

SIRE: Separate Intra- and Inter-sentential Reasoning for Document-level Relation Extraction

Shuang Zeng^{1,3}, Yuting Wu^{1,2} and Baobao Chang^{1*}

¹The MOE Key Laboratory of Computational Linguistics, Peking University, China

²Wangxuan Institute of Computer Technology, Peking University, China

³School of Software and Microelectronics, Peking University, China

{zengs, wytting, chbb}@pku.edu.cn

Abstract

Document-level relation extraction has attracted much attention in recent years. It is usually formulated as a classification problem that predicts relations for all entity pairs in the document. However, previous works indiscriminately represent intra- and inter-sentential relations in the same way, confounding the different patterns for predicting them. Besides, they create a document graph and use paths between entities on the graph as clues for logical reasoning. However, not all entity pairs can be connected with a path and have the correct logical reasoning paths in their graph. Thus many cases of logical reasoning cannot be covered. This paper proposes an effective architecture, SIRE, to represent intra- and inter-sentential relations in different ways. We design a new and straightforward form of logical reasoning module that can cover more logical reasoning chains. Experiments on the public datasets show SIRE outperforms the previous state-of-the-art methods. Further analysis shows that our predictions are reliable and explainable. Our code is available at <https://github.com/PKUnlp-icler/SIRE>.

1 Introduction

Relation Extraction (RE) is an important way of obtaining knowledge facts from natural language text. Many recent advancements (Sahu et al., 2019; Christopoulou et al., 2019; Yao et al., 2019b; Nan et al., 2020; Zeng et al., 2020; Wang et al., 2020) manage to tackle the document-level relation extraction (doc-level RE) that extracts semantic relations among entities across multiple sentences. Due to its strong correlation with real-world scenarios, doc-level RE has attracted much attention in the field of information extraction.

The doc-level RE task is usually formulated as a classification problem that predicts possible rela-

* Corresponding author.

ABBA Live [1] <i>ABBA Live</i> is an album of live recordings by <i>Swedish</i> pop group <i>ABBA</i> , released by <i>Polar Music</i> in <i>1986</i> [6] The tracks were mostly taken from <i>ABBA</i> 's concerts at <i>Wembley Arena</i> in <i>London</i> in <i>November 1979</i> [13] It was remastered ...
Head: <i>Polar Music</i> Tail: <i>Swedish</i> relation: <i>country of origin</i> evidence: [1]
Head: <i>Wembley Arena</i> Tail: <i>London</i> relation: <i>located in</i> evidence: [6]
IBM Research–Brazil [1] <i>IBM Research–Brazil</i> is one of twelve research laboratories comprising <i>IBM Research</i> , its first in <i>South America</i> [2] It was established in <i>June 2010</i> , with locations in <i>São Paulo</i> and <i>Rio de Janeiro</i> [5] In collaboration with <i>Brazil</i> 's government, it will help <i>IBM</i> [6] ... <i>IBM</i> has 4 priority areas in <i>Brazil</i> ...
<pre>graph LR SP[São Paulo] -- "country [1, 2, 5]" --> BR[Brazil] SP -- "continent [1, 2, 5]" --> SA[South America] BR -- "continent [1, 5]" --> SA BR -- "has part [1, 5]" --> SA</pre>

Figure 1: Two examples from DocRED (Yao et al., 2019b) for illustration of intra- and inter-sentential relations. Sentence numbers, entity mentions, and supporting evidence involved in these relation instances are colored. Other mentions are underlined for clarity.

tions for all entity pairs, using the information from the entire document. It has two different kinds of relations: intra-sentential relation and inter-sentential relation. We show examples of these two kinds of relations in Figure 1. When two entities have mentions co-occurred in the same sentence, they may express intra-sentential relations. Otherwise, they may express inter-sentential relations.

Previous methods do not explicitly distinguish these two kinds of relations in the design of the model and use the same method to represent them. However, from the perspective of linguistics, intra-sentential relations and inter-sentential relations are expressed in different patterns. For two intra-sentential entities, their relations are usually expressed from local patterns within their co-occurred

sentences. As shown in the first example in Figure 1, (*Polar Music, country of origin, Swedish*) and (*Wembley Arena, located in, London*) can be inferred based solely on the sentence they reside in, i.e., sentences 1 and 6 respectively. Unlike intra-sentential relations, inter-sentential relations tend to be expressed from the global interactions across multiple related sentences, also called supporting evidence. Moreover, cross-sentence relations usually require complex reasoning skills, e.g., logical reasoning. As shown in the second example in Figure 1, (*São Paulo, continent, South America*) can be inferred from the other two relation facts expressed in the document: (*São Paulo, country, Brazil*) and (*Brazil, continent, South America*). So the different patterns between intra- and inter-sentential relations show that it would be better for a model to treat intra- and inter-sentential relations differently. However, previous works usually use the information from the whole document to represent all relations, e.g., 13 sentences for predicting (*Polar Music, country of origin, Swedish*) in the first example in Figure 1. We argue that this will bring useless noises from unrelated sentences that misguide the learning of relational patterns.

Besides, previous methods (Christopoulou et al., 2019; Zeng et al., 2020) treat logical reasoning as a representation learning problem. They construct a document graph from the input document using entities as nodes. And the paths between two entities on their graphs, usually passing through other entities, could be regarded as clues for logical reasoning. However, since not all entity pairs can be connected with a path and have the correct logical reasoning paths available on the graph, many cases of logical reasoning cannot be covered. So their methods are somehow limited, and we should consider a new form of logical reasoning to better model and cover all possible reasoning chains.

In this paper, we propose a novel architecture called **Separate Intra- and inter-sentential REasoning (SIRE)** for doc-level RE. Unlike previous works in this task, we introduce two different methods to represent intra- and inter-sentential relations respectively. For an intra-sentential relation, we utilize a sentence-level encoder to represent it in every co-occurred sentence. Then we get the final representation by aggregating the relational representations from all co-occurred sentences. This will encourage intra-sentential entity pairs to focus on the local patterns in their co-occurred sen-

tences. For an inter-sentential relation, we utilize a document-level encoder and a mention-level graph proposed by Zeng et al. (2020) to capture the document information and interactions among entity mentions, document, and local context. Then, we apply an evidence selector to encourage inter-sentential entity pairs to selectively focus on the sentences that may signal their cross-sentence relations, i.e., finding supporting evidence. Finally, we develop a new form of logical reasoning module where one relation instance can be modeled by attentively fusing the representations of other relation instances in all possible logical chains. This form of logical reasoning could cover all possible cases of logical reasoning in the document.

Our contributions can be summarized as follows:

- We propose an effective architecture called SIRE that utilizes two different methods to represent intra-sentential and inter-sentential relations for doc-level RE.
- We come up with a new and straightforward form of logical reasoning module to cover all cases of logical reasoning chains.

We evaluate our SIRE on three public doc-level RE datasets. Experiments show SIRE outperforms the previous state-of-the-art models. Further analysis shows SIRE could produce more reliable and explainable predictions which further proves the significance of the separate encoding.

2 Separate Intra- and Inter-sentential Reasoning (SIRE) Model

SIRE mainly consists of three modules: intra- and inter-sentential relation representation module (Sec. 2.1), logical reasoning module (Sec. 2.2), classification module (Sec. 2.3), as is shown in Figure 2. Assume we have a document \mathcal{D} containing l sentences $\{\mathcal{S}_i\}_{i=1}^l$.

2.1 Intra- and Inter-sentential Relation Representation Module

As is discussed in Sec. 1, for two intra-sentential entities, their relations are usually determined by the local patterns from their co-occurred sentences, while for two inter-sentential entities, their relations are usually expressed across multiple related sentences that can be regarded as the supporting evidence for their relations. So in this module, we utilize two different methods to represent intra-sentential and inter-sentential relations separately.

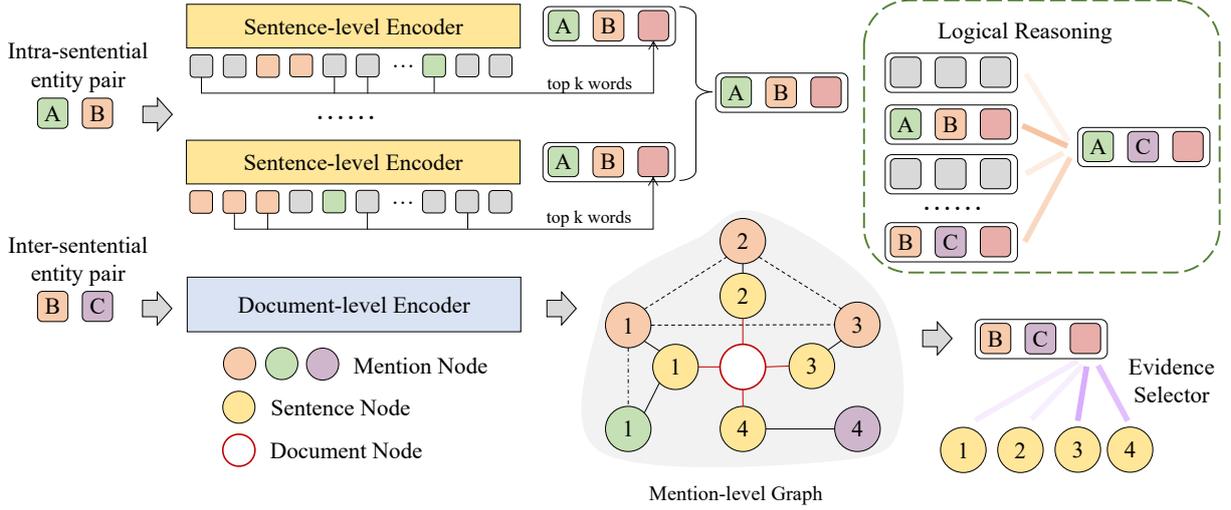


Figure 2: The architecture of SIRE. In the mention-level graph, the number in each circle is its sentence number. Mention nodes with the same color belong to the same entity. Different types of edges are in different styles of line. Our model uses different methods to represent intra- and inter-sentential relations and the self-attention mechanism to model the logical reasoning process. We use the logical reasoning chain: $e_A \rightarrow e_B \rightarrow e_C$ for illustration.

Our methods encourage intra-sentential entity pairs to focus on their co-occurred sentences as much as possible and encourage inter-sentential entity pairs to selectively focus on the sentences that may express their cross-sentence relations. We use three parts to represent the relation between two entities: head entity representation, tail entity representation and context representation.

2.1.1 Intra-sentential Relation Representation Module

Encoding. We use a sentence-level encoder to capture the context information for intra-sentential relations and produce contextualized word embedding for each word. Formally, we convert the i -th sentence \mathcal{S}_i containing n_i words $\{w_j^{\mathcal{S}_i}\}_{j=1}^{n_i}$ into a sequence of vectors $\{\mathbf{g}_j^{\mathcal{S}_i}\}_{j=1}^{n_i}$.

For each word w in \mathcal{S}_i , we first concatenate its word embedding with entity type embedding and co-reference embedding¹:

$$\mathbf{x} = [E_w(w); E_t(t); E_c(c)] \quad (1)$$

where $E_w(\cdot)$, $E_t(\cdot)$ and $E_c(\cdot)$ denote the word embedding layer, entity type embedding layer and

¹The existing doc-level RE datasets annotate which mentions belong to the same entity. So for each word in the document, it may belong to the i -th entity or non-entity in the document. We embed this co-reference information between entity mention (surface words) and entity (an abstract concept) into the initialized representation of a word.

co-reference embedding layer, respectively. t and c are named entity type and entity id.²

Then the vectorized word representations are fed into the sentence-level encoder to obtain the sentence-level context-sensitive representation for each word:

$$[\mathbf{g}_1^{\mathcal{S}_i}, \dots, \mathbf{g}_{n_i}^{\mathcal{S}_i}] = f_{enc}^{\mathcal{S}}([\mathbf{x}_1^{\mathcal{S}_i}, \dots, \mathbf{x}_{n_i}^{\mathcal{S}_i}]) \quad (2)$$

where the $f_{enc}^{\mathcal{S}}$ denotes sentence-level encoder, which can be any sequential encoder. We will also get the sentence representation $\mathbf{s}^{\mathcal{S}_i}$ for sentence \mathcal{S}_i from this encoder. For LSTM, $\mathbf{s}^{\mathcal{S}_i}$ is the hidden state of the last time step; for BERT, $\mathbf{s}^{\mathcal{S}_i}$ is the output representation of the special marker [CLS].

Representing. For i -th entity pair $(e_{i,h}, e_{i,t})$ which expresses intra-sentential relations, where $e_{i,h}$ is the head entity and $e_{i,t}$ is the tail entity, their mentions co-occur in C sentences $\mathcal{S}_{co-occur} = \{\mathcal{S}_{i_1}, \mathcal{S}_{i_2}, \dots, \mathcal{S}_{i_C}\}$ once or many times. In j -th co-occurred sentence \mathcal{S}_{i_j} , we use the entity mentions in \mathcal{S}_{i_j} to represent head and tail entity. And we define that the context representation of this relation instance in \mathcal{S}_{i_j} is the top K words correlated with the relations of these two mentions.

Specifically, head entity mention ranging from s -th to t -th word is represented as the average of the words it contains: $\mathbf{e}_{i,h}^{\mathcal{S}_{i_j}} = \frac{1}{t-s+1} \sum_{k=s}^t \mathbf{g}_k^{\mathcal{S}_{i_j}}$, so is the tail entity mention $\mathbf{e}_{i,t}^{\mathcal{S}_{i_j}}$. Then, we concatenate

²For those words not belonging to any entity, we introduce *None* entity type and id.

the representations of head and tail entity mentions and use it as a query to attend all words in \mathcal{S}_{i_j} and compute relatedness score for each word in \mathcal{S}_{i_j} :

$$s_{i,k} = \sigma((W_{intra} \cdot [\mathbf{e}_{i,h}^{S_{i_j}}; \mathbf{e}_{i,t}^{S_{i_j}}])^T \cdot \mathbf{g}_k^{S_{i_j}}) \quad (3)$$

$$\alpha_{i,k} = \text{Softmax}(s_{i,k}) \quad (4)$$

where $[\cdot; \cdot]$ is a concatenation operation. $W_{intra} \in \mathbb{R}^{d \times 2d}$ is a parameter matrix. σ is an activation function (e.g., ReLU).

Then, we average the representations of top K related words to represent the context information \mathbf{c}_i for intra-sentential entity pair $(e_{i,h}, e_{i,t})$ in \mathcal{S}_{i_j} . In order to make W_{intra} trainable during computing gradient, we also add an item which is the weighted average representation of all words:

$$\mathbf{c}_i^{S_{i_j}} = \beta \cdot \frac{1}{K} \sum_{k \in \text{top}K(\alpha_{i,*})} \mathbf{g}_k^{S_{i_j}} + (1-\beta) \cdot \sum_t^{n_{i_j}} \alpha_{i,t} \mathbf{g}_t^{S_{i_j}} \quad (5)$$

where β is a hyperparameter and we use 0.9 here to force model to focus on the top K words but still consider the subtle influence from other words.

Next, we concatenate the three parts obtained above to form the relational representation of intra-sentential entity pair $(e_{i,h}, e_{i,t})$ in \mathcal{S}_{i_j} and further average the representations in all co-occurred sentences $\mathcal{S}_{co-occur}$ to get our final relation representation \mathbf{r}_i for intra-sentential entity pair $(e_{i,h}, e_{i,t})$ ³:

$$\mathbf{r}_i = \frac{1}{C} \sum_{S_{i_j} \in \mathcal{S}_{co-occur}} [\mathbf{e}_{i,h}^{S_{i_j}}; \mathbf{e}_{i,t}^{S_{i_j}}; \mathbf{c}_i^{S_{i_j}}] \quad (6)$$

This way, we could force the intra-sentential entity pairs to focus on the semantic information from their co-occurred sentences and ignore the noise information from other sentences.

2.1.2 Inter-sentential Relation Representation Module

Encoding. According to the nature of inter-sentential relation, we use a document-level encoder to capture the global interactions for inter-sentential relations and produce contextualized word embedding for each word. Formally, we convert a document \mathcal{D} containing m words $\{w_i^{\mathcal{D}}\}_{i=1}^m$ into a sequence of vectors $\{\mathbf{g}_j^{\mathcal{D}}\}_{j=1}^m$.

³If a head entity mentioned N times in a sentence, we will get N intra-sentential relational representations for each of the other tail entities in this sentence.

Same as the embedding for intra-sentential relations, we use Equation 1 to embed each word in the document. Then the vectorized word representations are fed into the document-level encoder to obtain document-level context-sensitive representation for each word:

$$[\mathbf{g}_1^{\mathcal{D}}, \dots, \mathbf{g}_m^{\mathcal{D}}] = f_{enc}^{\mathcal{D}}([\mathbf{x}_1^{\mathcal{D}}, \dots, \mathbf{x}_m^{\mathcal{D}}]) \quad (7)$$

where $f_{enc}^{\mathcal{D}}$ denotes the document-level encoder. And we will also get the document representation $\mathbf{d}^{\mathcal{D}}$ from this encoder.

To further enhance the document interactions, we utilize the mention-level graph (MG) proposed by Zeng et al. (2020). MG in Zeng et al. (2020) contains two different nodes: mention node and document node. Each mention node denotes one particular mention of an entity. Furthermore, MG also has one document node that aims to model the document information. We argue that this graph only contains nodes concerning prediction, i.e., the mentions of the entities and document information. However, it does not contain the local context information, which is crucial for the interaction among entity mentions and the document. So we introduce a new type of node: sentence node and its corresponding new edges to infuse the local context information into MG.

So there are four types of edges⁴ in MG:

Intra-Entity Edge: Mentions referring to the same entity are fully connected. This models the interactions among mentions of the same entity.

Inter-Entity Edge: Mentions co-occurring in the same sentence are fully connected. This models the interactions among different entities via co-occurrences of their mentions.

Sentence-Mention Edge: Each sentence node connects with all entity mentions it contains. This models the interactions between mentions and their local context information.

Sentence-Document Edge: All sentence nodes are connected to the document node. This models the interactions between local context information and document information, acting as a bridge between mentions and document.

Next, we apply Relational Graph Convolutional Network (R-GCN, Schlichtkrull et al., 2017) on MG to aggregate the features from neighbors for each node. Given node u at the l -th layer, the graph

⁴Note that we remove the mention-document edges of original MG in (Zeng et al., 2020) and substitute them by introducing mention-sentence and sentence-document edges.

convolutional operation can be defined as:

$$\mathbf{h}_u^{(l+1)} = \text{ReLU} \left(\sum_{t \in \mathcal{T}} \sum_{v \in \mathcal{N}_u^t \cup \{u\}} \frac{1}{c_{u,t}} W_t^{(l)} \mathbf{h}_v^{(l)} \right) \quad (8)$$

where \mathcal{T} is a set of different types of edges, $W_t^{(l)} \in \mathbb{R}^{d \times d}$ is a trainable parameter matrix. \mathcal{N}_u^t denotes a set of neighbors for node u connected with t -th type edge. $c_{u,t} = |\mathcal{N}_u^t|$ is a normalization constant.

We then aggregate the outputs of all R-GCN layers to form the final representation of node u :

$$\mathbf{m}_u = \text{ReLU}(W_u \cdot [\mathbf{h}_u^{(0)}; \mathbf{h}_u^{(1)}; \dots; \mathbf{h}_u^{(N)}]) \quad (9)$$

where $W_u \in \mathbb{R}^{d \times Nd}$ is a trainable parameter matrix. $h_u^{(0)}$ is the initial representation of node u . For a mention ranging from the s -th word to the t -th word in the document, $\mathbf{h}_u^{(0)} = \frac{1}{t-s+1} \sum_{j=s}^t \mathbf{g}_j^{\mathcal{D}}$; for i -th sentence node, it is initialized with \mathbf{s}^i from sentence-level encoder; for the document node, it is initialized with $\mathbf{d}^{\mathcal{D}}$ from document-level encoder.

Representing. We argue that inter-sentential relations can be inferred from the following information sources: 1) the head and tail entities themselves; 2) the related sentences that signal their cross-sentence relations, namely supporting evidences; 3) reasoning information such as logical reasoning, co-reference reasoning, world knowledge, etc. We here only consider the first two information and leave the last in Sec. 2.2.

Different from intra-sentential relations, inter-sentential relations tend to be expressed from the global interactions. So for the i -th entity pair $(e_{i,h}, e_{i,t})$ which expresses inter-sentential relation, the head entity representation $\mathbf{e}_{i,h}$ and the tail entity representation $\mathbf{e}_{i,t}$ are defined as the average of their entity mentions from MG:

$$\mathbf{e}_i = \frac{1}{N} \sum_{j \in M(e_i)} \mathbf{m}_j \quad (10)$$

where the $M(e_i)$ is the mention set of e_i .

And we apply an evidence selector with attention mechanism (Bahdanau et al., 2015) to encourage the inter-sentential entity pair to selectively focus on the sentences that express their cross-sentence relations. This process could be regarded as finding supporting evidence for their relations. So the context representation \mathbf{c}_i for inter-sentential entity pair $(e_{i,h}, e_{i,t})$ is the weighted average of the sentence representations from MG:

$$P(\mathcal{S}_k | e_{i,h}, e_{i,t}) = \sigma(W_k \cdot [\mathbf{e}_{i,h}; \mathbf{e}_{i,t}; \mathbf{m}_{\mathcal{S}_k}]) \quad (11)$$

$$\alpha_{i,k} = \frac{P(\mathcal{S}_k | e_{i,h}, e_{i,t})}{\sum_l P(\mathcal{S}_l | e_{i,h}, e_{i,t})} \quad (12)$$

$$\mathbf{c}_i = \sum_k \alpha_{i,k} \cdot \mathbf{m}_{\mathcal{S}_k} \quad (13)$$

where $W_k \in \mathbb{R}^{1 \times 2d}$ is a trainable parameter matrix. σ is a *sigmoid* function.

Next, the final relation representation for inter-sentential entity pair $(e_{i,h}, e_{i,t})$ should be:

$$\mathbf{r}_i = [\mathbf{e}_{i,h}; \mathbf{e}_{i,t}; \mathbf{c}_i] \quad (14)$$

2.2 Logical Reasoning Module

In this module, we focus on logical reasoning modeling. As mentioned in Sec. 1, previous works usually use the paths between each entity pair as the clues for logical reasoning. Furthermore, they concatenate the path representations with entity pair representations to predict relations. However, since not all entity pairs are connected with a path and have the correct logical reasoning paths in their graph, many cases of logical reasoning cannot be covered. So their methods are somehow limited.

In this paper, we utilize self-attention mechanism (Vaswani et al., 2017) to model logical reasoning. Specifically, we can get the relational representations for all entity pairs from the above sections. For i -th entity pair (e_h, e_t) , we can assume there is a two-hop logical reasoning chains: $e_h \rightarrow e_k \rightarrow e_t$ in the document, where e_k can be any other entities in the document except e_h and e_t . So (e_h, e_t) can attend to all the relational representations of other entity pairs including (e_h, e_k) and (e_k, e_t) , termed as \mathcal{R}_{att} . Finally, the weighted sum of \mathcal{R}_{att} can be treated as a new relational representation for (e_h, e_t) , which considers all possible two-hop logical reasoning chains in the document.⁵

$$\mathbf{r}_i^{new} = \sum_{\mathbf{r}_k \in \mathcal{R}_{att} \cup \{\mathbf{r}_i\}} \gamma_k \cdot \mathbf{r}_k \quad (15)$$

$$\gamma_k = \text{Softmax}((W_{att} \cdot \mathbf{r}_i)^T \cdot \mathbf{r}_k) \quad (16)$$

where $W_{att} \in \mathbb{R}^{3d \times 3d}$ is a parameter matrix.

In this way, the path in the previous works could be converted into the individual attention on every entity pair in the logical reasoning chains. We argue that this form of logical reasoning is simpler

⁵This can be scaled to multi-hop logical reasoning by increasing the self-attention layers. We only consider two-hop logical reasoning in this paper following Zeng et al. (2020).

and more scalable because it will consider all possible logical reasoning chains without connectivity constraints in the graph structure.

2.3 Classification Module

We formulate the doc-level RE task as a multi-label classification task:

$$P(r|e_{i,h}, e_{i,t}) = \text{sigmoid}(W_1\sigma(W_2\mathbf{r}_i + b_1) + b_2) \quad (17)$$

where W_1, W_2, b_1, b_2 are trainable parameters, σ is an activation function (e.g., ReLU). We use binary cross entropy as objective to train our SIRE:

$$\begin{aligned} \mathcal{L}_{rel} = & - \sum_{D \in \mathcal{C}} \sum_{h \neq t} \sum_{r_i \in \mathcal{R}} \mathbb{I}(r_i = 1) \log P(r_i|e_{i,h}, e_{i,t}) \\ & + \mathbb{I}(r_i = 0) \log(1 - P(r_i|e_{i,h}, e_{i,t})) \end{aligned} \quad (18)$$

where \mathcal{C} denotes the whole corpus, \mathcal{R} denotes relation type set and $\mathbb{I}(\cdot)$ refers to indicator function.

3 Experiments

3.1 Dataset

We evaluate our proposed model on three document-level RE datasets:

DocRED: The largest human-annotated document-level relation extraction dataset was proposed by Yao et al. (2019b). It is constructed from Wikipedia and Wikidata and contains 96 types of relations, 132,275 entities, and 56,354 relational facts in total. Documents in DocRED have about 8 sentences on average. More than 40.7% relation facts can only be extracted from multiple sentences. 61.1% relation instances require various reasoning skills such as logical reasoning. 93.4% intra-sentential relations can be inferred based solely on their co-occurred sentences. We show two examples from DocRED in Figure 1. We follow the standard split of the dataset, 3,053 documents for training, 1,000 for development, and 1,000 for testing.

CDR (BioCreative V): The Chemical-Disease Reactions dataset was created by Li et al. (2016) manually. It contains one type of relation: *Chemical-Induced-Disease* between chemical and disease entities. We follow the standard split of the dataset, 500 documents for training, 500 for development, and 500 for testing.

GDA (DisGeNet): The Gene-Disease-Associations dataset was introduced by Wu et al. (2019). It contains one type of relation: *Gene-Induced-Disease* between gene and disease entities. We use standard

split of the dataset, 23,353 documents for training, 5,839 for development, and 1,000 for testing.

3.2 Experimental Settings

In our SIRE implementation, we use 3 layers of GCN, use ReLU as our activation function, and set the dropout rate to 0.3, learning rate to 0.001. We train SIRE using AdamW (Loshchilov and Hutter, 2019) as optimizer with weight decay 0.0001 and implement SIRE under PyTorch (Paszke et al., 2017) and DGL (Wang et al., 2019b) frameworks.

We implement two settings for our SIRE. **SIRE-GloVe** uses GloVe (100d, Pennington et al., 2014) and BiLSTM (512d, Schuster and Paliwal, 1997) as word embedding and encoder, respectively. **SIRE-BERT** use BERT-base (Devlin et al., 2019) as encoder on DocRED, cased BioBERT-Base v1.1 as the encoder on CDR/GDA, and the learning rate for BERT parameters is set to $1e^{-5}$ and learning rate for other parameters remains $1e^{-3}$. Detailed hyperparameter settings are in Appendix.

3.3 Baselines and Evaluation Metrics

We use the following models as our baselines:

Yao et al. (2019b) propose the **BiLSTM** (Schuster and Paliwal, 1997) as the encoder on DocRED and use the output from the encoder to represent all entity pairs to predict relations.

Wang et al. (2019a) propose **BERT** to replace the BiLSTM as the encoder on DocRED. Moreover, they also propose **BERT-Two-Step**, which first predicts whether two entities have a relation and then predicts the specific target relation.

Tang et al. (2020) propose the hierarchical inference networks **HIN-GloVe** and **HIN-BERT**, which make full use of multi-granularity inference information including entity level, sentence level, and document level to infer relations.

Similar to Wang et al. (2019a), Ye et al. (2020) propose a language representation model called **CorefBERT** as encoder on DocRED that can capture the coreferential relations in context.

Nan et al. (2020) propose the **LSR-GloVe** and **LSR-BERT** to dynamically induce the latent dependency tree structure to better model the document interactions for prediction.

Wang et al. (2020) propose a global-to-local network **GLRE**, which encodes the document information in terms of entity global and local representations as well as context relation representations.

Zeng et al. (2020) propose the graph aggregation-and-inference networks **GAIN-GloVe** and **GAIN-**

Model	Dev				Test	
	Ign F1	F1	Intra-F1	Inter-F1	Ign F1	F1
BiLSTM (Yao et al., 2019b)	48.87	50.94	57.05	43.49	48.78	51.06
HIN-GloVe (Tang et al., 2020)	51.06	52.95	-	-	51.15	53.30
LSR-GloVe (Nan et al., 2020)	48.82	55.17	60.83	48.35	52.15	54.18
GAIN-GloVe (Zeng et al., 2020)	53.05	55.29	61.67	48.77	52.66	55.08
SIRE-GloVe	54.10	55.91	62.94	48.97	54.04	55.96
-LR Module	53.73	55.58	62.77	47.87	53.75	55.55
-context	52.57	54.41	61.66	46.92	52.33	54.15
-inter4intra	52.23	54.26	60.81	48.36	51.77	53.30
BERT (Wang et al., 2019a)	-	54.16	61.61	47.15	-	53.20
BERT-Two-Step (Wang et al., 2019a)	-	54.42	61.80	47.28	-	53.92
HIN-BERT (Tang et al., 2020)	54.29	56.31	-	-	53.70	55.60
CorefBERT (Ye et al., 2020)	55.32	57.51	-	-	54.54	56.96
GLRE-BERT (Wang et al., 2020)	-	-	-	-	55.40	57.40
LSR-BERT (Nan et al., 2020)	52.43	59.00	65.26	52.05	56.97	59.05
GAIN-BERT (Zeng et al., 2020)	59.14	61.22	67.10	53.90	59.00	61.24
SIRE-BERT	59.82	61.60	68.07	54.01	60.18	62.05

Table 1: Performance on DocRED. Models above the double line do not use pre-trained model. *LR Module* is the logical reasoning module. *context* denotes context representations in Eq. 6 and Eq. 14. *inter4intra* denotes using the inter-sentential module also for intra-sentential entity pairs.

Model	CDR	GDA
BRAN (Verga et al., 2018)	62.1	-
EoG (Wang et al., 2020)	63.6	81.5
LSR (Nan et al., 2020)	64.8	82.2
GLRE-BioBERT (Wang et al., 2020)	68.5	-
SIRE-BioBERT	70.8	84.7

Table 2: Performance on CDR and GDA.

BERT which utilize two levels of graph structures: mention-level graph and entity-level graph to capture document interactions and conduct path logical reasoning mechanism, respectively.

Verga et al. (2018) propose a self-attention encoder **BRAN** to consider interactions across mentions and relations across sentence boundaries.

Following the previous works (Yao et al., 2019b; Zeng et al., 2020), we use the F1 and Ign F1 as the evaluation metrics to evaluate the overall performance of a model. The Ign F1 metric calculates F1 excluding the common relation facts in the training and dev/test sets. We also use the intra-F1 and inter-F1 metrics to evaluate a model’s performance on intra-sentential relations and inter-sentential relations on the dev set.

3.4 Results

The performances of SIRE and baseline models on the DocRED dataset are shown in Table 1. Among the model not using BERT encoding, SIRE outperforms the previous state-of-the-art model by 0.88/1.38 F1/Ign F1 on the test set. Among the model using BERT encoding, SIRE outperforms the previous state-of-the-art models by 1.18/0.81

F1/Ign F1 on the test set. The improvement on Ign F1 is larger than that on F1. This shows SIRE has a stronger generalization ability on the unseen relation instances. On intra-F1 and inter-F1, we can observe that SIRE is better than the previous models that indiscriminately represent the intra- and inter-sentential relations in the same way. This demonstrates that representing intra- and inter-sentential relations in different methods is better than representing them in the same way. The improvement on intra-F1 is greater than the improvement on inter-F1. This shows that SIRE mainly improves the performance of intra-sentential relations. The performances of SIRE and baseline models on the CDR/GDA dataset are shown in Table 2, which are consistent with the improvement on DocRED.

3.5 Ablation Study

To further analyze SIRE, we also conduct ablation studies to illustrate the effectiveness of different modules in SIRE. We show the results in Table 1.

1) **the importance of the logical reasoning module**: When we discard the logical reasoning module, the performance of SIRE-GloVe decreases by 0.41 F1 on the DocRED test set. This shows the effectiveness of our logical reasoning module, which can better model the reasoning information in the document. Moreover, it drops significantly on inter-F1 and drops fewer points on intra-F1. This shows our logical reasoning module mainly improves the performance of the inter-sentential relations that usually require reasoning skills.

2) **Ablation on context representations in Eq. 6**

Type	Examples
Intra-sentential relation instances	" <i>Your Disco Needs You</i> " is a song performed by Australian recording artist and songwriter <i>Kylie Minogue</i> , taken from her seventh studio album <i>Light Years</i> (2000). Relation: performer
	<i>Lark Force</i> was an <i>Australian Army</i> formation established in <i>March 1941</i> during <i>World War II</i> for service in <i>New Britain</i> and <i>New Ireland</i> . Relation: inception
	<i>Lake Hiawatha</i> is one of the few lakes through which <i>Minnehaha Creek</i> flows, and the last one before it reaches <i>Minnehaha Falls</i> and then the <i>Mississippi River</i> . Relation: mouth of the watercourse
Inter-sentential relation instances	[1] (0.87) <i>IBM Research–Brazil</i> is one of twelve research laboratories comprising <i>IBM Research</i> , its first in <i>South America</i> . [2] (0.66) It was established in <i>June 2010</i> , with locations in <i>São Paulo</i> and <i>Rio de Janeiro</i> . [3] (0.01) Research focuses on Industrial Technology and Science, ... [4] (0.04) The new lab, <i>IBM's</i> ninth ... [5] (0.38) In collaboration with <i>Brazil's</i> government, it will help <i>IBM</i> to develop ... Relation: continent Logical reasoning attention weight: (São Paulo, Brazil) 0.32 (São Paulo, June 2010) 0.03 ... (Brazil, South America) 0.45 (June 2010, South America) 0.02 ...

Figure 3: Cases for illustrating the reliable and explainable predictions of our SIRE. Head entities, tail entities, and sentence numbers along with the scores from evidence selector are colored in blue, red, green, respectively. In intra-sentential relations, words with pink background color are the top 4 words from Equation 5.

and Eq. 14: When we remove the context representations in intra- and inter-sentential relational representations, the performance of SIRE-GloVe on the DocRED test set drops by 1.81 F1. This shows context information (top K words for intra, evidence sentences for inter) is important for both intra- and inter-sentential relation representation.

3) **Using the inter-sentential module also for intra-sentential entity pairs:** In this experiment, we do not distinguish these two types of relations, using the encoding method for inter-sentential to encode all entity pairs, and remain the logical reasoning module unchanged. The performance of SIRE-GloVe drops by 2.66/2.13 F1/intra-F1 on the DocRED test set. This confirms the motivation that we cannot use global information to learn the local patterns for intra-sentential relations.

3.6 Reasoning Performance

Furthermore, we evaluate the reasoning ability of our model on the development set in Table 3. We use infer-F1 as the metric that considers only two-hop positive relation instances in the dev set. So it will naturally exclude many cases that do not belong to the two-hop logical reasoning process to strengthen the evaluation of reasoning performance. As Table 3 shows, SIRE is superior to previous models in handling the two-hop logical reasoning process. Moreover, after removing the logical reasoning module, our SIRE drops signif-

Model	Infer-F1	P	R
BiLSTM	38.73	31.60	50.01
GAIN-GloVe	40.82	32.76	54.14
SIRE-GloVe	42.72	34.83	55.22
- LR Module	39.18	31.97	50.59

Table 3: Infer-F1 results on dev set of DocRED. P: Precision, R: Recall.

icantly on infer-F1. This shows that our logical reasoning module plays a crucial role in modeling the logical reasoning process.

3.7 Case Study

Figure 3 shows the prediction cases of our SIRE. In intra-sentential relations, the top 4 words related to the relations of three entity pairs conform with our intuition. Our model correctly find the words by using Eq.5 that trigger the relations of these entity pairs. In inter-sentential relations, the supporting evidence that the model finds, i.e., sentences 1 and 2, indeed expresses the relations between São Paul and South America. We also conduct logical reasoning in terms of the logical reasoning chains: São Paul → other-entity → South America. Our SIRE could focus on the correct logical reasoning chains: São Paul → Brazil → South America. These cases show the predictions of SIRE are explainable.

4 Related Work

Document-level relation extraction. Many recent efforts (Quirk and Poon, 2017; Peng et al., 2017; Gupta et al., 2019; Song et al., 2018; Jia et al., 2019; Yao et al., 2019b; Wang et al., 2019a; Tang et al., 2020; Nan et al., 2020; Zeng et al., 2020; Wang et al., 2020; Dai et al., 2020) manage to tackle the document-level relation extraction. Most of them use graph-based models, such as Graph Convolutional Networks (GCNs, Kipf and Welling, 2017; Schlichtkrull et al., 2017) that has been used in many natural language processing tasks (Marcheggiani and Titov, 2017; Yao et al., 2019a; Liu et al., 2020). They construct a graph structure from the input document. This graph uses the word, mentions or entities as nodes and uses heuristic rules and semantic dependencies as edges. They use this graph to model document information and interactions and to predict possible relations for all entity pairs. Nan et al. (2020) proposed a latent structure induction to induce the dependency tree in the document dynamically. Zeng et al. (2020) proposed a double graph-based graph aggregation-and-inference network that constructs two graphs: mention-level graph and entity-level graph. They use the former to capture the document information and interactions among entity mentions and document and use the latter to conduct path-based logical reasoning. However, these works do not explicitly distinguish the intra- and inter-sentential relation instances in the design of the model and use the same way to encode them. So the most significant difference between our model and previous models is that we treat intra-sentential and inter-sentential relations differently to conform with the relational patterns for their prediction.

Reasoning in relation extraction. Reasoning problem has been extensively studied in the field of question answering (Dhingra et al., 2018). However, few works manage to tackle this problem in the document-level relation extraction task. Zeng et al. (2020) is the first to propose the explicit way of relational reasoning on doc-level RE, which mainly focuses on logical reasoning. They use the paths on their entity-level graph to provide clues for logical reasoning. However, since not all entity pairs are connected with a path and have the correct logical reasoning paths in their graph, their methods are somehow limited. In this work, we design a new form of logical reasoning to cover more cases of logical reasoning.

5 Conclusion

Intra- and inter-sentential relations are two types of relations in doc-level RE. We propose a novel architecture, SIRE, to represent these two relations in different ways separately in this work. We introduce a new form of logical reasoning module that models logical reasoning as a self-attention among representations of all entity pairs. Experiments show that our SIRE outperforms the previous state-of-the-art methods. The detailed analysis demonstrates that our predictions are explainable. We hope this work will have a positive effect on future research regarding new encoding schema, a more generalizable and explainable model.

Acknowledgments

The authors would like to thank the reviewers for their thoughtful and constructive comments. This paper is supported by the National Key R&D Program of China under Grand No.2018AAA0102003, the National Science Foundation of China under Grant No.61936012 and 61876004.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. [Neural machine translation by jointly learning to align and translate](#). In *3rd International Conference on Learning Representations, ICLR 2015*.
- Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. [Connecting the dots: Document-level neural relation extraction with edge-oriented graphs](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4925–4936.
- Damai Dai, Jing Ren, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2020. [Coarse-to-fine entity representations for document-level relation extraction](#). *Computing Research Repository*, arXiv:2012.02507.
- 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 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.
- Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2018. [Neural models for reasoning over multiple mentions using coreference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

- Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 42–48.
- Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, and Thomas A. Runkler. 2019. [Neural relation extraction within and across sentence boundaries](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 6513–6520.
- Robin Jia, Cliff Wong, and Hoifung Poon. 2019. [Document-level n-ary relation extraction with multiscale representation learning](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3693–3704.
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). In *5th International Conference on Learning Representations, ICLR 2017*.
- Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wieggers, and Zhiyong Lu. 2016. [Biocreative V CDR task corpus: a resource for chemical disease relation extraction](#). *Database*, 2016.
- Sennan Liu, Shuang Zeng, and Sujian Li. 2020. [Evaluating text coherence at sentence and paragraph levels](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1695–1703. European Language Resources Association.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations, ICLR 2019*.
- Diego Marcheggiani and Ivan Titov. 2017. [Encoding sentences with graph convolutional networks for semantic role labeling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515.
- Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. [Reasoning with latent structure refinement for document-level relation extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. [Automatic differentiation in pytorch](#). In *NIPS-W*.
- Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. [Cross-sentence n-ary relation extraction with graph LSTMs](#). *Transactions of the Association for Computational Linguistics*, 5:101–115.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Chris Quirk and Hoifung Poon. 2017. [Distant supervision for relation extraction beyond the sentence boundary](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182.
- 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 57th Annual Meeting of the Association for Computational Linguistics*, pages 4309–4316.
- Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. [Modeling relational data with graph convolutional networks](#). *Computing Research Repository*, arXiv:1703.06103.
- M. Schuster and K. K. Paliwal. 1997. [Bidirectional recurrent neural networks](#). *IEEE Transactions on Signal Processing*, 45(11):2673–2681.
- Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. [N-ary relation extraction using graph-state LSTM](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2226–2235.
- 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, Singapore, May 11-14, 2020, Proceedings, Part I*, volume 12084 of *Lecture Notes in Computer Science*, pages 197–209.
- 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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. [Simultaneously self-attending to all mentions for full-abstract biological relation extraction](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884.
- Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. [Global-to-local neural networks for document-level relation extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods*

in *Natural Language Processing (EMNLP)*, pages 3711–3721.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019a. [Fine-tune bert for docred with two-step process](#). *Computing Research Repository*, arXiv:1909.11898.

Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, Ziyue Huang, Qipeng Guo, Hao Zhang, Haibin Lin, Junbo Zhao, Jinyang Li, Alexander J Smola, and Zheng Zhang. 2019b. [Deep graph library: Towards efficient and scalable deep learning on graphs](#). *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Ye Wu, Ruibang Luo, Henry C. M. Leung, Hing-Fung Ting, and Tak Wah Lam. 2019. [RENET: A deep learning approach for extracting gene-disease associations from literature](#). In *Research in Computational Molecular Biology - 23rd Annual International Conference, RECOMB 2019, Washington, DC, USA, May 5-8, 2019, Proceedings*, volume 11467, pages 272–284.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019a. [Graph convolutional networks for text classification](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7370–7377.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019b. [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.

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 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. [Double graph based reasoning for document-level relation extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640.

A Hyperparameter settings

We use the development set to manually tune the optimal hyperparameters for SIRE, based on the Ign F1 score. Experiments are run on NVIDIA-RTX-3090-24GB GPU. Hyperparameter settings for SIRE-GloVe, SIRE-BERT on DocRED are

listed in Table 4, 5, respectively. The values of hyperparameters we finally adopted are in bold. Note that we do not tune all the hyperparameters.

Hyperparameter	Value
Batch Size	16, 32
Learning Rate	0.001
Activation Function	ReLU , Tanh
Positive v.s. Negative Ratio	1, 0.5, 0.25
Word Embedding Size	200
Entity Type Embedding Size	20
Coreference Embedding Size	20
Encoder Hidden Size	256, 512
Dropout	0.3, 0.5 , 0.7
Layers of GCN	1, 2, 3
GCN Hidden Size	1024
Weight Decay	0.0001
β	0.9
Numbers of Parameters	95M
Training Time	18 hours
Hyperparameter Search Trials	20

Table 4: Settings for SIRE-GloVe.

Hyperparameter	Value
Batch Size	16, 32
Learning Rate	0.001
Activation Function	ReLU , Tanh
Positive v.s. Negative Ratio	1, 0.5, 0.25
Entity Type Embedding Size	128
Coreference Embedding Size	128
Dropout	0.3 , 0.5, 0.7
Layers of GCN	1, 2, 3
GCN Hidden Size	1024
Weight Decay	0.0001
β	0.9
Numbers of Parameters	307M
Training Time	24 hours
Hyperparameter Search Trials	30

Table 5: Settings for SIRE-BERT.