Global-to-Local Neural Networks for Document-Level Relation Extraction

Difeng Wang[†] **Wei Hu**[†, ‡, *] **Ermei Cao**[†] **Weijian Sun**[§]
[†] State Key Laboratory for Novel Software Technology, Nanjing University, China
[‡] National Institute of Healthcare Data Science, Nanjing University, China
[§] Huawei Technologies Co., Ltd.
{dfwang,emcao}.nju@gmail.com, whu@nju.edu.cn, sunweijian@huawei.com

Abstract

Relation extraction (RE) aims to identify the semantic relations between named entities in text. Recent years have witnessed it raised to the document level, which requires complex reasoning with entities and mentions throughout an entire document. In this paper, we propose a novel model to document-level RE, by encoding the document information in terms of entity global and local representations as well as context relation representations. Entity global representations model the semantic information of all entities in the document, entity local representations aggregate the contextual information of multiple mentions of specific entities, and context relation representations encode the topic information of other relations. Experimental results demonstrate that our model achieves superior performance on two public datasets for document-level RE. It is particularly effective in extracting relations between entities of long distance and having multiple mentions.

1 Introduction

Relation extraction (RE) aims to identify the semantic relations between named entities in text. While previous work (Zeng et al., 2014; Zhang et al., 2015, 2018) focuses on extracting relations within a sentence, a.k.a. *sentence*-level RE, recent studies (Verga et al., 2018; Christopoulou et al., 2019; Sahu et al., 2019; Yao et al., 2019) have escalated it to the *document* level, since a large amount of relations between entities usually span across multiple sentences in the real world. According to an analysis on Wikipedia corpus (Yao et al., 2019), at least 40.7% of relations can only be extracted on the document level.

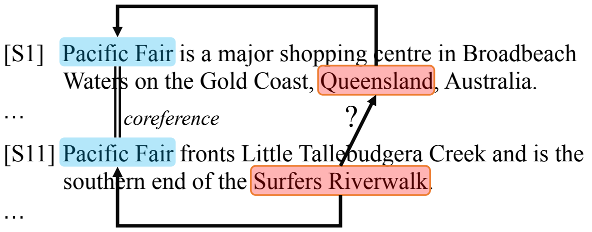

[S1] Pacific Fair is a major shopping centre in Broadbeach Waters on the Gold Coast, Queensland, Australia.
…
[S11] Pacific Fair fronts Little Tallebudgera Creek and is the southern end of the Surfers Riverwalk.
…

Figure 1: An example of document-level RE excerpted from the DocRED dataset (Yao et al., 2019). Arrows denote intra/inter-sentential relations.

Compared with sentence-level RE, document-level RE requires more complex reasoning, such as logical reasoning, coreference reasoning and common-sense reasoning. A document often contains many entities, and some entities have multiple mentions under the same phrase of alias. To identify the relations between entities appearing in different sentences, document-level RE models must be capable of modeling the complex interactions between multiple entities and synthesizing the context information of multiple mentions.

Figure 1 shows an example of document-level RE. Assume that one wants to extract the relation between *"Surfers Riverwalk"* in S11 and *"Queensland"* in S1. One has to find that *"Surfers Riverwalk"* contains *"Pacific Fair"* (from S11), and *"Pacific Fair"* (coreference) is located in *"Queensland"* (from S1). This chain of interactions helps infer the inter-sentential relation *"located in"* between *"Surfers Riverwalk"* and *"Queensland"*.

State-of-the-art. Early studies (Peng et al., 2017; Quirk and Poon, 2017) confined document-level RE to short text spans (e.g., within three sentences). Some other studies (Nguyen and Verspoor, 2018; Gupta et al., 2019) were restricted to handle two entity mentions in a document. We argue that they are incapable of dealing with the example in Figure 1, which needs to consider multiple mentions of entities integrally. To encode the semantic interactions of multiple entities in long distance, recent

[*] Corresponding author

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3711–3721,
November 16–20, 2020. ©2020 Association for Computational Linguistics

work defined document-level graphs and proposed graph-based neural network models. For example, Sahu et al. (2019); Gupta et al. (2019) interpreted words as nodes and constructed edges according to syntactic dependencies and sequential information. However, there is yet a big gap between word representations and relation prediction. Christopoulou et al. (2019) introduced the notion of document graphs with three types of nodes (mentions, entities and sentences), and proposed an edge-oriented graph neural model for RE. However, it indiscriminately integrated various information throughout the whole document, thus irrelevant information would be involved as noise and damages the prediction accuracy.

Our approach and contributions. To cope with the above limitations, we propose a novel graph-based neural network model for document-level RE. Our key idea is to make full use of document semantics and predict relations by learning the representations of involved entities from both coarse-grained and fine-grained perspectives as well as other context relations. Towards this goal, we address three challenges below:

First, *how to model the complex semantics of a document?* We use the pre-trained language model BERT (Devlin et al., 2019) to capture semantic features and common-sense knowledge, and build a heterogeneous graph with heuristic rules to model the complex interactions between all mentions, entities and sentences in the document.

Second, *how to learn entity representations effectively?* We design a global-to-local neural network to encode coarse-grained and fine-grained semantic information of entities. Specifically, we learn entity global representations by employing R-GCN (Schlichtkrull et al., 2018) on the created heterogeneous graph, and entity local representations by aggregating multiple mentions of specific entities with multi-head attention (Vaswani et al., 2017).

Third, *how to leverage the influence from other relations?* In addition to target relation representations, other relations imply the topic information of a document. We learn context relation representations with self-attention (Sorokin and Gurevych, 2017) to make final relation prediction.

In summary, our main contribution is twofold:

- We propose a novel model, called *GLRE*, for document-level RE. To predict relations between entities, GLRE synthesizes entity global representations, entity local representations and context relation representations integrally. For details, please see Section 3.
- We conducted extensive experiments on two public document-level RE datasets. Our results demonstrated the superiority of GLRE compared with many state-of-the-art competitors. Our detailed analysis further showed its advantage in extracting relations between entities of long distance and having multiple mentions. For details, please see Section 4.

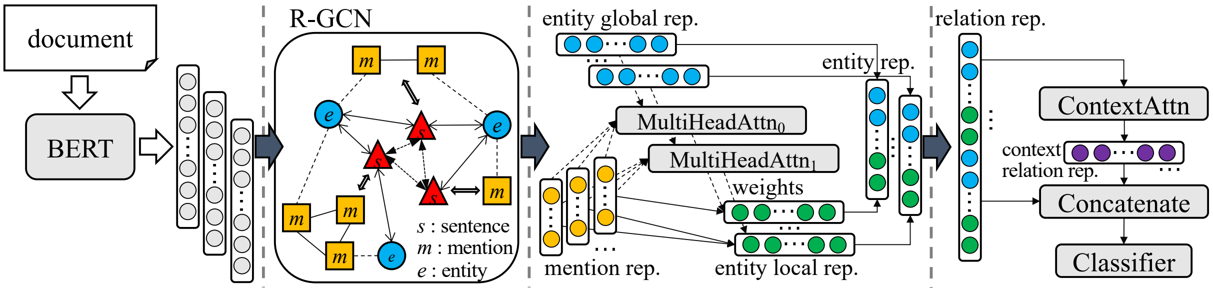
2 Related Work

RE has been intensively studied in a long history. In this section, we review closely-related work.

Sentence-level RE. Conventional work addressed sentence-level RE by using carefully-designed patterns (Soderland et al., 1995), features (Kambhatla, 2004) and kernels (Culotta and Sorensen, 2004). Recently, deep learning-based work has advanced the state-of-the-art without heavy feature engineering. Various neural networks have been exploited, e.g., CNN (Zeng et al., 2014), RNN (Zhang et al., 2015; Cai et al., 2016) and GNN (Zhang et al., 2018). Furthermore, to cope with the wrong labeling problem caused by distant supervision, Zeng et al. (2015) adopted Piecewise CNN (PCNN), Lin et al. (2016); Zhang et al. (2017) employed attention mechanisms, and Zhang et al. (2019); Qu et al. (2019) leveraged knowledge graphs as external resources. All these models are limited to extracting intra-sentential relations. They also ignore the interactions of entities outside a target entity pair.

Document-level RE. As documents often provide richer information than sentences, there has been an increasing interest in document-level RE. Gu et al. (2017); Nguyen and Verspoor (2018); Gupta et al. (2019); Wang et al. (2019) extended the sentence-level RE models to the document level. Ye et al. (2020) explicitly incorporated coreference information into language representation models (e.g., BERT). Zheng et al. (2018); Tang et al. (2020) proposed hierarchical networks to aggregate information from the word, sentence and document levels.

Quirk and Poon (2017) proposed the notion of document-level graphs, where nodes denote words and edges incorporate both syntactic dependencies and discourse relations. Following this, Peng et al. (2017) first splitted a document-level graph into two directed acyclic graphs (DAGs), then used a graph LSTM for each DAG to learn the contextual representation of each word, which was concate-



(a) Encoding layer (b) Global rep. layer (c) Local rep. layer (d) Classifier layer

Figure 2: Architecture of the proposed model.

nated and finally fed to the relation classifier. Differently, Song et al. (2018) kept the original graph structure and directly modeled the whole document-level graph using graph-state LSTM. These models only predict the relation of a single mention pair in a document at a time, and ignore multiple mentions of a target entity pair as well as other entities.

Several models predict the relation of a target entity pair by aggregating the scores of all mention pairs with multi-instance learning. Verga et al. (2018) proposed a Transformer-based model. Later, Sahu et al. (2019) switched Transformer to GCN. The two models only consider one target entity pair per document, and construct the document-level graphs relying on external syntactic analysis tools. Christopoulou et al. (2019) built a document graph with heterogeneous types of nodes and edges, and proposed an edge-oriented model to obtain global representations for relation classification. Our model differs in further learning entity local representations to reduce the influence of irrelevant information and considering other relations in the document to refine the prediction. Recently, Nan et al. (2020) defined a document graph as a latent variable and induced it based on the structured attention. Unlike our work, it improves the performance of document-level RE models by optimizing the structure of the document graph.

Besides, a few models (Levy et al., 2017; Qiu et al., 2018) borrowed the reading comprehension techniques to document-level RE. However, they require domain knowledge to design question templates, and may perform poorly in zero-answer and multi-answers scenarios (Liu et al., 2019), which are very common for RE.

3 Proposed Model

We model document-level RE as a *classification* problem. Given a document annotated with entities and their corresponding textual mentions, the objective of document-level RE is to identify the relations of all entity pairs in the document.

Figure 2 depicts the architecture of our model, named GLRE. It receives an entire document with annotations as input. First, in (a) *encoding layer*, it uses a pre-trained language model such as BERT (Devlin et al., 2019) to encode the document. Then, in (b) *global representation layer*, it constructs a global heterogeneous graph with different types of nodes and edges, and encodes the graph using a stacked R-GCN (Schlichtkrull et al., 2018) to capture entity global representations. Next, in (c) *local representation layer*, it aggregates multiple mentions of specific entities using multi-head attention (Vaswani et al., 2017) to obtain entity local representations. Finally, in (d) *classifier layer*, it combines the context relation representations obtained with self-attention (Sorokin and Gurevych, 2017) to make final relation prediction. Please see the rest of this section for technical details.

3.1 Encoding Layer

Let $\mathcal{D} = [w_1, w_2, \dots, w_k]$ be an input document, where w_j ($1 \leq j \leq k$) is the j^{th} word in it. We use BERT to encode $\mathcal{D}$ as follows:

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_k] = \text{BERT}([w_1, w_2, \dots, w_k]), \quad (1)$$

where $\mathbf{h}_j \in \mathbb{R}^{d_w}$ is a sequence of hidden states at the output of the last layer of BERT. Limited by the input length of BERT, we encode a long document sequentially in form of short paragraphs.

3.2 Global Representation Layer

Based on $\mathbf{H}$, we construct a *global heterogeneous graph*, with different types of nodes and edges to capture different dependencies (e.g., co-occurrence dependencies, coreference dependencies and order dependencies), inspired by Christopoulou et al. (2019). Specifically, there are three types of nodes:

- *Mention nodes,* which model different mentions of entities in $\mathcal{D}$. The representation of a mention node m_i is defined by averaging the representations of contained words. To distinguish node types, we concatenate a node type representation $\mathbf{t}_m \in \mathbb{R}^{d_t}$. Thus, the representation of m_i is $\mathbf{n}_{m_i} = [\text{avg}_{w_j \in m_i}(\mathbf{h}_j); \mathbf{t}_m]$, where $[\,;]$ is the concatenation operator.
- *Entity nodes,* which represent entities in $\mathcal{D}$. The representation of an entity node e_i is defined by averaging the representations of the mention nodes to which they refer, together with a node type representation $\mathbf{t}_e \in \mathbb{R}^{d_t}$. Therefore, the representation of e_i is $\mathbf{n}_{e_i} = [\text{avg}_{m_j \in e_i}(\mathbf{n}_{m_j}); \mathbf{t}_e]$.
- *Sentence nodes,* which encode sentences in $\mathcal{D}$. Similar to mention nodes, the representation of a sentence node s_i is formalized as $\mathbf{n}_{s_i} = [\text{avg}_{w_j \in s_i}(\mathbf{h}_j); \mathbf{t}_s]$, where $\mathbf{t}_s \in \mathbb{R}^{d_t}$.

Then, we define five types of edges to model the interactions between the nodes:

- *Mention-mention edges.* We add an edge for any two mention nodes in the same sentence.
- *Mention-entity edges.* We add an edge between a mention node and an entity node if the mention refers to the entity.
- *Mention-sentence edges.* We add an edge between a mention node and a sentence node if the mention appears in the sentence.
- *Entity-sentence edges.* We create an edge between an entity node and a sentence node if at least one mention of the entity appears in the sentence.
- *Sentence-sentence edges.* We connect all sentence nodes to model the non-sequential information (i.e., break the sentence order).

Note that there are no entity-entity edges, because they form the relations to be predicted.

Finally, we employ an L-layer stacked R-GCN (Schlichtkrull et al., 2018) to convolute the global heterogeneous graph. Different from GCN, R-GCN considers various types of edges and can better model multi-relational graphs. Specifically, its node forward-pass update for the $(l+1)^{\text{th}}$ layer is defined as follows:

$$\mathbf{n}_i^{l+1} = \sigma\Big(\sum_{x \in \mathcal{X}} \sum_{j \in \mathcal{N}_i^x} \frac{1}{|\mathcal{N}_i^x|} \mathbf{W}_x^l \mathbf{n}_j^l + \mathbf{W}_0^l \mathbf{n}_i^l\Big), \quad (2)$$

where $\sigma(\cdot)$ is the activation function. $\mathcal{N}_i^x$ denotes the set of neighbors of node i linked with edge x, and $\mathcal{X}$ denotes the set of edge types. $\mathbf{W}_x^l, \mathbf{W}_0^l \in \mathbb{R}^{d_n \times d_n}$ are trainable parameter matrices (d_n is the dimension of node representations).

We refer to the representations of entity nodes after graph convolution as *entity global representations*, which encode the semantic information of entities throughout the whole document. We denote an entity global representation by $\mathbf{e}_i^{\text{glo}}$.

3.3 Local Representation Layer

We learn *entity local representations* for specific entity pairs by aggregating the associated mention representations with multi-head attention (Vaswani et al., 2017). The "local" can be understood from two angles: (i) It aggregates the original mention information from the encoding layer. (ii) For different entity pairs, each entity would have multiple different local representations w.r.t. the counterpart entity. However, there is only one entity global representation.

Multi-head attention enables a RE model to jointly attend to the information of an entity composed of multiple mentions from different representation subspaces. Its calculation involves the sets of queries $\mathcal{Q}$ and key-value pairs $(\mathcal{K}, \mathcal{V})$:

$$\text{MHead}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = [\text{head}_1; \ldots; \text{head}_z]\mathbf{W}^{\text{out}}, \quad (3)$$

$$\text{head}_i = \text{softmax}\Big(\frac{\mathcal{Q}\mathbf{W}_i^{\mathcal{Q}}(\mathcal{K}\mathbf{W}_i^{\mathcal{K}})'}{\sqrt{d_v}}\Big)\mathcal{V}\mathbf{W}_i^{\mathcal{V}}, \quad (4)$$

where $\mathbf{W}^{\text{out}} \in \mathbb{R}^{d_n \times d_n}$ and $\mathbf{W}_i^{\mathcal{Q}}, \mathbf{W}_i^{\mathcal{K}}, \mathbf{W}_i^{\mathcal{V}} \in \mathbb{R}^{d_n \times d_v}$ are trainable parameter matrices. z is the number of heads satisfying that $z \times d_v = d_n$.

In this paper, $\mathcal{Q}$ is related to the entity global representations, $\mathcal{K}$ is related to the initial sentence node representations before graph convolution (i.e., the input features of sentence nodes in R-GCN), and $\mathcal{V}$ is related to the initial mention node representations. Specifically, given an entity pair (e_a, e_b), we define their local representations as follows:

$$\begin{aligned} \mathbf{e}_a^{\text{loc}} &= \text{LN}\big(\text{MHead}_0(\mathbf{e}_b^{\text{glo}}, \{\mathbf{n}_{s_i}\}_{s_i \in \mathcal{S}_a}, \{\mathbf{n}_{m_j}\}_{m_j \in \mathcal{M}_a})\big), \\ \mathbf{e}_b^{\text{loc}} &= \text{LN}\big(\text{MHead}_1(\mathbf{e}_a^{\text{glo}}, \{\mathbf{n}_{s_i}\}_{s_i \in \mathcal{S}_b}, \{\mathbf{n}_{m_j}\}_{m_j \in \mathcal{M}_b})\big), \end{aligned} \quad (5)$$

where $\text{LN}(\cdot)$ denotes layer normalization (Ba et al., 2016). $\mathcal{M}_a$ is the corresponding mention node set of e_a, and $\mathcal{S}_a$ is the corresponding sentence node set in which each mention node in $\mathcal{M}_a$ is located. $\mathcal{M}_b$ and $\mathcal{S}_b$ are similarly defined for e_b. Note that MHead_0 and MHead_1 learn independent model parameters for entity local representations.

Intuitively, if a sentence contains two mentions m_a, m_b corresponding to e_a, e_b, respectively, then the mention node representations $\mathbf{n}_{m_a}, \mathbf{n}_{m_b}$ should contribute more to predicting the relation of (e_a, e_b) and the attention weights should be greater in getting $\mathbf{e}_a^{\text{loc}}, \mathbf{e}_b^{\text{loc}}$. More generally, a higher semantic similarity between the node representation of a sentence containing m_a and $\mathbf{e}_b^{\text{glo}}$ indicates that this sentence and m_b are more semantically related, and $\mathbf{n}_{m_a}$ should get a higher attention weight to $\mathbf{e}_a^{\text{loc}}$.

3.4 Classifier Layer

To classify the target relation r for an entity pair (e_a, e_b), we firstly concatenate entity global representations, entity local representations and relative distance representations to generate entity final representations:

$$\begin{aligned} \hat{\mathbf{e}}_a &= [\mathbf{e}_a^{\text{glo}}; \mathbf{e}_a^{\text{loc}}; \mathbf{\Delta}(\delta_{ab})], \\ \hat{\mathbf{e}}_b &= [\mathbf{e}_b^{\text{glo}}; \mathbf{e}_b^{\text{loc}}; \mathbf{\Delta}(\delta_{ba})], \end{aligned} \tag{6}$$

where δ_{ab} denotes the relative distance from the first mention of e_a to that of e_b in the document. δ_{ba} is similarly defined. The relative distance is first divided into several bins $\{1, 2, \dots, 2^b\}$. Then, each bin is associated with a trainable distance embedding. $\mathbf{\Delta}(\cdot)$ associates each δ to a bin.

Then, we concatenate the final representations of e_a, e_b to form the *target relation representation* $\mathbf{o}_r = [\hat{\mathbf{e}}_a; \hat{\mathbf{e}}_b]$.

Furthermore, all relations in a document implicitly indicate the topic information of the document, such as *"director"* and *"character"* often appear in movies. In turn, the topic information implies possible relations. Some relations under similar topics are likely to co-occur, while others under different topics are not. Thus, we use self-attention (Sorokin and Gurevych, 2017) to capture *context relation representations*, which reveal the topic information of the document:

$$\mathbf{o}_c = \sum_{i=0}^{p} \theta_i \mathbf{o}_i = \sum_{i=0}^{p} \frac{\exp(\mathbf{o}_i \mathbf{W} \mathbf{o}'_r)}{\sum_{j=0}^{p} \exp(\mathbf{o}_j \mathbf{W} \mathbf{o}'_r)} \mathbf{o}_i, \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{d_r \times d_r}$ is a trainable parameter matrix. d_r is the dimension of target relation representations. $\mathbf{o}_i$ ($\mathbf{o}_j$) is the relation representation of the i^{th} (j^{th}) entity pair. θ_i is the attention weight for $\mathbf{o}_i$. p is the number of entity pairs.

Finally, we use a feed-forward neural network (FFNN) over the target relation representation $\mathbf{o}_r$ and the context relation representation $\mathbf{o}_c$ to make the prediction. Besides, considering that an entity pair may hold several relations, we transform the multi-classification problem into multiple binary classification problems. The predicted probability distribution of r over the set $\mathcal{R}$ of all relations is defined as follows:

$$\mathbf{y}_r = \text{sigmoid}(\text{FFNN}([\mathbf{o}_r; \mathbf{o}_c])), \tag{8}$$

where $\mathbf{y}_r \in \mathbb{R}^{|\mathcal{R}|}$.

We define the loss function as follows:

$$\mathcal{L} = -\sum_{r \in \mathcal{R}} \Big(y_r^* \log(y_r) + (1 - y_r^*) \log(1 - y_r) \Big), \tag{9}$$

where $y_r^* \in \{0, 1\}$ denotes the true label of r. We employ Adam optimizer (Kingma and Ba, 2015) to optimize this loss function.

Datasets		#Doc.	#Rel.	#Inst.	#N/A Inst.
CDR	Train	500	1	1,038	4,280
	Dev.	500	1	1,012	4,136
	Test	500	1	1,066	4,270
DocRED	Train	3,053	96	38,269	1,163,035
	Dev.	1,000	96	12,332	385,263
	Test	1,000	96	12,842	379,316

Table 1: Dataset statistics (Inst.: relation instances excluding N/A relation; N/A Inst.: negative examples).

4 Experiments and Results

We implemented our GLRE with PyTorch 1.5. The source code and datasets are available online.[1] In this section, we report our experimental results.

4.1 Datasets

We evaluated GLRE on two public document-level RE datasets. Table 1 lists their statistical data:

- The Chemical-Disease Relations (*CDR*) data set (Li et al., 2016) was built for the BioCreative V challenge and annotated with one relation *"chemical-induced disease"* manually.
- The *DocRED* dataset (Yao et al., 2019) was built from Wikipedia and Wikidata, covering various relations related to science, art, personal life, etc. Both manually-annotated and distantly-supervised data are offered. We only used the manually-annotated data.

4.2 Comparative Models

First, we compared GLRE with five sentence-level RE models adapted to the document level:

- Zhang et al. (2018) employed GCN over pruned dependency trees.

[1] `https://github.com/nju-websoft/GLRE`

- Yao et al. (2019) proposed four baseline models. The first three ones are based on CNN, LSTM and BiLSTM, respectively. The fourth context-aware model incorporates the attention mechanism into LSTM.

We also compared GLRE with nine document-level RE models:

- Zhou et al. (2016) combined feature-, tree kernel- and neural network-based models.
- Gu et al. (2017) leveraged CNN and maximum entropy.
- Nguyen and Verspoor (2018) integrated character-based word representations in CNN.
- Panyam et al. (2018) exploited graph kernels.
- Verga et al. (2018) proposed a bi-affine network with Transformer.
- Zheng et al. (2018) designed a hierarchical network using multiple BiLSTMs.
- Christopoulou et al. (2019) put forward an edge-oriented graph neural model with multi-instance learning.
- Wang et al. (2019) applied BERT to encode documents, and used a bilinear layer to predict entity relations. It improved performance by two phases. First, it predicted whether a relation exists between two entities. Then, it predicted the type of the relation.
- Tang et al. (2020) is a sequence-based model. It also leveraged BERT and designed a hierarchical inference network to aggregate inference information from entity level to sentence level, then to document level.

4.3 Experiment Setup

Due to the small size of CDR, some work (Zhou et al., 2016; Verga et al., 2018; Zheng et al., 2018; Christopoulou et al., 2019) created a new split by unionizing the training and development sets, denoted by *"train + dev"*. Under this setting, a model was trained on the train + dev set, while the best epoch was found on the development set. To make a comprehensive comparison, we also measured the corresponding precision, recall and F1 scores.

For consistency, we used the same experiment setting on DocRED. Additionally, the gold standard of the test set of DocRED is unknown, and only F1 scores can be obtained via an online interface. Besides, it was noted that some relation instances are present in both training and development/test sets (Yao et al., 2019). We also measured F1 scores ignoring those duplicates, denoted by *Ign F1*.

Models	**Train**			**Train + Dev**		
	P	R	F1	P	R	F1
Zhang et al.¶	52.3	72.0	60.6	58.1	**74.6**	65.3
Zhou et al.	64.9	49.3	56.0	55.6	68.4	61.3
Gu et al.	55.7	68.1	61.3	-	-	-
Nguyen and Verspoor	57.0	68.6	62.3	-	-	-
Panyam et al.	55.6	68.4	61.3	-	-	-
Verga et al.	55.6	70.8	62.1	63.3	67.1	65.1
Zheng et al.	45.2	68.1	54.3	56.2	68.0	61.5
Christopoulou et al.¶	62.7	66.3	64.5	61.5	73.6	67.0
Wang et al.¶	61.9	68.7	65.1	66.0	68.3	67.1
GLRE (ours)	**65.1**	**72.2**	**68.5**	**70.5**	74.5	**72.5**

¶ denotes that we performed hyperparameter tuning. For others, we reused the reported results due to the lack of source code.

Table 2: Result comparison on CDR.

Models	**Train**		**Train + Dev**	
	Ign F1	F1	Ign F1	F1
Zhang et al.¶	49.9	52.1	52.5	54.6
Yao et al. (CNN)	40.3	42.3	-	-
Yao et al. (LSTM)	47.7	50.1	-	-
Yao et al. (BiLSTM)	48.8	51.1	-	-
Yao et al. (Context-aware)	48.4	50.7	-	-
Christopoulou et al.¶	49.1	50.9	48.3	50.4
Wang et al.¶	53.1	55.4	54.5	56.5
Tang et al.	53.7	55.6	-	-
GLRE (ours)	**55.4**	**57.4**	**56.7**	**58.9**

Table 3: Result comparison on DocRED.

For GLRE and Wang et al. (2019), we used different BERT models in the experiments. For CDR, we chose BioBERT-Base v1.1 (Lee et al., 2019), which re-trained the BERT-Base-cased model on biomedical corpora. For DocRED, we picked up the BERT-Base-uncased model. For the comparative models without using BERT, we selected the PubMed pre-trained word embeddings (Chiu et al., 2016) for CDR and GloVe (Pennington et al., 2014) for DocRED. For the models with source code, we used our best efforts to tune the hyperparameters. Limited by the space, we refer interested readers to the appendix for more details.

4.4 Main Results

Tables 2 and 3 list the results of the comparative models and GLRE on CDR and DocRED, respectively. We have four findings below:

(1) The sentence-level RE models (Zhang et al., 2018; Yao et al., 2019) obtained medium performance. They still fell behind a few document-level models, indicating the difficulty of directly applying them to the document level.

(2) The graph-based RE models (Panyam et al., 2018; Verga et al., 2018; Christopoulou et al., 2019) and the non-graph models (Zhou et al., 2016; Gu et al., 2017; Nguyen and Verspoor,

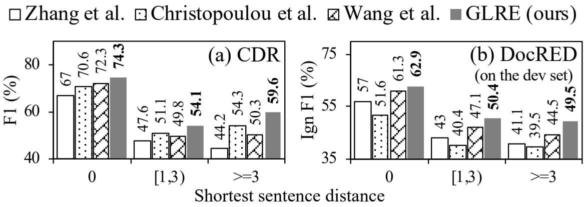


Figure 3: Results w.r.t. entity distance.

2018; Zheng et al., 2018) achieved comparable results, while the best graph-based model (Christopoulou et al., 2019) outperformed the best non-graph (Nguyen and Verspoor, 2018). We attribute it to the document graph on the entity level, which can better model the semantic information in a document.

(3) From the results of Wang et al. (2019); Tang et al. (2020), the BERT-based models showed stronger prediction power for document-level RE. They outperformed the other comparative models on both CDR and DocRED.

(4) GLRE achieved the best results among all the models. We owe it to entity global and local representations. Furthermore, BERT and context relation representations also boosted the performance. See our analysis below.

4.5 Detailed Analysis

Entity distance. We examined the performance of the open-source models in terms of entity distance, which is defined as the shortest sentence distance between all mentions of two entities. Figure 3 depicts the comparison results on CDR and DocRED using the training set only. We observe that:

(1) GLRE achieved significant improvement in extracting the relations between entities of long distance, especially when distance ≥ 3. This is because the global heterogeneous graph can effectively model the interactions of semantic information of different nodes (i.e., mentions, entities and sentences) in a document. Furthermore, entity local representations can reduce the influence of noisy context of multiple mentions of entities in long distance.

(2) According to the results on CDR, the graph-based model (Christopoulou et al., 2019) performed better than the sentence-level model (Zhang et al., 2018) and the BERT-based model (Wang et al., 2019) in extracting inter-sentential relations. The main reason is that it leveraged heuristic rules to construct the document graph at the entity level, which can better model the semantic information across sentences and avoid error accumulation involved by NLP tools, e.g., the dependency parser used in Zhang et al. (2018).

(3) On DocRED, the models (Wang et al., 2019; Zhang et al., 2018) outperformed the model (Christopoulou et al., 2019), due to the power of BERT and the increasing accuracy of dependency parsing in the general domain.

Number of entity mentions. To assess the effectiveness of GLRE in aggregating the information of multiple entity mentions, we measured the performance in terms of the average number of mentions for each entity pair. Similar to the previous analysis, Figure 4 shows the results on CDR and DocRED using the training set only. We see that:

(1) GLRE achieved great improvement in extracting the relations with average number of mentions ≥ 2, especially ≥ 4. The major reason is that entity local representations aggregate the contextual information of multiple mentions selectively. As an exception, when the average number of mentions was in $[1, 2)$, the performance of GLRE was slightly lower than Christopoulou et al. (2019) on CDR. This is because both GLRE and Christopoulou et al. (2019) relied on modeling the interactions between entities in the document, which made them indistinguishable under this case. In fact, the performance of all the models decreased when the average number of mentions was small, because less relevant information was provided in the document, which made relations harder to be predicted. We will consider external knowledge in our future work.

(2) As compared with Zhang et al. (2018) and Christopoulou et al. (2019), the BERT-based model (Wang et al., 2019) performed better in general, except for one interval. When the average number of mentions was in $[1, 2)$ on CDR, its performance was significantly lower than other models. The reason is twofold. On one hand, it is more difficult to capture the latent knowledge in the biomedical field. On the other hand, the model (Wang et al., 2019) only relied on the semantic information of the mentions of target entity pairs to predict the relations. When the average number was small, the prediction became more difficult. Furthermore, when the average number was large, its performance increase was not significant. The

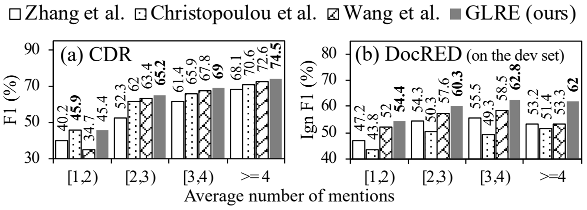


Figure 4: Results w.r.t. number of entity mentions.

Models	**CDR**			**DocRED**	
	P	R	F1	Ign F1	F1
GLRE	65.1	72.2	**68.5**	**55.4**	**57.4**
w/o BERT	**69.6**	66.5	68.0	51.6	53.6
w/o Entity global rep.	67.0	65.4	66.2	54.7	56.6
w/o Entity local rep.	60.9	68.5	64.5	54.6	56.4
w/o Context rel. rep.	60.5	**75.1**	67.1	54.6	56.8

Table 4: Results of ablation study.

main reason is that, although BERT brought rich knowledge, the model (Wang et al., 2019) indiscriminately aggregated the information of multiple mentions and introduced much noisy context, which limited its performance.

Ablation study. To investigate the effectiveness of each layer in GLRE, we conducted an ablation study using the training set only. Table 4 shows the comparison results. We find that: (1) BERT had a greater influence on DocRED than CDR. This is mainly because BERT introduced valuable linguistic knowledge and common-sense knowledge to RE, but it was hard to capture latent knowledge in the biomedical field. (2) F1 scores dropped when we removed entity global representations, entity local representations or context relation representations, which verified their usefulness in document-level RE. (3) Particularly, when we removed entity local representations, F1 scores dropped more dramatically. We found that more than 54% and 19% of entities on CDR and DocRED, respectively, have multiple mentions in different sentences. The local representation layer, which uses multi-head attention to selectively aggregate multiple mentions, can reduce much noisy context.

Pre-trained language models. To analyze the impacts of pre-trained language models on GLRE and also its performance upper bound, we replaced BERT-Base with BERT-Large, XLNet-Large (Yang et al., 2019) or ALBERT-xxLarge (Lan et al., 2020). Table 5 shows the comparison results using the training set only, from which we observe that larger models boosted the performance of GLRE to some extent. When the "train + dev" setting was used on DocRED, the Ign F1 and F1 scores of XLNet-Large even reached to 58.5 and 60.5, respectively. However, due to the lack of biomedical versions, XLNet-Large and ALBERT-xxLarge did not bring improvement on CDR. We argue that selecting the best pre-trained models is not our primary goal.

GLRE	**CDR**			**DocRED**	
	P	R	F1	Ign F1	F1
BERT-Base	65.1	72.2	68.5	55.4	57.4
BERT-Large	65.3	72.3	**68.6**	**56.8**	58.9
XLNet-Large	**66.1**	70.5	68.2	**56.8**	**59.0**
ALBERT-xxLarge	57.5	**80.6**	67.1	56.3	58.3

Table 5: Results w.r.t. different pre-training models.

Case study. To help understanding, we list a few examples from the CDR test set in Table 6. See Appendix for more cases from DocRED.

(1) From Case 1, we find that logical reasoning is necessary. Predicting the relation between *"rofecoxib"* and *"GI bleeding"* depends on the bridge entity *"non-users of aspirin"*. GLRE used R-GCN to model the document information based on the global heterogeneous graph, thus it dealt with complex inter-sentential reasoning better.

(2) From Case 2, we observe that, when a sentence contained multiple entities connected by conjunctions (such as *"and"*), the model (Wang et al., 2019) might miss some associations between them. GLRE solved this issue by building the global heterogeneous graph and considering the context relation information, which broke the word sequence.

(3) Prior knowledge is required in Case 3. One must know that *"fatigue"* belongs to *"adverse effects"* ahead of time. Then, the relation between *"bepridil"* and *"dizziness"* can be identified correctly. Unfortunately, both GLRE and Wang et al. (2019) lacked the knowledge, and we leave it as our future work.

We analyzed all 132 inter-sentential relation instances in the CDR test set that were incorrectly predicted by GLRE. Four major error types are as follows: (1) Logical reasoning errors, which occurred when GLRE could not correctly identify the relations established indirectly by the bridge entities, account for 40.9%. (2) Component missing errors, which happened when some component of a sentence (e.g., subject) was missing, account for 28.8%. In this case, GLRE needed the whole document information to infer the lost component and

... [S8] Among non-users of aspirin, the adjusted hazard ratios were: **rofecoxib** 1.27, naproxen 1.59, diclofenac 1.17 and ibuprofen 1.05. ... [S10] CONCLUSION: Among non-users of aspirin, naproxen seemed to carry the highest risk for AMI / **GI bleeding**. ...

Case 1 Label: CID GLRE: CID Wang et al.: N/A

... [S2] **S-53482** and **S-23121** are N-phenylimide herbicides and produced **embryolethality**, teratogenicity. ...

Case 2 Label: CID GLRE: CID Wang et al.: N/A

[S1] Clinical evaluation of **adverse effects** during **bepridil** administration for atrial fibrillation and flutter. ... [S8] There was marked QT prolongation greater than 0.55 s in 13 patients ... and general **fatigue** in 1 patient each. ...

Case 3 Label: CID GLRE: N/A Wang et al.: N/A

Table 6: Case study on the CDR test set. CID is short for the *"chemical-induced disease"* relation. **Target entities** and **related entities** are colored accordingly.

predict the relation, which was not always accurate. (3) Prior knowledge missing errors account for 13.6%. (4) Coreference reasoning errors, which were caused by pronouns that could not be understood correctly, account for 12.9%.

5 Conclusion

In this paper, we proposed GLRE, a global-to-local neural network for document-level RE. Entity global representations model the semantic information of an entire document with R-GCN, and entity local representations aggregate the contextual information of mentions selectively using multi-head attention. Moreover, context relation representations encode the topic information of other relations using self-attention. Our experiments demonstrated the superiority of GLRE over many comparative models, especially the big leads in extracting relations between entities of long distance and with multiple mentions. In future work, we plan to integrate knowledge graphs and explore other document graph modeling ways (e.g., hierarchical graphs) to improve the performance.

Acknowledgments. This work is supported partially by the National Key R&D Program of China (No. 2018YFB1004300), the National Natural Science Foundation of China (No. 61872172), and the Water Resource Science & Technology Project of Jiangsu Province (No. 2019046).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv:1607.06450*.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *ACL*, pages 756–765, Berlin, Germany. ACL.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *BioNLP*, pages 166–174, Berlin, Germany. ACL.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *EMNLP-IJCNLP*, pages 4925–4936, Hong Kong. ACL.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *ACL*, pages 423–429, Barcelona, Spain. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, Minneapolis, MN, USA. ACL.

Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2017. Chemical-induced disease relation extraction via convolutional neural network. *Database*, page bax024.

Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, and Thomas Runkler. 2019. Neural relation extraction within and across sentence boundaries. In *AAAI*, pages 6513–6520, Honolulu, HI, USA. AAAI Press.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *ACL 2004*, pages 178–181, Barcelona, Spain. ACL.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*, San Diego, CA, USA. OpenReview.net.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*, Addis Ababa, Ethiopia. OpenReview.net.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *CoNLL*, pages 333–342, Vancouver, Canada. ACL.

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*, page baw068.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *ACL*, pages 2124–2133, Berlin, Germany. ACL.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *ACL*, pages 1546–1557, Online. ACL.

Dat Quoc Nguyen and Karin Verspoor. 2018. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. In *BioNLP*, pages 129–136, Melbourne, Australia. ACL.

Nagesh C Panyam, Karin Verspoor, Trevor Cohn, and Kotagiri Ramamohanarao. 2018. Exploiting graph kernels for high performance biomedical relation extraction. *Journal of Biomedical Semantics*, 9:7.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. *TACL*, 5:101–115.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543, Doha, Qatar. ACL.

Lin Qiu, Hao Zhou, Yanru Qu, Weinan Zhang, Suoheng Li, Shu Rong, Dongyu Ru, Lihua Qian, Kewei Tu, and Yong Yu. 2018. QA4IE: A question answering based framework for information extraction. In *ISWC*, pages 198–216, Monterey, CA, USA. Springer.

Jianfeng Qu, Wen Hua, Dantong Ouyang, Xiaofang Zhou, and Ximing Li. 2019. A fine-grained and noise-aware method for neural relation extraction. In *CIKM*, pages 659–668, Beijing, China. ACM.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *EACL*, pages 1171–1182, Valencia, Spain. ACL.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *ACL*, pages 4309–4316, Florence, Italy. ACL.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *ESWC*, pages 593–607, Heraklion, Greece. Springer.

Stephen Soderland, David Fisher, Jonathan Aseltine, and Wendy Lehnert. 1995. CRYSTAL: Inducing a conceptual dictionary. In *IJCAI*, pages 1314–1319, Montréal, Canada. Morgan Kaufmann Publishers.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. N-ary relation extraction using graph-state LSTM. In *EMNLP*, pages 2226–2235, Brussels, Belgium. ACL.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *EMNLP*, pages 1784–1789, Copenhagen, Denmark. ACL.

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 *PAKDD*, pages 197–209, Singapore. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008, Long Beach, CA, USA. Curran Associates, Inc.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *NAACL*, pages 872–884, New Orleans, LA, USA. ACL.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune Bert for DocRED with two-step process. *arXiv:1909.11898*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5753–5763, Vancouver, Canada. Curran Associates, Inc.

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 *ACL*, pages 764–777, Florence, Italy. ACL.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. In *EMNLP*, Online. ACL.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, pages 1753–1762, Lisbon, Portugal. ACL.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344, Dublin, Ireland. ACL.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In *NAACL-HLT*, pages 3016–3025, Minneapolis, MN, USA. ACL.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *PACLIC*, pages 73–78, Shanghai, China. ACL.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *EMNLP*, pages 2205–2215, Brussels, Belgium. ACL.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *EMNLP*, pages 35–45, Copenhagen, Denmark. ACL.

Wei Zheng, Hongfei Lin, Zhiheng Li, Xiaoxia Liu, Zhengguang Li, Bo Xu, Yijia Zhang, Zhihao Yang, and Jian Wang. 2018. An effective neural model extracting document level chemical-induced disease relations from biomedical literature. *Journal of Biomedical Informatics*, 83:1–9.

Huiwei Zhou, Huijie Deng, Long Chen, Yunlong Yang, Chen Jia, and Degen Huang. 2016. Exploiting syntactic and semantics information for chemical-disease relation extraction. *Database*, page baw048.