra F1 than Inter F1. The reason is that the intra-sentential entity pairs can easily capture the high-quality information of relations. And inter F1 of SRLR is better than other baselines. This shows that our method can use high-quality intra-sentential information to enhance the inter-sentential information, which employs the dense relational information (intra-sentence) to infer the low-density relational information (inter-sentence).
— In Table 5, among the models not using BERT, SRLR-GloVe consistently outperforms all sequential-based and graph-based strong baselines by 0.76-2.64 F1 scores on the Test. Among the models using BERT or BERT variants, SRLR-BERT$_{Base}$ yields a great improvement of F1/Ign F1 on the Dev and Test by 1.13/0.91 and 0.58/0.43, in comparison with the strong ATLOP$_{Base}$, respectively. It suggests that SRLR is more effective in document-level RE tasks. We can also show that ATLOP$_{Base}$ improves Intra F1/Inter F1 by 0.29 and 2.36 on the Dev. These results demonstrate that separate relation representation (mention-to-mention, entity-to-mention, and entity-to-entity) and logical reasoning can efficiently capture the useful relational features and employ the reasoning to infer the target relation.
— We performed five training runs on both ATLOP and SRLR to obtain the mean and standard deviation of F1 in the multi-run setting. This is because ATLOP, which is an important attempt on DocRED, produces different results in each run. Since our method is based on ATLOP, we face the same issue and therefore conduct multiple runs. Our experimental results indicate that SRLR can achieve competitive performance.
— In Table 4, SRLR already outperforms all existing methods. Compared with CNN, BRAN, EoG, LSR, ATLOP, SRLR$_{Base}$ improves F1 by 18.10, 8.60, 7.10, 5.90, and 1.70 on the Test. The results demonstrate that our method can learn crucial relation representations from the document. More details about the information are extracted to enhance the performance of SRLR.

[7]Since the previous work on CDR only used F1 as the evaluation metric, we ignored the results of Intra F1, Inter F1, and Ign F1. And some methods (e.g., GAIN, HIN) are difficult to reproduce and do not use this dataset, the experimental results of these methods are missing.

Table 5. Main Results on the Dev and Test of DocRED

Model (%)	Dev				Test	
	Ign F1	Intra F1	F1	Inter F1	Ign F1	F1
Sequence-based Models						
CNN	41.58	51.87	43.45	37.58	40.33	42.26
LSTM	48.44	-	50.68	-	47.71	50.07
Bi-LSTM	48.87	57.05	50.94	50.94	48.78	51.06
Context-Aware	48.94	-	51.09	-	48.40	50.70
HIN_{Glove}	51.06	60.83	52.95	48.35	51.15	53.30
Graph-based Models						
GAT	45.17	58.14	51.44	43.94	47.36	49.51
GCNN	46.22	57.78	51.52	44.11	49.59	51.62
EoG	45.94	58.90	52.15	44.60	49.48	51.82
AGGCN	46.29	58.76	52.47	45.45	48.89	51.45
LSR_{Glove}	48.82	60.83	55.17	48.35	52.15	54.18
GAIN-BERT_{Glove}	53.05	61.67	55.29	48.77	52.66	55.08
SRLR_{Glove}	**53.74**	**61.98**	**55.91**	**49.52**	**53.66**	**55.84**
Transformer-based Models						
BERT-RE_{Base}	-	61.61	54.16	47.15	-	53.20
RoBERTa_{Base}	53.85	-	56.05	-	53.52	55.77
$\text{BERT-Two-Step}_{Base}$	-	61.80	54.42	47.28	-	53.92
HIN-BERT_{Base}	54.29	-	56.31	-	53.70	55.60
CorefBERT_{Base}	55.32	-	57.51	-	54.54	56.96
LSR-BERT_{Base}	52.43	65.26	59.00	52.05	56.97	59.05
ATLOP-BERT_{Base}†	59.11	67.26	61.01	53.20	59.31	61.30
GAIN-BERT_{Base}	59.14	-	61.22	-	59.00	61.24
$\text{KD-DocRED-BERT}_{Base}$†	**60.08**	**67.91**	<u>62.03</u>	<u>55.01</u>	<u>59.72</u>	<u>61.71</u>
SRLR-BERT_{Base}	<u>60.02</u>	<u>67.55</u>	**62.14**	**55.56**	**59.74**	**61.88**
$\text{ATLOP-BERT}_{Large}$†	61.32	68.05	63.18	56.10	61.39	63.40
GAIN-BERT_{Large}	60.87	-	63.09	-	60.31	62.76
$\text{KD-DocRED-BERT}_{Large}$†	62.09	68.21	64.04	56.84	**62.49**	**64.09**
SRLR-BERT_{Large}	**62.15**	**68.42**	**64.17**	**57.56**	<u>61.86</u>	<u>63.92</u>
Multi-run setting						
ATLOP-BERT_{Base}†	59.02±0.41	67.16±0.15	60.82±0.37	52.91±0.51	-	-
SRLR-BERT_{Base}	**59.93±0.32**	**67.35±0.23**	**61.87±0.58**	**55.26±0.27**	-	-

Results with † are computed based on re-trained models. Other results are reported in their original articles. The second-best results are <u>underlined</u>.

6.5 Ablation Study

To further analyze SRLR, we also conduct the ablation analysis to illustrate the effectiveness of different modules and mechanisms in SRLR. We show the results of the ablation study in Table 6. SRLR_{Base} had a more excellent performance on DocRED. This is because our model can capture the multi-level relational information and infer useful information.

Separate Relation Representation. We examined the performance of the models in terms of Hierarchical Aggregation. In Table 6, we find that:

— RE performance is sharply poor without **Separate Relation Representation**. Specifically, the experimental results are reduced by 5.83/5.30 F1/Ign F1 and 5.09/5.26 F1/Ign F1 on the Dev and Test setting in SRLR_{Base}, respectively. There are two reasons: (1) Entities and mentions contain rich and useful information for identifying relations. (2) The same evidence

Table 6. An Ablation Study for the SRLR$_{base}$ Model on Dev

Model	Dev				Test	
	Ign F1	F1	Intra F1	Inter F1	Ign F1	F1
Our Method	**60.02**	**62.14**	**67.55**	**55.56**	**59.74**	**61.88**
w/o Separate Relation Representation	54.72	56.31	61.22	50.04	54.48	56.79
w/o Mention-to-Mention	59.58	61.47	67.57	54.20	59.49	61.25
w/o Entity-to-Mention	59.11	61.36	67.14	53.17	59.25	61.29
w/o Entity-to-Entity	58.91	60.61	66.58	53.47	59.53	61.04
w/o Logical Reasoning	59.20	61.17	67.31	53.42	59.42	61.09
w/o Evidence Extraction	59.64	61.43	67.42	54.48	59.57	61.29
w/o Infer Relation	59.40	61.22	67.29	53.98	59.40	61.18

sentences can support different labels of entity pairs. Therefore, it is difficult to predict the relation based on logical reasoning, ignoring separative relation representation.

— Without **Mention-to-Mention**, **Entity-to-Mention**, and **Entity-to-Entity**, the experimental results all showed different degrees of decrease. These results demonstrate that the representation of relational facts in the document is complex, and using a single representation (entity pairs) can lead to useful relational information being ignored, resulting in poor experimental results. Therefore, there is a gap between the entity pairs and mention pairs, and this gap should be bridged to improve the performance of document-level RE.

Logical Reasoning. To assess the effectiveness of SRLR in Noise Reduction, we measured the performance in different components. In Table 6, we can see that:

— When the **Logical Reasoning** is removed, the results of the experiment are reduced by 0.97/0.82 F1/Ign F1 and 0.79/0.32 F1/Ign F1 on the Dev and Test setting in SRLR$_{Base}$, respectively. The major reason is that the candidate evidence sentences contain useful relational information, and the reasoning can use multiple candidate evidence sentences of more relational facts to infer the correct relation, which is different from the information of **Mention-to-Mention**, **Entity-to-Mention**, and **Entity-to-Entity**.

— The experimental effect is reduced by 0.71/0.38 F1/Ign F1 and 0.59/0.17 F1/Ign F1 on the Dev and Test settings, respectively, without the **Evidence Extraction** module. The reason is that the evidence sentences contain rich context and relational information that is different from entities.

— When we further remove the **Infer Relation** module, the results are dropped to 0.92/0.62 F1/Ign F1 and 0.70/0.34 F1/Ign F1 on the Dev and Test settings. There are two reasons. On the one hand, reasoning allows the model to capture candidate evidence sentences for multiple related relational facts, extending the perceptual range and depth of the model. On the other hand, some entity pairs contain the same relation and context information, even evidence sentences. Using the relational information of these entity pairs to infer the target entity pair can greatly improve the effectiveness of the experiment.

6.6 Effect of Evidence Sentences

To further analyze the effect of evidence sentences on the DocRED, the dataset is divided into three components according to the coverage of the evidence sentences: **Over**, **Equal**, and **Under**, which can be detailed description as follows:

— **Over** means that the number of the candidate evidence sentences is over the true evidence sentences, and the candidate evidence sentences contain all true evidence sentences.

Table 7. Statistics of Relations in the Train and Dev of DocRED

		Equal	Over	Under	Total
Train	Count	19,341	15,366	1,946	38,180
	Percent (%)	50.7	40.2	5.1	100
Dev	Count	6,354	4,864	606	12,323
	Percent (%)	51.6	39.5	4.9	100

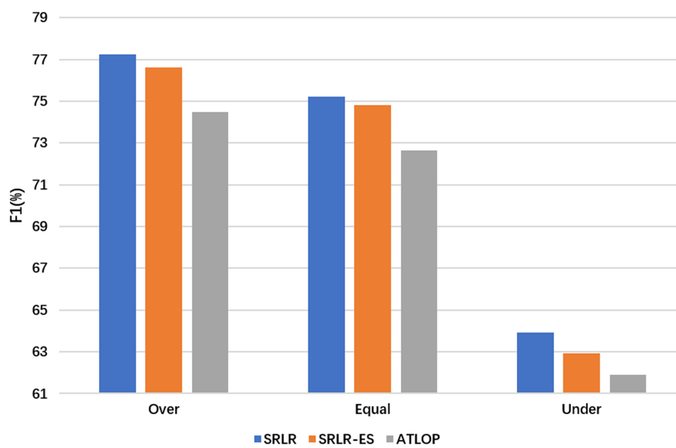

Fig. 4. Performance of SRLR, SRLR-ES, and ATLOP in F1 with three categories on the Dev. SRLR-ES is without evidence sentences.

— **Equal** means that the number of the candidate evidence sentences is equal to the true evidence sentences, and the candidate evidence sentences are exactly the same as true evidence sentences.
— **Under** means that the number of the candidate evidence sentences is less than the true evidence sentences, and the candidate evidence sentences don't contain all true evidence sentences.

The number and percentage of relations covered are listed in Table 7. We can clearly see that the candidate evidence sentences cover over 90% of relational facts in the Dev and Test. The results are shown in Figure 4. We can observe that:

— Our model has the best performance. The results demonstrate that evidence sentences are very important for identifying relations, and our method can extract key information from noisy candidate evidence sentences.
— For all our methods, the improvements over ATLOP is `Over>Equal>Under`, and the results of **Under** is worst than **Over** and **Equal**. The reason is that the **Under** evidence sentences are incomplete and contain more noisy sentences.
— When removing the evidence sentence, the experimental results demonstrate the validity of our approach in two ways: (1) The experimental results of SRLR were better than those of SRLR-ES, indicating that the evidence sentences play a positive role in inferring the entity pair. (2) SRLR-ES achieved higher performance than ATLOP, showings the effectiveness of multi-level information in identifying document-level relations. In addition, all experimental results were higher because we removed all entity pairs about the "NA" class due to their absence of evidence sentences.

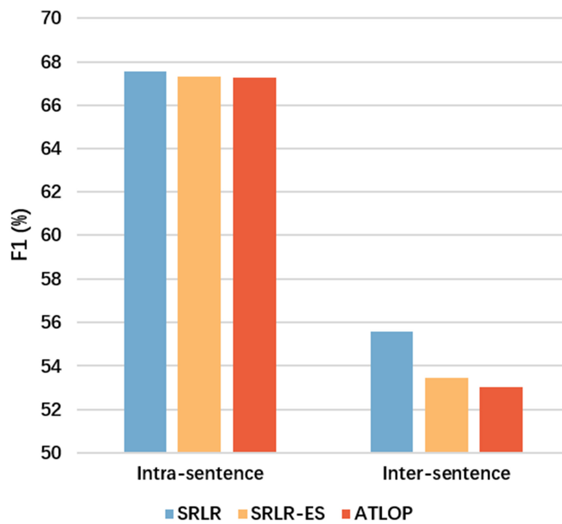

Fig. 5. Performance of SRLR, SRLR-LR, and ATLOP in F1 on the intra-sentence and inter-sentence of Dev. SRLR-LR is without logical reasoning.

6.7 Effect of Cross-sentence

To demonstrate the ability of our approach in dealing with cross-sentence relational facts, we divided the dataset into intra-sentence and inter-sentence for experiments in Figure 5 and obtained the following:

- On the intra-sentence, we are able to observe that SRLR and SRLR-LR have the similar experimental results. Because when the relational fact is intra-sentence, the entity pair is usually located in the evidence sentence, and the evidence sentence is the sentence where the entity pair is located. Therefore, the information of the entity pair contains the information of the evidence sentence, and ignoring the evidence sentence and logical reasoning has little effect on the experiment.
- Compared with ATLOP and SRLR-LR, they also has the similar performance. As Entity pairs and mention pairs are usually the same on the intra-sentence, the experimental effects of ATLOP and SRLR-LR are similar. Therefore, the effect of separate relation representation on intra-sentence is small.
- On the inter-sentence, SRLR has achieved the best performance than SRLR-LR and ATLOP. There are two reasons: (1) The inter-sentential relational information is complex and relies on information from multiple evidence sentences to correctly identify the relations of relational facts. (2) Reasonable reasoning of evidence sentences can also improve the accuracy of relation recognition to a great extent.
- The experimental results of SRLR-LR is better than ATLOP on the inter-sentence, even though both SRLR-LR and ATLOP ignore the evidence sentences. Because SRLR-LR solves the problem of indirect relation represenation, it also supports the validity of our approach to sink the relational identification down to the mention-level.

6.8 Effect of Entity pair≠Mention pair.

To analyze the effect of our method about indirect relational facts, we divided the dataset into **Mention pair=Entity pair** and **Mention pair≠Entity pair** for experiments, which can be detailed description as follows:

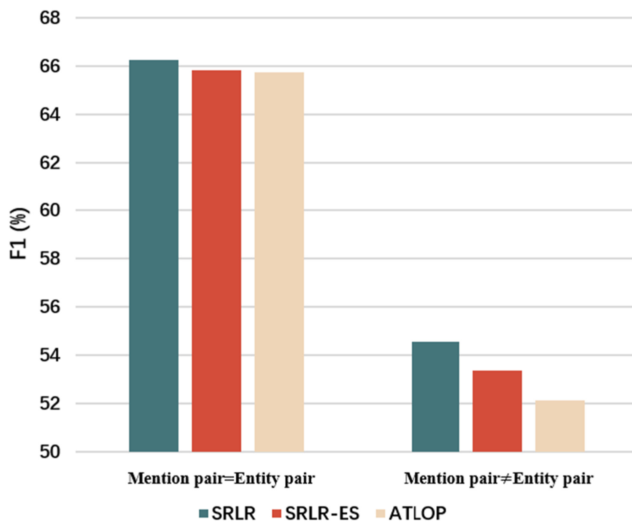

Fig. 6. Performance of SRLR, SRLR-LR, and ATLOP in F1 on the different setting of Dev. SRLR-LR is without logical reasoning.

- **Mention pair=Entity pair.** This denotes that mention pairs and entity pairs are exactly the same. The each entity in the entity pair only has a mention.
- **Mention pair≠Entity pair.** This denotes that all entity pairs except **Mention pair=Entity pair**, where at least one entity in the entity pair contains more than one mention.

We can find that in Figure 6:

- The experimental results of SRLR are better than SRLR-LR and ATLOP on the **Mention pair=Entity pair**, and SRLR-LR and ATLOP have the similar performance. Because **Mention pair=Entity pair** contains two cases, one where the entity pairs are intra-sentence and the other where the entity pairs are inter-sentence. Ignoring the logical reasoning leads to the drop of inter-sentential entity pairs. The inter-sentential entity pairs should focus more on multiple sentences and logical reasoning to predict the relation.
- Compared with **Intra-sentence** (Figure 5), their results of **Mention pair=Entity pair** (Figure 6) significantly drop. Because **Mention pair=Entity pair** contains Intra-sentence to a large degree and also contains entity pairs across many sentences. The relational information of cross-sentence needs to consider the multiple evidence sentences. But the Intra-sentence only considers one evidence sentence, and the entity contains contextual information, which is easy to learn.
- SRLR also has achieved the best performance than SRLR-LR and ATLOP on the **Mention pair≠Entity pair**. Entity pairs in the **Mention pair≠Entity pair** are multi-mention and cross-sentence. This demonstrates that our method can greatly improve the performance of the model when identifying cross-sentence entity pairs in the document.
- Compared with ATLOP, the result of SRLR-LR is better on the **Mention pair≠Entity pair**. Because SRLR-LR utilizes separative relation representation to reduce the impact of indirect relation representation. Moreover, the result of SRLR is better than SRLR-LR. Because SRLR employs logical reasoning to capture the potential reasoning and enhance the performance of document-level RE.

[0] Washington Place (William Washington House) is one of the first homes built by freed slaves after the Emancipation Proclamation of 1863 in ***Hampshire County*** , ***West Virginia*** , ***United States*** . … [2] William … Route 28 and became the first African - American land developer in the state of ***West Virginia*** . [3] One of his subdivisions is the " Blacks Hill " neighborhood of ***Romney*** , adjacent to the Washington …

Entity pair	Ground Truth	SRLR	BERT-RE
West Virginia, United States	*P131, P17*	*P17, P131*	*P17,* ***P131***
United States, West Virginia	*P150*	*P17* (✘)	*P17* (✘)
Hampshire County, West Virginia	*P131*	*P131*	*P131*
West Virginia, Hampshire County	*P150*	*P150*	*P150*
Romney, United States	*P17*	*P17*	*P27* (✘)

Fig. 7. The case study of our proposed SRLR and baseline models. ✘ denotes that the predicted label is wrong. The red words in the labels are not identified. The relations of P17, P131, and P150 are country, `located in the administrative territorial entity`, and `contains administrative territorial entity`, respectively.

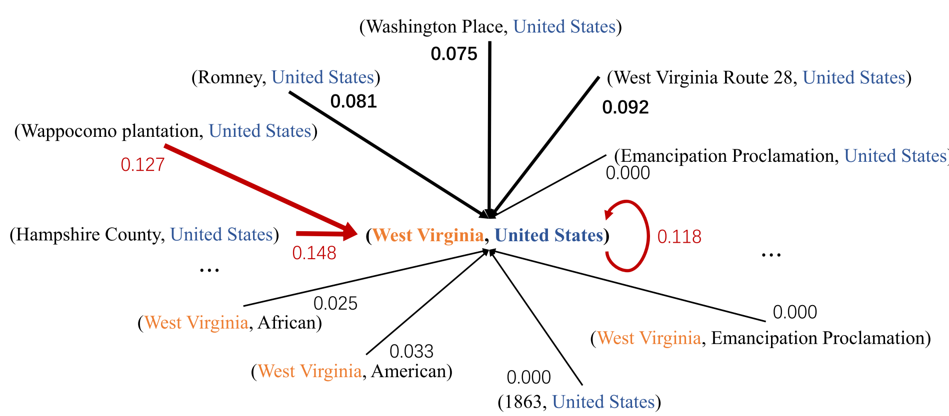

Fig. 8. The weight of reasoning attention about the event pair (`West Virginia`, `United States`) by SRLR.

6.9 Case Study

A qualitative analysis of our model on the DocRED dataset is shown in Figures 7 and 8.

The Figure 7 shows five entity pairs and their labels. We can observe that: Both SRLR and BERT-RE can identify the correct relations about (`United States`,`Hampshire Country`) and (`Hampshire Country`, `United States`). The reason may be that they are all in the same sentence, and each entity contains only one mention, more like a sentence-level RE. It is relatively easy to identify correct relations. However, BERT-RE and SRLR-BERT$_{base}$ do not recognize the relation between `United States` and `West Virginia`. SRLR-BERT$_{base}$ can obtain the correct relation of (`West Virginia` and `United States`). This indicates that different orders of the entity pair will have a serious influence, and this is a difficult point to research the document-level RE. Because `West Virginia` has two mentions in different sentences, and it is difficult to extract useful information across sentences. Overall, entities (`West Virginia`, `United States`) have different orders with three relations. BERT-RE did not predict the correct label when predicting multi-label the entity pair. This demonstrates that our method can capture more exact information and identify the right label in the multi-label problem. Meanwhile, (Romney, United States) and (West Virginia, United States) have the same relation (P17) and entity type (Loc, Loc). BERT-RE fails to identify the relation between Romney and United States, while SRLR-BERT$_{base}$ infer it successful. This indicates that our method can infer the target relation by other entity pairs, and logical reasoning is necessary.

Figure 8 shows the weights of logical reasoning for (West Virginia, United States), we can observe that: (1) Entity pairs with the same entity at the head or tail can effectively support the relation of the target entity pair. For example, (Hampshire County, United States) and (Wappocomo plantation, United States) have the same relation with (West Virginia, United States), and they have higher weights than other entity pairs. (2) The entity types of entity pairs is able to effectively reduce unnecessary reasoning. For example, the entity type of (1863, United States) is (Time, Loc), but (West Virginia, United States) is (Loc, Loc). Therefore, the reasoning weight of (1863, United States) is 0. In this document, there are 34 candidate entity pairs to reason the relation of (West Virginia, United States), but the number of entity pairs that can really reason the relation is six entity pairs, which is only 17.6%. When using the entity type to constrain the candidate entity pairs, the candidate entity pairs are 16, and the number of entity pairs that can really reason the relation is five entity pairs, which is only 31.3% of the candidate entity pairs. Although the restriction of entity type reduces the number of entity pairs that can actually reason the target entity pair, it greatly increases the proportion of candidate entity pairs that can reason correctly about the relation.

7 CONCLUSIONS

In this article, we exploit a novel Separate Relation Representation and Logical Reasoning Model, which can capture multi-level relation representation about the mentions and infer the information for the evidence containing the relational features. Experimental results show that our method is effective and significantly better than competitive baselines. The extensive analysis confirms that the evidence of multi-level information can enhance relation representation. And reducing the noisy information and inferring the relations further boosts the performance. Besides, it is an urgent issue to be solved that the same mentions, the same entities, and the same evidence sentences have more than one relation.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP*. 3613–3618.

[2] Sergey Brin. 1998. Extracting patterns and relations from the world wide web. In *Proceedings of the World Wide Web and Databases, International Workshop*. 172–183.

[3] Razvan C. Bunescu and Raymond J. Mooney. 2005. Subsequence kernels for relation extraction. In *Proceedings of the Advances in Neural Information Processing Systems*. 171–178.

[4] Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, and Wei Bi. 2021. Uncertainty-aware self-training for semi-supervised event temporal relation extraction. In *Proceedings of the Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2900–2904.

[5] Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zheng-Yu Niu. 2005. Unsupervised feature selection for relation extraction. In *Proceedings of the Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts*.

[6] Xiaolin Chen, Xuemeng Song, Ruiyang Ren, Lei Zhu, Zhiyong Cheng, and Liqiang Nie. 2020. Fine-grained privacy detection with graph-regularized hierarchical attentive representation learning. *ACM Transactions on Information Systems* 38, 4 (2020), 37:1–37:26.

[7] Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. UHop: An unrestricted-hop relation extraction framework for knowledge-based question answering. In *Proceedings of the NAACL-HLT*. 345–356.

[8] Stavros Christodoulakis, M. Theodoridou, F. Ho, Maria Pia Papa, and A. Pathria. 1986. Multimedia document presentation, information extraction, and document formation in MINOS: A model and a system. *ACM Transactions on Information Systems* 4, 4 (1986), 345–383.

[9] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2018. A walk-based model on entity graphs for relation extraction. In *Proceedings of the ACL*. 81–88.

[10] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the EMNLP*. 4924–4935.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT*. 4171–4186.

[12] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* 89, 1–2 (1997), 31–71.

[13] Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th International Conference on World Wide Web*. 100–110.

[14] Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. *ACL* 1 (2019), 241–251.

[15] Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the ACL*. 241–251.

[16] Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of the ACL/IJCNLP*. 998–1004.

[17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. CCNet: Criss-cross attention for semantic segmentation. In *Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision*. 603–612.

[18] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3060–3066.

[19] Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level N-ary relation extraction with multiscale representation learning. In *Proceedings of the NAACL*. 3693–3704.

[20] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database J. Biol. Databases Curation* 2016 (2016), baw068 pages.

[21] Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. 2021. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1359–1370.

[22] Rui Li, Cheng Yang, Tingwei Li, and Sen Su. 2022. MiDTD: A simple and effective distillation framework for distantly supervised relation extraction. *ACM Transactions on Information Systems* 40, 4 (2022), 83:1–83:32.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692

[24] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the ICLR*. OpenReview.net.

[25] Erin MacDonald and Denilson Barbosa. 2020. Neural relation extraction on wikipedia tables for augmenting knowledge graphs. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. 2133–2136.

[26] Kohei Makino, Makoto Miwa, and Yutaka Sasaki. 2021. A neural edge-editing approach for document-level relation graph extraction. In *Findings of the ACL/IJCNLP*. 2653–2662.

[27] Mstislav Maslennikov and Tat-Seng Chua. 2010. Combining relations for information extraction from free text. *ACM Transactions on Information Systems* 28, 3 (2010), 14:1–14:35.

[28] Mintz, Mike, Steven, Jurafsky, and Dan. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 1003–1011.

[29] Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the ACL*. 1105–1116.

[30] Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the ACL*. 1546–1557.

[31] Alistair Plum, Tharindu Ranasinghe, Spencer Jones, Constantin Orasan, and Ruslan Mitkov. 2022. Biographical Semi-supervised relation extraction dataset. In *Proceedings of SIGIR*. 3121–3130.
[32] Meng Qu, Xiang Ren, Yu Zhang, and Jiawei Han. 2018. Weakly-supervised relation extraction by pattern-enhanced embedding learning. In *Proceedings of the 2018 World Wide Web Conference*. 1257–1266.
[33] Petar Ristoski, Anna Lisa Gentile, Alfredo Alba, Daniel Gruhl, and Steve Welch. 2020. Large-scale relation extraction from web documents and knowledge graphs with human-in-the-loop. *Journal of Web Semantics* 60 (2020), 100546.
[34] 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 EMNLP*. 1239–1250.
[35] Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of the ACL*. 4309–4316.
[36] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the ACL*.
[37] Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. arXiv:1904.05255. Retrieved from https://arxiv.org/abs/1904.05255
[38] Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2022. Learning implicit and explicit multi-task interactions for information extraction. *ACM Transactions on Information Systems* 41 (2022), 29.
[39] Kumutha Swampillai and Mark Stevenson. 2011. Extracting relations within and across sentences. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. 25–32.
[40] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.), Association for Computational Linguistics, 1672–1681. DOI: https://doi.org/10.18653/v1/2022.findings-acl.132
[41] 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 *Proceedings of the Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 1114, 2020*. 197–209.
[42] Siliang Tang, Jinjian Zhang, Ning Zhang, Fei Wu, Jun Xiao, and Yueting Zhuang. 2017. ENCORE: External neural constraints regularized distant supervision for relation extraction. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1113–1116.
[43] Hieu Minh Tran, Trung Minh Nguyen, and Thien Huu Nguyen. 2020. The dots have their values: Exploiting the node-edge connections in graph-based neural models for document-level relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 4561–4567.
[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*. 6000–6010.
[45] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of ICLR*.
[46] Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of NAACL-HLT*. 872–884.
[47] Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *Proceedings of the EMNLP*. 3711–3721.
[48] Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune Bert for DocRED with two-step process. arXiv:1909.11898. Retrieved from https://arxiv.org/abs/1909.11898
[49] Wang Xu, Kehai Chen, and Tiejun Zhao. 2021. Discriminative reasoning for document-level relation extraction. In *Findings of the ACL/IJCNLP*. 1653–1663.
[50] Xu Yan, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency path. *Computer Science* 42, 1 (2015), 56–61.
[51] Yulan Yan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang, and Mitsuru Ishizuka. 2009. Unsupervised relation extraction by mining wikipedia texts using information from the web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 1021–1029.
[52] Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the NAACL*. 289–299.
[53] 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 ACL*. 764–777.
[54] Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. In *Proceedings of the EMNLP*. 7170–7186.

[55] Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *Proceedings of NAACL*. 2810–2819.

[56] Bowen Yu, Mengge Xue, Zhenyu Zhang, Tingwen Liu, Yubin Wang, and Bin Wang. 2020. Learning to prune dependency trees with rethinking for neural relation extraction. In *Proceedings of COLING*. 3842–3852.

[57] Haoze Yu, Haisheng Li, Dianhui Mao, and Qiang Cai. 2020. A relationship extraction method for domain knowledge graph construction. *World Wide Web* 23, 2 (2020), 735–753.

[58] Changsen Yuan, Heyan Huang, and Chong Feng. 2021. Multi-graph cooperative learning towards distant supervised relation extraction. *ACM Transactions on Intelligent Systems and Technology* 12, 5 (2021), 52:1–52:21.

[59] Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of the ACL/IJCNLP*. 524–534.

[60] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the EMNLP*. 1630–1640.

[61] Sheng Zhang, Cliff Wong, Naoto Usuyama, Sarthak Jain, Tristan Naumann, and Hoifung Poon. 2021. Modular self-supervision for document-level relation extraction. In *Proceedings of the EMNLP*. 5291–5302.

[62] Yue Zhang, Hongliang Fei, and Ping Li. 2021. ReadsRE: Retrieval-augmented distantly supervised relation extraction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2257–2262.

[63] Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the EMNLP*. 2205–2215.

[64] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 35–45.

[65] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of AAAI*. 14612–14620.