

MRN: A Locally and Globally Mention-Based Reasoning Network for Document-Level Relation Extraction

Jingye Li¹, Kang Xu¹, Fei Li^{1*}, Hao Fei¹, Yafeng Ren², Donghong Ji¹

¹Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China

²Laboratory of Language and Artificial Intelligence, Guangdong University of Foreign Studies, China

{theodorelee, xukang0531, lifei_csnlp, hao.fei, renyafeng, dhji}@whu.edu.cn

Abstract

Document-level relation extraction aims to detect the relations within one document, which is challenging since it requires complex reasoning using mentions, entities, local and global contexts. Few previous studies have distinguished local and global reasoning explicitly, which may be problematic because they play different roles in intra- and inter-sentence relations. Moreover, the interactions between local and global contexts should be considered since they could help relation reasoning based on our observation. In this paper, we propose a novel mention-based reasoning (MRN) module based on explicitly and collaboratively local and global reasoning. Based on MRN, we design a co-predictor module to predict entity relations based on local and global entity and relation representations jointly. We evaluate our MRN model on three widely-used benchmark datasets, namely DocRED, CDR, and GDA. Experimental results show that our model outperforms previous state-of-the-art models by a large margin.

1 Introduction

Relation extraction (RE), identifying the semantic relations among target entities in the text, has long been a fundamental task in the natural language processing (NLP) community (Zeng et al., 2014; Xu et al., 2015). Prior efforts largely focus on sentence-level RE (Lin et al., 2016; Zhang et al., 2018). However, recent studies reveal that a large number of relations can actually be expressed through multiple sentences, which necessitates document-level RE (Yao et al., 2019). Compared with sentence-level RE, the entities for document-level relations may be mentioned in multiple sentences across a document. Therefore, document-level RE requires capturing the complex interactions between all enti-

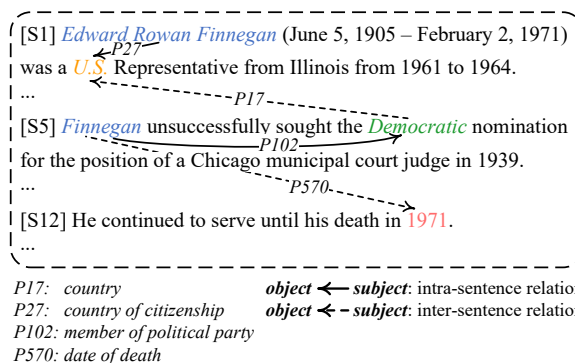


Figure 1: An example of document-level RE from the DocRED dataset. We use the same color to denote the mentions of the same entity.

ties in the entire document (Nan et al., 2020; Zeng et al., 2020; Wang et al., 2020).

It is well known that local and global contexts are two key performance enhancers for the task. Intuitively, the former can benefit the identification of nearer (e.g., intra-sentence) relations, while the latter is more useful for distant (e.g., inter-sentence) relations. For document-level RE, such context information is closely related to the ubiquitous mentions in a document, so **mention-based reasoning** with different context granularities is highly important for the task. However, most previous studies have not distinguished local and global reasoning explicitly (Peng et al., 2017; Christopoulou et al., 2019; Sahu et al., 2019; Nan et al., 2020; Zhou et al., 2020; Zeng et al., 2020; Li et al., 2020; Zhang et al., 2020). Recently Wang et al. (2020) investigate local and global contexts for document-level RE by performing global and local reasoning consecutively. However, their pipeline method can be problematic because it ignores the interactions and communications of local and global contexts,¹ which limits the performance of the task.

¹From Figure 1, it is shown that the intra-sentence relations P27 and P102 can help to identify the inter-sentence relation P17, since ‘U.S.’ and ‘Democratic’ are linked through ‘Finnegan’.

*Corresponding author.

In this paper, we aim to address the above issues for improving document-level RE by presenting a novel **Mention-based Reasoning Network (MRN)**. As shown in Figure 2, MRN consists of several innovative modules for modeling the relation reasoning locally and globally, including (1) a two-dimensional windowed convolution for capturing the local mention-to-mention interactions between the subject and object arguments of relations, and (2) a co-attention module for capturing the global interaction between each pair of entity mentions. Note that the above two modules also provide mutual information to each other, in order to capture the interactions between local and global contexts.

Moreover, different from previous work that expresses entities with just one kind of representation (Peng et al., 2017; Christopoulou et al., 2019; Sahu et al., 2019; Nan et al., 2020; Zhou et al., 2020; Zeng et al., 2020; Li et al., 2020; Zhang et al., 2020), our method distinguishes mentions into subjects and objects, and generates entity representations from both local and global mention-based reasoning. Specifically, we design a novel module, called **co-predictor**, to utilize both local and global entity representations for jointly reasoning the relations between close and distant entities.

We conduct extensive experiments on three widely-used benchmarks, including DocRED (Yao et al., 2019), CDR (Li et al., 2016) and GDA (Wu et al., 2019). The results show that our MRN model outperforms the current best model by a large margin, demonstrating its advances. In summary, we make the following contributions:

- We propose a mention-based reasoning network (MRN), to distinguish the impacts of close and distant entity mentions in relation extraction and meanwhile consider the interactions between local and global contexts, which we call locally and globally mention-based reasoning.
- We also propose a co-predictor to work in concert with the mention reasoning block, and predict the relation of a pair of entities using local and global features simultaneously.
- Our model achieves the state-of-the-art performances on three benchmark datasets for document-level RE. We also conduct extensive analyses of our model to better understand its working mechanism.²

²Codes are publicly available at <https://github.com/ljynlp/MRN>

2 Related Work

Relation extraction (RE), including sentence-level RE and document-level RE, plays a crucial role in a wide variety of knowledge-based applications, such as question answering (Hixon et al., 2015), dialogue generation (He et al., 2017), etc. Recent studies largely focus on sentence-level RE by various neural network methods, such as CNN (Zeng et al., 2014; dos Santos et al., 2015), BiLSTM (Zhang et al., 2015; Cai et al., 2016), attention mechanism (Wang et al., 2016; Lin et al., 2016), and neural graph models (Zhang et al., 2018; Zhu et al., 2019). However, in practice, many relational facts need to be inferred across multiple sentences in a document, so researchers have shown a growing interest in document-level RE.

Compared with sentence-level RE, document-level RE needs to consider the complicated interactions between entities across multiple sentences. With this in mind, researchers begin to use graph neural networks to reason intra- and inter-sentence relations and make certain progress in extracting inter-sentence relations with document-level graph convolutional neural network (Peng et al., 2017; Velickovic et al., 2018; Christopoulou et al., 2019). Sahu et al. (2019). For example, Zhou et al. (2020) use entities as nodes and the context between entity pairs as edges to construct graphs. Nan et al. (2020) treat the graph as a latent structure and perform relational reasoning. However, most existing approaches only use entity-level information and ignore mention-level information.

Some studies also take mentions into account by adding mention nodes in the graph. For instance, Christopoulou et al. (2019) put mentions and entities in the same graph. Li et al. (2020) utilize a dual-tier heterogeneous graph to propagate relational information among entity mentions, and then summarize them into corresponding entities. More recently, Wang et al. (2020) use multi-head attention to aggregate multiple mentions of specific entities. Different from the above methods, Zeng et al. (2020) propose a graph aggregation and inference network that includes two graphs, one for capturing complex interactions among different mentions, and the other for integrating mention-level information of the same entities. Although these methods introduce mention-level nodes or graphs, none of them consider local mention-based contextual information or mention-level relative distances, which are all considered in our model.

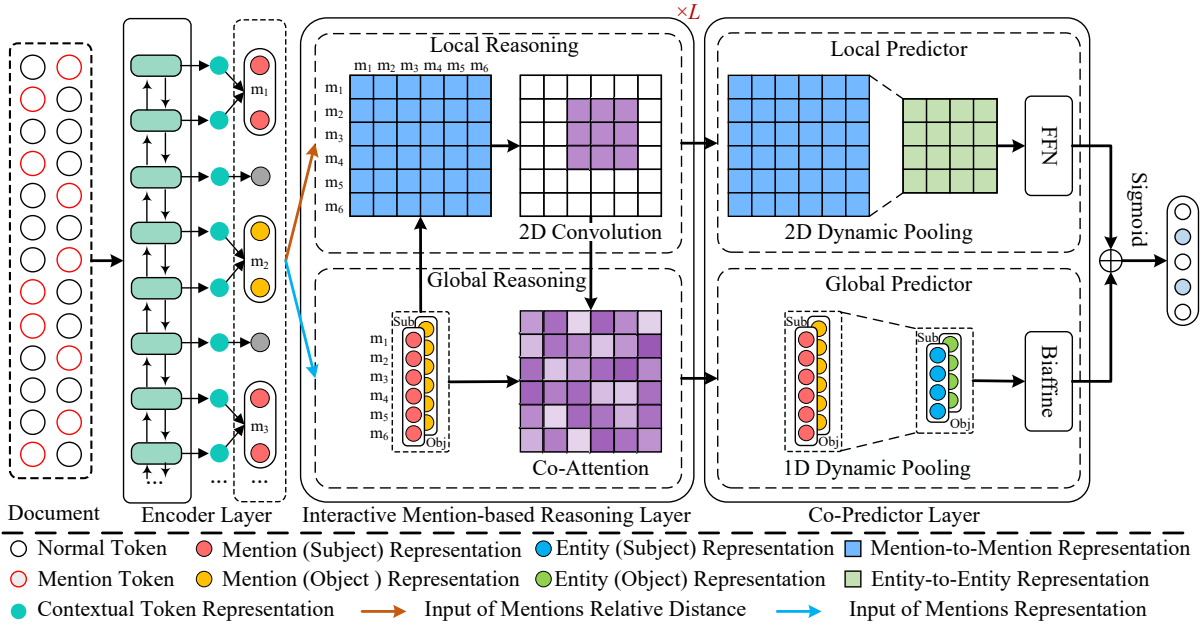


Figure 2: The overall architecture of our MRN, including three key components: encoder (§ 3.1), interactive mention-based reasoning (§ 3.2) and co-predictor (§ 3.3).

3 Methodology

Task Formulation Given an input document consisting of N tokens $\mathcal{D} = \{w_i\}_{i=1}^N$ and P entities $\mathcal{E} = \{e_i\}_{i=1}^P$, the task aims to extract a subset of relations from \mathcal{R} between the entity pairs (e_i^s, e_j^o) , where \mathcal{R} is a pre-defined relation type set, e_i^s and e_j^o are identified as subject and object entities. An entity e_i can appear multiple times in the document via K_{e_i} mentions $\mathcal{M}_{e_i} = \{m_j\}_{j=1}^{K_{e_i}}$.

Framework As illustrated in Figure 2, the overall architecture consists of three tiers. An encoder first yields contextual representations, and then the mention-based reasoning block (in multiple layers) performs local-level feature extraction with a two-dimensional (2D) convolution, and global-level feature retrieval with a co-attention module. Afterward, a co-predictor layer that contains a 1D and 2D dynamic pooling aggregates the entity-level features from the mention-based reasoning block. Finally, a multi-layer perceptron and a biaffine classifier are leveraged for jointly reasoning the relations between subject and object entities.

3.1 Encoding Layer

We first map each word w_i into a vector, and concatenate it with its corresponding entity type t_i :³

³ t_i denotes the type of entity mention that contains this word (e.g. if an entity type is `Person`, its mention word type is also `Person`).

$$\mathbf{x}_i = [\mathbf{x}_i^w; \mathbf{x}_i^t]. \quad (1)$$

Then, we adopt BiLSTM to encode the vectorial word representations into contextualized word representations:

$$\mathbf{h}_i = \text{BiLSTM}(\mathbf{x}_i), \quad (2)$$

where \mathbf{h}_i is the token hidden representation. Note that we also can use the BERT (Devlin et al., 2019) as an alternative to improve performances. Based on \mathbf{h}_i , we can obtain the mention representation:

$$\mathbf{m}_i = \text{Max}\{\mathbf{h}_j\}_{j=a_i}^{b_i}, \quad (3)$$

where a_i and b_i are the start and end positions of the i -th mention, respectively.

3.2 Interactive Mention-based Reasoning Layer

As we argued earlier, local and global context information is closely related to the ubiquitous mentions in a document. We thus propose a mention-based module for multi-hop reasoning among the relationships of all the mentions. Considering that there exist overlapping relations where multiple relations share the same mention, we distinguish mentions into subjects and objects by their directions. Moreover, near neighbors are more informative than distant ones for determining relations. To this end,

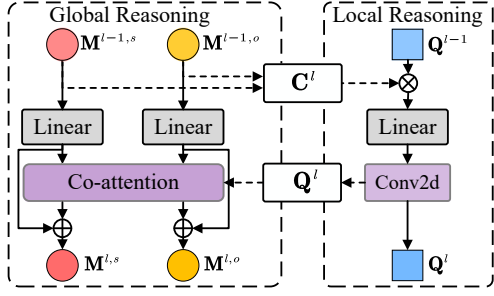


Figure 3: The interactive mention-based reasoning block. \oplus and \otimes represent the addition and concatenation operations.

we adopt a two-dimensional (2D) windowed convolution to capture local interactions between close subject and object mentions.

As shown in Figure 3, we perform multi-hop reasoning by stacking multiple layers (i.e., L) of our interactive mention-based reasoning blocks. The inputs of the l -th layer block include: (1) a set of subject mention representations $\mathbf{M}^{l-1,s} = \{\mathbf{m}_i^{l-1,s}\}_{i=1}^T$, (2) a set of object mention representations $\mathbf{M}^{l-1,o} = \{\mathbf{m}_i^{l-1,o}\}_{i=1}^T$, and (3) a set of mention relation representations \mathbf{Q}^{l-1} , where T is the number of mentions and \mathbf{Q}_{ij}^{l-1} denotes the type of implicit relation between the subject i and object j . For initialization, $\mathbf{m}_i^{0,s} = \mathbf{m}_i^{0,o} = \mathbf{m}_i$, and \mathbf{Q}^0 denotes relative distance embeddings of mentions (Zeng et al., 2014).

Local Reasoning To model the interactions between near subjects and objects, we first build a relationship tensor \mathbf{C}^l from $\mathbf{M}^{l-1,s}$ and $\mathbf{M}^{l-1,o}$, in which $\mathbf{C}_{ij}^l = [\mathbf{m}_i^{l-1,s}; \mathbf{m}_j^{l-1,o}]$. Next, we concatenate \mathbf{C}^l and \mathbf{Q}^{l-1} , and adopt a Feedforward Network (FFN) to reduce their dimensions. We then employ a 2D convolution to capture local contextual interactions, which can be regarded as extracting subgraph representations from a fully-connected bipartite graph containing subject and object nodes. The overall process can be formulated as:

$$\mathbf{Q}^l = \sigma(\text{Conv2d}(\text{FFN}([\mathbf{Q}^{l-1}; \mathbf{C}^l]))) , \quad (4)$$

where σ is a LeakyReLU activation function.

Global Reasoning We introduce a co-attention mechanism to compute attention coefficients that indicate the importance of subjects to objects and vice versa, so that the interaction between each pair of mentions can be considered. Inspired by the

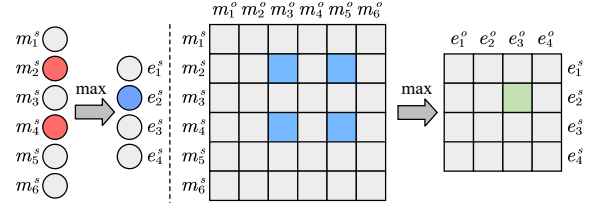


Figure 4: The 1D and 2D dynamic pooling.

success of graph attention network (GAT) (Velickovic et al., 2018), we apply two learnable linear transformations to transform subject and object mention representations $\mathbf{M}^{l-1,s}$ and $\mathbf{M}^{l-1,o}$ into higher-level features. Then, we leverage the mention relation representation \mathbf{Q}^l to calculate the coefficients and inject them into the co-attention process:

$$\hat{\mathbf{m}}_i^{l,\phi} = \sum_{j \in \mathcal{M}^{l,\psi}} \alpha_{ij}^{l,\psi} \mathbf{W}^{l,\psi} \mathbf{m}_j^{l-1,\psi} , \quad (5)$$

where $\mathbf{W}^{l,\psi}$ is a learnable parameter matrix, and $\alpha_{ij}^{l,\psi} = \text{Softmax}(\text{FFN}(\mathbf{Q}^l))$. Here $\phi, \psi \in \{s, o\}$. Afterward, the mention representation of the next layer $\mathbf{m}_i^{l,\phi}$ is generated by adding the residual of the last layer $\mathbf{m}_i^{l-1,\phi}$ and the non-linear transformation of the co-attention output $\hat{\mathbf{m}}_i^{l,\phi}$, which can be formulated as:

$$\mathbf{m}_i^{l,\phi} = \sigma(\hat{\mathbf{m}}_i^{l,\phi} + \mathbf{W}^{l,\phi} \mathbf{m}_i^{l-1,\phi}) + \mathbf{m}_i^{l-1,\phi} . \quad (6)$$

3.3 Co-Predictor Layer

After the last reasoning block (i.e., the L -th block), we obtain the final mention representations $\mathbf{M}^{L,s}$ and $\mathbf{M}^{L,o}$, as well as the mention relation representation \mathbf{Q}^L . Since different mentions may belong to different entities, we apply 1D and 2D max-pooling to aggregate mention-level features into the entity level (Figure 4). Then, we apply two predictors to calculate two relation distributions for entity pair (e_i, e_j) and then combine them for obtaining the final prediction.

Local Predictor Based on the mention relation representation \mathbf{Q}^L generated from the mention-based reasoning block, we adopt a 2D dynamic max-pooling (cf. Figure 4 right) to aggregate mention-level features into entity-level features:

$$\mathbf{G}_{ij} = \text{Max}\{\mathbf{Q}_{kt}^L\}_{k \in \mathcal{M}_{e_i}, t \in \mathcal{M}_{e_j}} , \quad (7)$$

where \mathcal{M}_{e_i} and \mathcal{M}_{e_j} are the mention sets corresponding to the i -th and j -th entities, respectively.

Then we employ a FFN to generate prediction scores for entity pair:

$$y'_{ij} = \text{FFN}(\mathbf{G}_{ij}). \quad (8)$$

Global Predictor Based on the mention representations $\mathbf{M}^{L,s}$ and $\mathbf{M}^{L,o}$, the representations of the i -th and j -th entities can also be generated by the 1D dynamic max-pooling (Figure 4 left).

$$\mathbf{e}_i^\phi = \text{Max}\{\mathbf{m}_k^{L,\phi}\}_{k \in \mathcal{M}_{e_i}}, \quad (9)$$

where \mathcal{M}_{e_i} is the mention set of the i -th entity, ϕ can be $\{s, o\}$ and $\mathbf{m}^{L,\phi} \in \mathbf{M}^{L,\phi}$. Then, a bi-affine classifier (Dozat and Manning, 2017) is used to compute the relation scores between a pair of subject and object entities:

$$\mathbf{z}_i^s = \text{MLP}^s(\mathbf{e}_i^s), \quad (10)$$

$$\mathbf{z}_j^o = \text{MLP}^o(\mathbf{e}_j^o), \quad (11)$$

$$y''_{ij} = \mathbf{z}_i^{s\top} \mathbf{U} \mathbf{z}_j^o + \mathbf{W}[\mathbf{z}_i^s; \mathbf{z}_j^o] + b, \quad (12)$$

where \mathbf{U} , \mathbf{W} and b are trainable parameters.

Joint Prediction The final relation probability for entity pair (e_i, e_j) with regards to the relationship r comes from the combination of the scores from both local and global predictors:

$$P(r|e_i, e_j) = \text{Sigmoid}(y'_{ij} + y''_{ij}). \quad (13)$$

3.4 Learning

Considering the imbalance of positive and negative samples in document-level RE, we use asymmetric loss (ASL) (Ben-Baruch et al., 2020) instead of binary cross-entropy loss:

$$L_+ = (1 - P(r|e_i, e_j))^{\gamma_+} \log(P(r|e_i, e_j)), \quad (14)$$

$$L_- = (P_n(r|e_i, e_j))^{\gamma_-} \log(1 - P_n(r|e_i, e_j)), \quad (15)$$

where γ_+ and γ_- are the focusing hyper-parameters for positive and negative samples, which aim to emphasize the contribution of positive samples and meanwhile down-weight the contribution of easy negatives samples ($\gamma_- > \gamma_+$). $P_n(r|e_i, e_j) = \max(P(r|e_i, e_j) - n, 0)$ is a probability shift mechanism that further filters out easy negative samples (probability margin $n \geq 0$ is a hyper-parameter). Here, the final loss function can be formulated as:

$$L = - \sum_{\mathcal{D} \in \mathcal{S}} \sum_{i \neq j} \sum_{r \in \mathcal{R}} \mathbb{I}(r = 1) L_+ + \mathbb{I}(r = 0) L_-, \quad (16)$$

where \mathcal{S} denotes the whole dataset, and $\mathbb{I}(\cdot)$ refers to the indicator function.

Statistics	DocRED	CDR	GDA
# Train	3,053	500	23,353
# Dev	1,000	500	5,839
# Test	1,000	500	1,000
# Relations	97	2	2
Relations instances	63,443	3,116	46,343
Avg.# entities per Doc.	19.52	6.82	4.84
Avg.# mentions per Doc.	26.20	19.23	18.45

Table 1: Statistics of the three datasets.

4 Experimental Settings

4.1 Dataset and Evaluation

We conduct experiments on three benchmark datasets, including DocRED (Yao et al., 2019), CDR (Li et al., 2016) and GDA (Wu et al., 2019), and the statistical information is shown in Table 1. DocRED is a large-scale human-annotated dataset for document-level RE, including 3,053 documents for training, 1,000 for development and 1,000 for test. CDR and GDA are the widely-used document-level RE in the biomedical domain. CDR contains 1,500 PubMed abstracts with 3,116 relational facts, and GDA consists of 30,192 MEDLINE abstracts and 46,343 relational facts.

Following previous studies (Yao et al., 2019; Nan et al., 2020), we adopt F1 and Ign F1 as the evaluation metrics, where Ign F1 is calculated by excluding the common relation facts shared by the training, development and test sets. Depending on whether relation arguments occur within one sentence or not, F1 can be further split into intra-F1 and inter-F1. The results of our model are presented after a significant test ($p \leq 0.03$).

4.2 Baselines

We make comparisons with the current state-of-the-art systems, including two categories. **1) Sequence-based methods**, which use different neural architectures to encode the entire document, include CNN (Zeng et al., 2014), BiLSTM (Cai et al., 2016), Context-Aware (Sorokin and Gurevych, 2017) and HIN (Tang et al., 2020). **2) Graph-based methods**, which construct homogeneous or heterogeneous graphs based on the whole document, include GAT (Velickovic et al., 2018), GCNN (Sahu et al., 2019), EoG (Christopoulou et al., 2019), LSR (Nan et al., 2020), GAIN (Zeng et al., 2020) and GLRE (Wang et al., 2020). Besides, some models leverage BERT (Devlin et al., 2019) for task improvements, including two-phase+BERT (Wang et al., 2019) and Coref+BERT

		Dev				Test	
		Ign F1	F1	Intra-F1	Inter-F1	Ign F1	F1
• Without BERT							
Sequence-based	CNN (Yao et al., 2019)	41.58	43.45	51.87	37.58	40.33	42.26
	BiLSTM (Yao et al., 2019)	48.87	50.94	57.05	43.49	48.78	51.06
	Context-Aware (Yao et al., 2019)	48.94	51.09	56.74	42.26	48.40	50.70
	HIN-GloVe (Tang et al., 2020)	51.06	52.95	-	-	51.15	53.30
Graph-based	GAT (Velickovic et al., 2018)	45.17	51.44	58.14	43.94	47.36	49.51
	GCNN (Sahu et al., 2019)	46.22	51.52	57.78	44.11	49.59	51.62
	EoG (Christopoulou et al., 2019)	45.94	52.15	58.90	44.60	49.48	51.82
	LSR (Nan et al., 2020)	48.82	55.17	60.83	48.35	52.15	54.18
Ours	GAIN (Zeng et al., 2020)	53.05	55.29	61.67	48.77	52.66	55.08
Ours	MRN	56.62	58.69	65.24	50.91	56.19	58.46
• With BERT							
Sequence-based	BERT (Wang et al., 2019)	-	54.16	61.61	47.15	-	53.20
	Two-Phase+BERT (Wang et al., 2019)	-	54.42	61.80	47.28	-	53.92
	Coref+BERT (Ye et al., 2020)	55.32	57.51	-	-	54.54	56.96
Graph-based	LSR+BERT (Nan et al., 2020)	52.43	59.00	65.26	52.05	56.97	59.05
	GLRE (Wang et al., 2020)	-	-	-	-	55.40	57.40
	GAIN+BERT (Zeng et al., 2020)	59.14	61.22	67.10	53.90	59.00	61.24
Ours	MRN+BERT	59.74	61.61	67.74	54.43	59.52	61.74

Table 2: Performances on DocRED. The results of baselines are from their related papers.

(Ye et al., 2020).

5 Results and Analyses

5.1 Results on DocRED

Table 2 shows the experimental results on the DocRED dataset. We can find that MRN achieves 56.19% Ign F1 and 58.46% F1 on the test set, outperforming the sequence-based models with large margins. Specifically, MRN outperforms the best sequence-based model HIN by 5.04% and 5.16% in terms of Ign F1 and F1, demonstrating the effectiveness of capturing the mention-level contextual information for document-level RE.

Moreover, we observe that graph-based models (e.g., LSR and GAIN) generally perform better than sequence-based models (e.g., BiLSTM and HIN). This verifies the usefulness of constructing informative graphs. In particular, our MRN achieves the best Ign F1 and F1 scores, and outperforms GAIN by 3.53% and 3.38%, respectively. This shows the importance of emphasizing local mention-level interactions for document-level RE.

Furthermore, we observe that the performance can be substantially boosted with the help of BERT, where the Ign-F1 and F1 increase by 3.33% and 3.28% on the test set. Notably, MRN with GloVe embeddings is able to achieve better results than several BERT-based models, such as CorefBERT and GLRE. This suggests that our model is more

	F1	Intra-F1	Inter-F1
• CDR data			
ME-CNN (Gu et al., 2017)	61.3	57.2	11.7
BRAN (Verga et al., 2018)	62.1	-	-
C-CHAR (Nguyen and Verspoor, 2018)	62.3	-	-
GCNN (Sahu et al., 2019)	58.6	-	-
EoG (Christopoulou et al., 2019)	63.6	68.2	50.9
DHG (Zhang et al., 2020)	64.7	68.6	54.1
LSR (Nan et al., 2020)	64.8	68.9	53.1
MRN	65.9	70.4	54.2
• GDA data			
EoG (Christopoulou et al., 2019)	81.5	85.2	50.0
DHG (Zhang et al., 2020)	82.2	85.4	52.4
LSR (Nan et al., 2020)	82.2	85.4	51.1
MRN	82.9	86.1	53.5

Table 3: Results on CDR and GDA datasets.

effective in capturing complex interactions between close and distant mentions even without the help of pre-trained embeddings.

5.2 Results on CDR and GDA

Table 3 shows the results on two biomedical datasets. Here, the baselines are also divided into sequence-based models (ME-CNN, BRAN, and C-CHAR) and graph-based models (GCNN, EoG, DHG, and LSR). Similar to the DocRED dataset, the graph-based models generally outperform the sequence-based models on CDR, which reveals the effectiveness of incorporating structural information and reasoning mechanisms in document-level

Model	Ign F1	F1	Intra-F1	Inter-F1
MRN	56.62	58.69	65.24	50.91
- entity type	55.70	57.61	64.25	49.74
- relative distance	55.61	57.64	64.33	49.70
- mention-based reasoning	53.71	55.66	61.55	49.14
- global predictor	52.77	54.53	60.81	46.99
- local predictor	53.73	55.70	61.95	49.01
dynamic average pooling	52.43	54.51	60.85	48.02
sharing object and subject	55.76	57.70	64.77	49.40
binary cross-entropy	56.08	58.02	64.52	50.29

Table 4: Ablation studies on the DocRED dataset.

RE. Besides, our MRN model achieves better performances than the state-of-the-art models on CDR and GDA datasets, outperforming LSR by 1.1% and the DHG by 0.7% respectively.

5.3 Intra- and Inter-sentence Relation Extraction

According to recent work (Yao et al., 2019), identifying 40.7% relations need the information of multiple sentences, which indicates that the reasoning ability of a model plays an important role in document-level RE. Thus, we also report the performances of intra- and inter-sentence relation extraction on three datasets in Table 2 and 3. We find that our model outperforms the current best models on all datasets in regard to both intra- and inter-F1. For example, MRN improves the intra-F1 by 3.57% and inter-F1 by 2.14% compared with GAIN on the development set of DocRED. This shows that mention-level reasoning is highly effective to capture complex interactions between mention objects and subjects, especially when not only local contexts but also long-range dependencies are considered.

5.4 Ablation Studies

We ablate each part of our MRN model on the development set of DocRED, as shown in Table 4. First, without entity type embeddings at the encoding layer, we observe slight performance drops. By removing relative distance information, the performance also decreases in a small degree. Furthermore, after removing interactive mention-based reasoning layer, global or local predictors, the performance goes down significantly. In particular, we find that the decrease of inter-F1 after removing global predictor (3.92%=50.91%-46.99%) is obviously higher than that for local predictor (1.90%=50.91%-49.01%), which verifies the usefulness of global features for long-dependency relation reasoning. A significant drop can be found

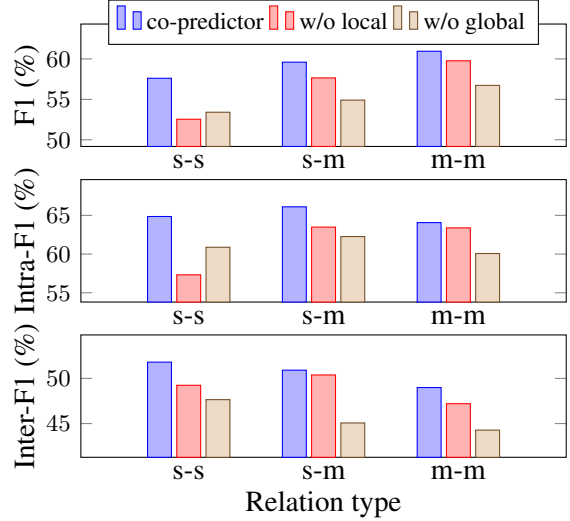


Figure 5: F1 scores of different relation types according to their entity arguments (e.g., (s)ingle mention and (m)ultiple mentions).

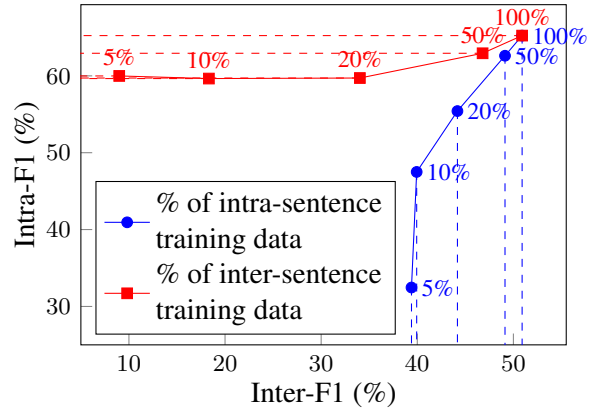


Figure 6: Intra- and inter-F1 scores vs. the proportions of inter- and intra-sentence training instances.

when replacing the dynamic max pooling with average pooling. Finally, sharing object and subject representations and using binary cross-entropy loss, the model also has a certain degree of degradation.

5.5 Effect Analysis for Co-Predictor

In this section, we investigate the effect of global and local predictors for MRN. We divide the relation instances in the development set of DocRED into three groups: the one where both subject and object arguments have single mention, the one where either subject or object argument has single mention, and the one where both subject and object arguments have multiple mentions. We also evaluate our model using different predictor configurations.

As shown in Figure 5, the model with both local and global predictors consistently outperforms

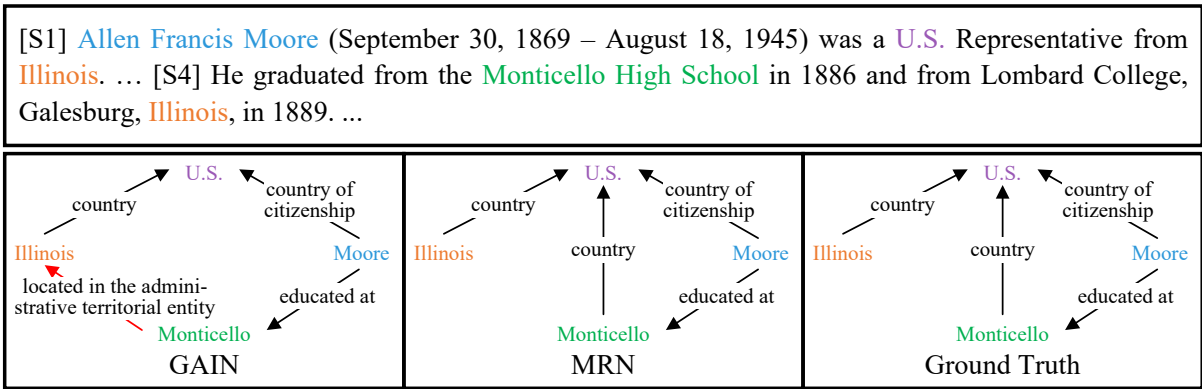


Figure 7: A case study for our model MRN and the state-of-the-art model, GAIN (Zeng et al., 2020). Only partial entities in the document are shown due to the space limitation. The red arrow denotes the wrong prediction.

the other ones. The F1s increase as the times that subjects and objects are mentioned increase (from s-s to m-m), indicating that multiple mentions appeared in various positions can provide more information to models. When removing the local predictor, we can observe that there is a huge drop for the s-s group, especially in Intra-F1. This demonstrates that the s-s group are mostly intra-sentence relations and they depend on local reasoning mostly. Moreover, if the global predictor is discarded, inter-F1 scores for the groups where subjects and objects are mentioned multiple times (s-m and m-m) drop the most. This reveals that the global predictor is more beneficial for extracting relations from multiple-mention entities or inter-sentence entities.

5.6 Effect Analysis for Inter- and Intra-sentence Training Data

As shown in Figure 6, we analyze the variation of inter- and intra-F1 scores when increasing or decreasing the proportions of intra- and inter-sentence training instances on the DocRED dataset. Note that the proportions of intra- and inter-sentence relation instances are 54.5% and 45.5% in the training set. The experimental setting is as below: first, we use 5% of inter-sentence training instances and observe the intra-F1. Then we gradually increase the percentage such as 10%, 20%, 50% and 100%. During this process, all intra-sentence training instances are used. The object for the above steps is to observe the effect of inter-sentence training instances on the intra-F1. In addition, we conduct similar experiments to observe the inter-F1 by gradually increasing the proportion of intra-sentence training instances.

As the red line shows in Figure 6, the inter-F1 is influenced slightly by the number of intra-sentence

training instances. In contrast, the number of intra-sentence training instances has a significant impact on the inter-F1, since the inter-F1 grows dramatically (the blue line) when more intra-sentence training instances are added. This suggests that the interactions between intra- and inter-sentence relations indeed exist, and one may be helpful for reasoning the other.

5.7 Case Study

As shown in Figure 7, we present a case study to better understand the effect of our proposed MRN, in comparison with previous state-of-the-art baseline GAIN. We can observe that *Monticello* is the object of the intra-sentence relation triple (*Moore*, *educated at*, *Monticello*) and also the subject of the inter-sentence (*Monticello*, *country*, *U.S.*). However, GAIN fails to identify the relation between *U.S.* and *Monticello*, while MRN deduces it successfully. This demonstrates that the effectiveness of distinguishing subjects and objects at the inference stage, and MRN has strong capability for inter-sentence reasoning. Meanwhile, GAIN has made a wrong prediction between *Monticello* and *Illinois* in the 4-th sentence, indicating that our model has better local inference ability.

6 Conclusion

We propose a novel mention-based reasoning network (MRN) for document-level relation extraction. Our model is capable of capturing local and global contextual information as well as close and distant mention interactions, via multiple mechanisms such as a multi-hop mention-level reasoning block and collaborative predictors. Experimental results show that our proposed model achieves new state-

of-the-art on three widely-used datasets. Through empirical analyses, we find that it is reasonable for document-level RE models to pay more attention on local context and close mentions. Meanwhile, global context and distant mention interactions are also highly important for document-level RE. Last but not least, joint reasoning with local and global context information is a reasonable and effective method for the task.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61772378), the National Key Research and Development Program of China (No. 2017YFC1200500), the Research Foundation of Ministry of Education of China (No. 18JZD015).

References

- Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2020. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*.
- Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the ACL*, pages 756–765.
- Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174.
- Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the EMNLP-IJCNLP*, pages 4925–4936.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL*, pages 4171–4186.
- Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the ICLR*.
- Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2017. Chemical-induced disease relation extraction via convolutional neural network. *Database*, 2017.
- He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the ACL*, pages 1766–1776.
- Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning knowledge graphs for question answering through conversational dialog. In *Proceedings of the NAACL*, pages 851–861.
- Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. 2020. Graph enhanced dual attention network for document-level relation extraction. In *Proceedings of the COLING*, pages 1551–1560.
- 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*, 2016.
- Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the ACL*, pages 2124–2133.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the ICLR*.
- Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the ACL*, pages 1546–1557.
- Dat Quoc Nguyen and Karin Verspoor. 2018. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. In *Proceedings of the BioNLP 2018 workshop*, pages 129–136.
- 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 EMNLP*, pages 1532–1543.
- Sunil Kumar Sahu, Fenja Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of the ACL*, pages 4309–4316.
- Cícero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the ACL-IJCNLP*, pages 626–634.

- Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the EMNLP*, pages 1784–1789.
- Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. Hin: Hierarchical inference network for document-level relation extraction. In *Proceedings of the PAKDD*, pages 197–209.
- Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the ICLR*.
- Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the NAACL*, 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 EMNLP*, pages 3711–3721.
- Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *arXiv preprint arXiv:1909.11898*.
- Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the ACL*, pages 1298–1307.
- Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *Proceedings of the RECOMB*, pages 272–284.
- Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015. Semantic relation classification via convolutional neural networks with simple negative sampling. In *Proceedings of the EMNLP*, pages 536–540.
- Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the ACL*, 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 EMNLP*, pages 7170–7186.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the COLING*, pages 2335–2344.
- Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the EMNLP*, pages 1630–1640.
- Shu Zhang, Dequan Zheng, Xinchun Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the PACLIC*, pages 73–78.
- Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the EMNLP*, pages 2205–2215.
- Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level relation extraction with dual-tier heterogeneous graph. In *Proceedings of the COLING*, pages 1630–1641.
- Huiwei Zhou, Yibin Xu, Weihong Yao, Zhe Liu, Chengkun Lang, and Haibin Jiang. 2020. Global context-enhanced graph convolutional networks for document-level relation extraction. In *Proceedings of the COLING*, pages 5259–5270.
- Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the ACL*, pages 1331–1339.

A Effect Analysis for Sentence Number

As shown in Figure 8, we display the F1 scores of MRN and GAIN in the development set of DocRED with regards to the document length. As seen, we count the document length using the number of sentences within a document, which varies from 3 to 13. Results show that MRN attains better performances than GAIN, no matter that how the document length changes. In addition, the performance difference between MRN and GAIN becomes larger, when the document becomes longer. For instance, MRN outperforms GAIN by about 10% with regards to the group where the document length is 13. This demonstrates that our model is more robust for long documents.

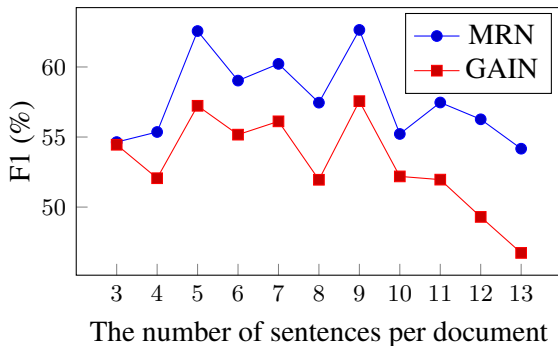


Figure 8: Comparisons between MRN and GAIN about the effect of the document length on the F1.

B Architecture Analysis for MRN Block

We conduct experiments for the interactive mention-based reasoning block based on the development set of DocRED, to understand which configuration works better. As shown in the left of Figure 9, our model performs the best when the kernel size is 3 in terms of all evaluation metrics. Meanwhile, the right part of Figure 9 shows that 3 seems to be a reasonable choice for the number of the MRN block.

C Analysis of Loss Functions

To compare the ASL and BCE loss, we compare their learning curves on DocRED and keep other settings of our model the same, as shown in Figure 10. With the comparisons of 80 epochs, we observe that the ASL loss helps our model converge to better performance than the BCE loss at a faster speed, demonstrating the effectiveness of the asymmetric strategies for positive and negative samples.

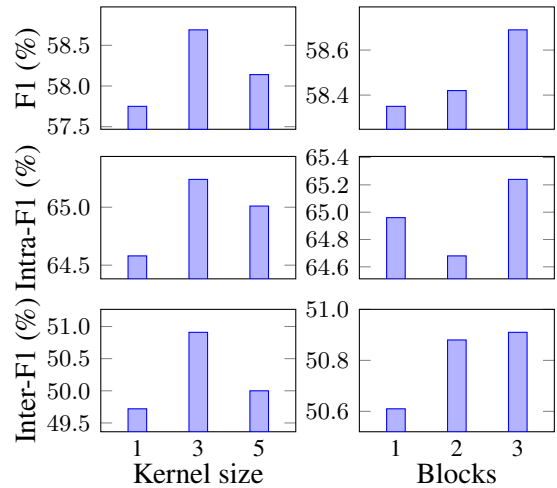


Figure 9: Analysis for the MRN block. The x-axis denotes the kernel size of 2D convolution in the local reasoning module (Left) and the number of used MRN blocks (Right).

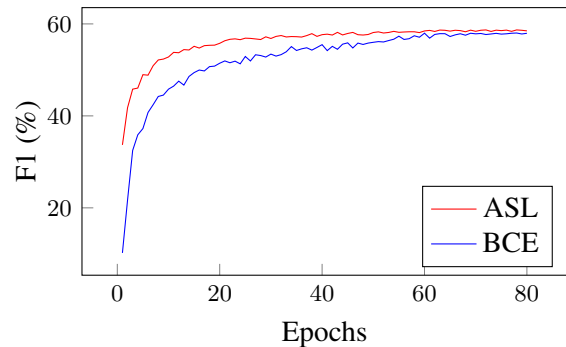


Figure 10: Comparison between the asymmetric loss (ASL) and binary cross-entropy (BCE) loss.

Hyper-parameter	DocRED	CDR	GDA
Batch size	32	16	32
Learning rate	1e-3	1e-3	1e-3
Gradient clipping	1	1	1
Weight decay	1e-5	1e-5	1e-5
Dropout for MRN block	0.5	0.5	0.5
Dropout for co-predictor	0.33	0.5	0.5
Blocks of MRN	3	3	3
Kernel size	3	3	3
Word embedding size	100	200	256
Entity type embedding size	20	20	20
Relative distance embedding size	20	20	20
LSTM hidden size	256	256	320
Positive focusing parameter	3	3	3
Negative focusing parameter	1	0	0
Probability margin	0.05	0.05	0.02

Table 5: Hyper-parameter settings for three datasets.

D Implementation Details

In this section, we provide more details of our experiments. We implemented MRN with PyTorch

and trained it with a NVIDIA Tesla V100 GPU. Hyper-parameter settings for DocRED, CDR and GDA are listed in Table 5.

We adopt AdamW (Loshchilov and Hutter, 2019) as our optimizer. For fair comparisons, we utilize the GloVe embeddings (Pennington et al., 2014) for DocRED, pre-trained PubMed embeddings (Chiu et al., 2016) for CDR, and randomly initialized embeddings for GDA, following Yao et al. (2019) and Christopoulou et al. (2019). Moreover, we apply the uncased BERT-base model as our encoder for comparing with some baselines. We use the initial learning rate $1e-5$ to fine-tune BERT.