Improving Document-level Relation Extraction via Context Guided Mention Integration and Inter-pair Reasoning

Anonymous ACL submission

Abstract

Document-level Relation Extraction (DRE) aims to recognize the relations between two entities. The entity may correspond to multiple mentions that span beyond sentence boundary. Few previous studies have investigated the mention integration, which may be problematic because coreferential mentions do not equally contribute to a specific relation. Moreover, prior efforts mainly focus on reasoning at entity-level rather than capturing the global interactions between entity pairs. In this paper, we propose two novel techniques, **C**ontext **G**uided **M**ention **I**ntegration and **I**nter-pair **R**easoning (CGM2IR), to improve the DRE. Instead of simply applying average pooling, the contexts are utilized to guide the integration of coreferential mentions in a weighted sum manner. Additionally, inter-pair reasoning executes an iterative algorithm on the entity pair graph, so as to model the interdependency of relations. We evaluate our CGM2IR 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[1].

1 Introduction

Relation extraction is a fundamental problem in natural language processing, which aims to identify the semantic relation between a pair of entities mentioned in the text. Recent progress in neural relation extraction has achieved great success (Zeng et al., 2015; Baldini Soares et al., 2019), but these approaches usually focus on binary relations (relations that only involve two entities) within a single sentence. While in practice, a large number of relations in entity pairs span sentence boundaries[2]. Many recent works (Yao et al., 2019; Zhou et al., 2021) pay emphasis on document-level scene that requires a larger context to identify relations, making it a more practical but also more challenging task.

[1] *Britain* 's Prince *Harry* is engaged to his US partner *Meghan Markle* ··· [2] *Harry* spent 10 years in the army and has this year, with his elderly brother *William* , ··· [3] The last major royal wedding took place in 2011, when *Kate Middleton* and Prince *William* were married ...

Relations: *royalty_of(Harry,Britain), sibling_of(William, Harry), spouse_of(kate,William), royalty_of(kate,Britain)* ...

Figure 1: An example of DRE. Note that mentions of the same entity are marked with identical color.

Document-level Relation Extraction (DRE) poses unique challenges compared to its sentence-level counterpart. First, it is more complex to model a document with rich entity structure for relation extraction. The entities engaged in a relation may appear in different sentences, and some entities are repeated with the same phrases or aliases, the occurrences of which are often named entity mentions. For example, as shown in Figure 1, *Britain* and *Kate* appear in the first and third sentences, respectively. *Harry* and *William* also appear more than once in this example. We are therefore confronted to deal with cross-sentence dependencies and synthesizing the information of multiple mentions, in contrast to two entities in one sentence. Second, there are intrinsic interactions among relational facts. The identification of relations between two entities requires reasoning beyond the contextual features. Specifically, in Figure 1, we can determine that the *royalty_of* relation exists between *William* and *Britain* from the context word *Prince*. *Kate* is also a member of the royal family, as she is married to *William*. Logical reasoning plays a dominant role when extract the fact $\langle Kate; royalty\_of; Britain \rangle$.

Many previous works have tried to fulfill DRE

[1] Our code is available at https://anonymous.4open.science/r/CGM2IR-F582.

[2] According to Yao et al. (2019), at least 40.7% of relations can only be identified from multiple sentences.

and tackle the above challenges. In order to exploit the document structure and capture cross-sentence dependencies, most current approaches construct a delicately designed document graph with syntactic structures (coreference, dependency, etc.) (Sahu et al., 2019), heuristics rules[3], or structured attention (Nan et al., 2020). The constructed graphs bridge entities that spread far apart in the document. Besides, as Transformers for NLP can be considered as a graph neural network with multi-head attention as the neighbourhood aggregation function. It implicitly models long-distance dependencies. There are also some works (Xu et al., 2021a; Zhou et al., 2021) that attempt to use Pre-trained Models (PTMs) directly for DRE without involving graph structure. Afterwards, researchers simply apply average (max) pooling to the representation of coreferential entity mentions. Unfortunately, this is obviously not in accordance with intuition and fact. All mentions are equally treated, ignoring the corresponding mention pair contexts for a specific relation.

In this paper, instead of simply synthesizing multiple coreferential mentions, we propose a novel context guided attention mechanism for mention integration. Similarly to Zhou et al. (2021), after encoding through PTMs, we directly get the contexts for each entity pair from the attention heads. Then, the contexts are guided as query to obtain the weights of mentions through cross-attention. This process makes the representation of an entity change dynamically according to the entity pair in which it is located.

In light of the necessity of reasoning, message passing algorithms on graph are employed to update the entity representations accordingly. Thus, it conducts reasoning in an implicit way (Christopoulou et al., 2019). Otherwise, a special reasoning network is designed for relation inference (Zeng et al., 2020; Li et al., 2021). Despite their great success, these methods mainly focus on entity-level or contextual information propagation rather than entity pair interactions, ignoring the global interdependency among multiple relational facts.

In this paper, we propose a novel inter-pair reasoning approach to achieve this purpose. The head and tail entity representations obtained by context guided integration are merged with their contextual information to get the representations of multiple entity pairs. Then, the entity pair representations are formed as the nodes of **G**raph **N**eural **N**etworks (GNNs). The inter-pair interactions are captured through an iterative algorithm over entity pairs, so as to complete reasoning.

By combining the proposed two techniques, we propose a simple yet effective document level relation extraction model, dubbed **CGM2IR** (**C**ontext **G**uided **M**ention **I**ntegration and **I**nter-pair **R**easoning), to fully utilize the power of PTMs. To demonstrate the effectiveness of the proposed approach, we conduct comprehensive experiments on three widely used document level relation extraction datasets. The experimental results reveal that our CGM2IR model significantly outperforms the state-of-the-art methods. Our contributions can be summarized as follows:

- We propose a context guided attention mechanism to dynamically merge mentions that refer to the same entity in a weighted sum manner. Our approach innovatively uses contextual information to guide the entity representation.
- We propose an inter-pair reasoning approach to model interactions among entity pairs rather than entities. Reasoning based on entity pairs is more rational and consistent with the human way of intelligence and learning.
- We conduct experiments on three public DRE datasets. Experimental results demonstrate the effectiveness of our CGM2IR model that achieves the new state-of-the-art performance.

2 Related Work

Relation extraction, also known as relational facts extraction, plays an essential role in a variety of applications in Natural Language Processing (NLP), especially for the automatic construction of Knowledge Graph (KG). Early researchers mainly concentrate on the sentence-level task, i.e. predicting the relations between two entities within a sentence. Many approaches (Zeng et al., 2014; Cai et al., 2016) have been proposed to effectively fulfill **S**entence-level **R**elation **E**xtraction (SRE), especially the pre-training-then-fine-tuning paradigm of PTMs (Zheng et al., 2021). SRE faces an inevitable restriction in practice, where many relation facts can only be extracted from multiple sentences. Re-

[3]For example, EoG (Christopoulou et al., 2019) builds a graph network with sentences, entities, and entity mentions as nodes and edges connected between different nodes according to heuristical rules.

cently, researchers gradually push SRE forward to DRE (Yao et al., 2019).

In DRE, an entity may correspond to multiple mentions, which are scattered in different sentences. We need to classify the relations of multiple entity pairs all at once, which usually requires complex reasoning skills and inter-sentential information. DRE can be cast as a problem with multiple entity pairs to classify and multiple labels to assign (Zhou et al., 2021). To fulfill this task, most current approaches (Christopoulou et al., 2019) adopt appropriate models to first learn the contextual representation of an input document and encode the tokens in it. Then the representation of entity pairs is obtained by different strategies. Finally, a sigmoid classifier is used for multi-label classification.

There are two types of mainstream for document representation. On the one hand, researchers construct a delicately designed document graph. Quirk and Poon (2017) attempted a first step toward constructing document-level graph that augments conventional intra-sentential dependencies with new dependencies introduced by adjacent sentences and discourse relations. Following this, Christopoulou et al. (2019) built a document graph with heterogeneous types of nodes and edges. Nan et al. (2020) proposed a latent structure induction to induce the dependency tree in the document dynamically. Wang et al. (2020), Zeng et al. (2020), Li et al. (2020), Zhang et al. (2020) integrated similar structural dependencies to model documents. Afterwards, graph based algorithm was employed to pass messages and conduct reasoning in an implicit way (Christopoulou et al., 2019). Otherwise, a special reasoning network was designed for relation inference (Zeng et al., 2020; Xu et al., 2021c; Li et al., 2021; Xu et al., 2021b; Zeng et al., 2021). On the other hand, as Transformers for NLP can be considered as a graph neural network with multi-head attention as the neighbourhood aggregation function, it implicitly models long-distance dependencies. There are also some works (Wang et al., 2019; Ye et al., 2020) that attempt to use PTMs directly for DRE without involving graph structure. Xu et al. (2021a) incorporated entity structure dependencies within Transformers encoding part and throughout the overall system. Zhou et al. (2021) proposed an adaptive-thresholding loss and a localized context pooling to improve the performance. Transformer-based approaches implicitly integrate reasoning into the encoding process. These methods are simple but very effective, and have yielded the state-of-the-art performance.

Among the various amounts of prior works, Zhou et al. (2021) and Zhang et al. (2021) are the most two relevant to our approach. Zhou et al. (2021) also considered context to enhance the entity representation. Zhang et al. (2021) captured global interdependency among relation facts at entity pair level. However, the differences are substantial. First and foremost, these two approaches both equally treated all the mentions. In contrast, we use a context guided intra-pair attention mechanism to weigh the mentions. Moreover, we adopt a GNN that forms entity pairs as nodes to learn the inter-pair interactions.

3 Methodology

In this section, we describe the proposed model CGM2IR that incorporates context guided mention integration and inter-pair reasoning to improve DRE. As illustrated in figure 2, CGM2IR mainly consists of four parts, namely (i) the document encoding; (ii) the context guided mention integration; (iii) the inter-pair reasoning; (iv) and the final classification layer.

3.1 Document Encoding Module

To model the semantics of input document better, CGM2IR adopts BERT (Devlin et al., 2019) as the document encoder, which has recently been proven surprisingly effective by presenting state-of-the-art results in various NLP tasks.

Given a document D as input, it is comprised of l tokens $x = \{t_i\}_1^l$ and a set of annotated entities $e_i = \{m_j\}_1^t$ where entity e_i may have multiple mentions that scatter across the document. Borrowing the idea of entity marker (Baldini Soares et al., 2019), we first insert a special marker "*" at the start and end of mentions to mark the mention's span by the entity's annotation. Then, the document encoder is responsible to map each token and mention markers to a sequence of contextualized embedding representations $\boldsymbol{H} = \{\boldsymbol{h_1}, \boldsymbol{h_2}, \cdots, \boldsymbol{h_n}\}$.

$$\boldsymbol{H} = PTMs(\{x_1, x_2, \cdots, x_n\}) \quad (1)$$

where n is the length of tokens with all markers. For each mention, we take the embedding of start marker as the mention embeddings. Limited by the input length of BERT, we use a dynamic window (Zhou et al., 2021) to sequentially encode the whole documents when $n > 512$.

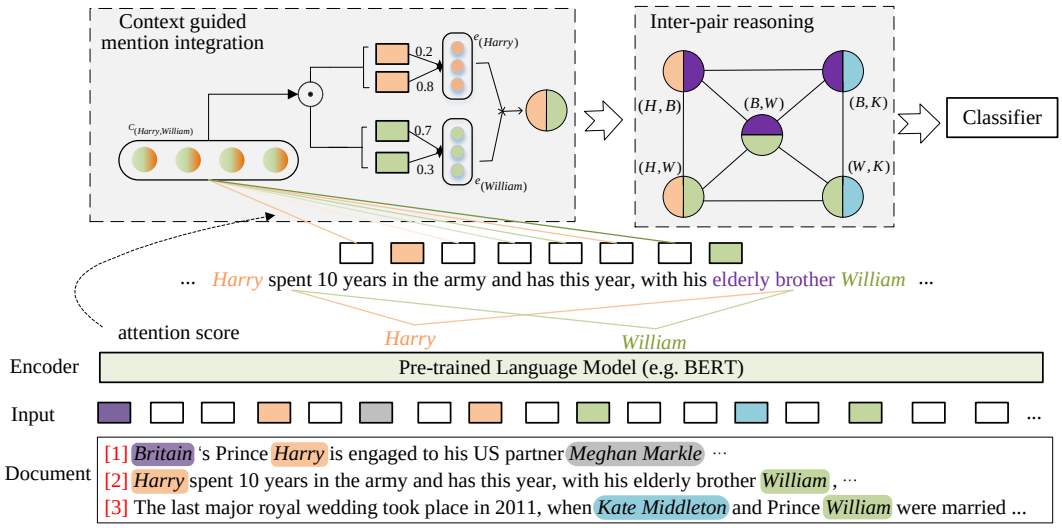

Figure 2: The overall architecture of CGM2IR. First, the input document is viewed as a long sequence of words, which are subsequently encoded through BERT. Then, the context guided mention integration module dynamically generates the head and tail entity embeddings for each entity pair. Next, we construct a homogeneous entity pair graph and use GNNs to model the inter-pair interaction. Finally, the classifier predicts relations of all the entity pairs in a parallel way.

3.2 Context Guided Mention Integration Module

As argued in Section 1, an entity may be mentioned under the same phrase or alias in multiple sentences throughout the document. To obtain entity-level representation, previous works usually synthesize the embeddings of all mentions of an entity. These methods equally treat each mention and only generate one global embedding for an entity. Then, the entity embedding is used in the relation classification of all entity pairs.

Unfortunately, it is obvious that some mentions may not be relevant to the relation when categorizing a particular entity pair. Therefore, we propose a context guided attention mechanism that can generate fine-grained entity representations for each pair. Different from the previous approaches, our motivation is to first get the entity-aware context through the average of mention attention matrices. Then, the contexts involved in both head and tail entities are located to steer the model for mention integration. Following Zhou et al. (2021), we explicitly use the token-level attention score $\boldsymbol{A}$ in the last encoder block of BERT to compute the pair-specific context embedding $\boldsymbol{c}_{h,t}$ for entity pair (e_h, e_t) as follows:

$$\begin{aligned} \boldsymbol{c}_{(h,t)} &= \boldsymbol{H}\boldsymbol{a}_{(h,t)} \\ \boldsymbol{a}_{(h,t)} &= \frac{\boldsymbol{A}_h \cdot \boldsymbol{A}_t}{\mathbf{1}^\top(\boldsymbol{A}_h \cdot \boldsymbol{A}_t)} \end{aligned} \tag{2}$$

where $\boldsymbol{A}_h = avg_{m_i \in e_h}(A_{m_i})$, A_{m_i} is the attention matrix for i-th mention of head entity e_h to all tokens in the document. A similar operation yields $\boldsymbol{A}_t$ for the tail entity. Since the transformer-based PTMs have learned token-level dependencies well by training in a large-scale corpus, we attend all the tokens that are important to both entities in pair (e_h, e_t) by multiplying their entity-level attentions score with a normalization.

After obtaining the contextual features of entity pairs in the first step, we use them as queries and perform cross-attention to pool the entity representations related to the entity pair from the mention embeddings of head or tail entity. Specifically, given an entity pair (e_h, e_t) and a sequence of mention embeddings $\boldsymbol{h}_{m_1}, \boldsymbol{h}_{m_2}, \cdots, \boldsymbol{h}_{m_p}$ of the head or tail entity, where $\boldsymbol{h}_{m_i} \in R^d$, p is the number of mentions. Guided by context feature $\boldsymbol{c}_{(h,t)} \in R^d$ for this pair, the head entity $\boldsymbol{e}^h_{(h,t)}$ is computed as follows:

$$\begin{aligned} \boldsymbol{e}^h_{(h,t)} &= \sum_{i=1}^{p} \boldsymbol{\alpha}^i_{(h,t)} \boldsymbol{h}_{\boldsymbol{m_i}} \\ \boldsymbol{a}^i_{(h,t)} &= \frac{\boldsymbol{W}_Q \boldsymbol{c}^\top_{(h,t)} \boldsymbol{W}_K \boldsymbol{h}_{\boldsymbol{m_i}}}{\sqrt{d}} \\ \boldsymbol{\alpha}^i_{(h,t)} &= \frac{\exp\left(\boldsymbol{a}^i_{(h,t)}\right)}{\sum_{j=1}^{p} \exp\left(\boldsymbol{a}^j_{(h,t)}\right)} \end{aligned} \tag{3}$$

where $\boldsymbol{W}_Q \in \mathbb{R}^{d\times d}$, $\boldsymbol{W}_K \in \mathbb{R}^{d\times d}$ denotes the query and key transformation matrixes, d is the dimension of hidden states. In a similar way, we

can obtain the representation of the tail entity $\boldsymbol{e}^t_{(h,t)}$. We can observe that the representation of each entity is not fixed. It is guided by the trade-off between the context and the entity pair in which it is located. The head entity and the tail entity are combined to dynamically determine the respective representations.

3.3 Inter-pair Reasoning Module

To model interactions among entity pairs in a document, we construct a homogeneous entity pair graph and use GNNs to perform reasoning.

For each document D with m entities, we formulate a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $m \times (m-1)$ entity pairs form the nodes of the graph. Each node representation is computed by the following steps: Given the embeddings $(\boldsymbol{e}^h_{(h,t)}, \boldsymbol{e}^t_{(h,t)})$ of an entity pair (e_h, e_t) and its context features $\boldsymbol{c}_{(h,t)}$, we first combine the entity embeddings with their context embedding, and then map them to hidden representations $\boldsymbol{z}^h_{(h,t)}$ and $\boldsymbol{z}^t_{(h,t)}$ respectively with a feedforward neural network. Finally, the entity pair embedding, $\boldsymbol{p}_{(h,t)}$, is calculated through a group bilinear[4] function as follows:

$$\begin{aligned}
\boldsymbol{z}^h_{(h,t)} &= tanh(W_h \boldsymbol{e}^h_{(h,t)} + W_{c_1} \boldsymbol{c}_{(h,t)}) \\
\boldsymbol{z}^t_{(h,t)} &= tanh(W_t \boldsymbol{e}^t_{(h,t)} + W_{c_2} \boldsymbol{c}_{(h,t)}) \\
\boldsymbol{p}_{(h,t)} &= \sigma(\sum_{i=1}^{k} \boldsymbol{z}^{h^i\top}_{(h,t)} W_p^i \boldsymbol{z}^{t^i}_{(h,t)})
\end{aligned} \tag{4}$$

where $W_h \in \mathbb{R}^{d\times d}$, $W_t \in \mathbb{R}^{d\times d}$, $W_{c_1} \in \mathbb{R}^{d\times d}$, $W_{c_2} \in \mathbb{R}^{d\times d}$ and $W_p^i \in \mathbb{R}^{d/k\times d/k}$ are learnable parameters. Furthermore, we concatenate the entity pair embedding with coreference embedding for head and tail entities to get the initial node representations following Yao et al. (2019):

$$\boldsymbol{P}^0_{(h,t)} = [\boldsymbol{p}^{\boldsymbol{h}}; \boldsymbol{p}_{(h,t)}; \boldsymbol{p}^{\boldsymbol{t}}] \tag{5}$$

In contrast to the fully-connected case, we link each node to the nodes that have overlapping entities with it, since the clues for logical reasoning are usually passed on the chain of entities as it is approved in Xu et al. (2021b) and Zeng et al. (2021). After the graph is constructed, We use GNNs to learn the inter-pair interactions. In each layer l, The GNNs selectively aggregate all entity pair embeddings passed from neighbors through an attention mechanism to update its representation in the next layer $l+1$. Formally, we have:

$$\begin{aligned}
\boldsymbol{P}^{l+1}_u &= FFN(\boldsymbol{W}_r \sum_{\boldsymbol{v}\in\mathcal{N}_{(u)}} \alpha_{(u,v)} \boldsymbol{P}^l_u) \\
\alpha_{(u,v)} &= \frac{\exp[\boldsymbol{Q}\boldsymbol{P}^l_v (\boldsymbol{K}\boldsymbol{P}^l_u)^\top]}{\sum_{\boldsymbol{v'}\in\mathcal{N}_{(u)}} \exp[\boldsymbol{Q}\boldsymbol{P}^l_{v'} (\boldsymbol{K}\boldsymbol{P}^l_u)^\top]}
\end{aligned} \tag{6}$$

where $\boldsymbol{W}_r \in \mathbb{R}^{d\times d}$, $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$, $\boldsymbol{K} \in \mathbb{R}^{d\times d}$ are learnable weight matrices, $FFN(\Delta)$ denotes a feed-forward network, $\mathcal{N}_{(u)}$ is the set of neighbor nodes to the vertex u. In addition, we employ residual connection between two layers and perform layer normalization.

3.4 Classification Module

To determine the semantic relations for an entity pair (e_h, e_t), we first concatenate two pair-specific entity representations and the corresponding final entity pair representation.

$$\boldsymbol{r}_{(h,t)} = [\boldsymbol{e}^h_{(h,t)}; \boldsymbol{e}^t_{(h,t)}; \boldsymbol{P}_{(h,t)}] \tag{7}$$

Then, we use a feed-forward neural network to calculate the probability for each relation:

$$\boldsymbol{P}(r|e_h, e_t) = sigmoid(\boldsymbol{W}_b \sigma(\boldsymbol{W}_a \boldsymbol{r}_{(h,t)} + \boldsymbol{b}_a) + \boldsymbol{b}_b) \tag{8}$$

where $\boldsymbol{W}_a \in \mathbb{R}^{3d\times d}$, $\boldsymbol{W}_b \in \mathbb{R}^{d\times r}$, $\boldsymbol{b}_a$, $\boldsymbol{b}_b$ are learnable parameters, σ is an elementwise activation function (e.g., tanh).

To address the multi-label and sample imbalance problem more effectively, we adopt an adaptive-thresholding loss (Zhou et al., 2021) as the classification loss to train our model in an end-to-end way. Specifically, it introduces an additional threshold relation category *TH*, and optimizes the loss by increasing the logits of the positive relations $\mathcal{P}_T$ higher than the *TH* relation and decreasing the logits of the negative relations $\mathcal{N}_T$ lower than the *TH* relation.

$$\begin{aligned}
\mathcal{L} = &- \sum_{r\in\mathcal{P}_T} log(\frac{\exp(logit_r))}{\sum_{r'\in\mathcal{P}_T\cup\{TH\}} \exp(logit_r)}) \\
&- log(\frac{\exp(logit_{TH}))}{\sum_{r'\in\mathcal{N}_T\cup\{TH\}} \exp(logit_r)})
\end{aligned} \tag{9}$$

where $logit$ is the output in the last layer before Sigmoid function.

4 Experiments

4.1 Datasets

We evaluate the effectiveness of our CGM2IR model on three public DRE datasets: DoRED,

[4]Group bilinear (Zheng et al., 2019) splits the embedding dimensions into k equal-sized groups and applies bilinear within the groups.

Statistics / Dataset	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
Avg. # Ment. per Ent.	1.4	2.7	3.3
Avg. # Ents per Doc.	19.5	7.6	2.4
Avg. # Facts. per Doc.	12.6	2.1	1.5

Table 1: Statistics of the datasets.

CDR, and GDA. The dataset statistics are shown in Table 1.

DocRED is a large-scale human-annotated dataset for document-level RE proposed by (Yao et al., 2019). It contains 97 types of relations and 5,053 annotated documents in total which are constructed from Wikipedia and Wikidata. Documents in DocRED contain about 12.6 positive relational facts on average, which is several times that of the common sentence-level RE dataset. **CDR** (Chemical-Disease Reactions) (Li et al., 2016) and **GDA** (Gene-Disease Associations) (Wu et al., 2019) are two widely-used DRE datasets in the biomedical domain. They both contain only one type of positive relation, *Chemical-Induced-Disease* between chemical and disease entities and *Gene-Induced-Disease* between gene and disease entities respectively. For a fair comparison, We follow the standard split of the three datasets as Zeng et al. (2020) and Zhou et al. (2021).

4.2 Experiment Settings and Evaluation Metrics

In our CGM2IR implementation, we use cased BERT-base (Devlin et al., 2019) or RoBERTa-large (Liu et al., 2019) the encoder on DocRED and cased SciBERT-base (Beltagy et al., 2019) on CDR and GDA. AdamW (Loshchilov and Hutter, 2019) is used to optimize the neural networks with a linear warmup and cosine decay learning rate schedule. We set the initial learning rate for all encoder modules to $2e^{-5}$, the initial learning rate for other modules to $1e^{-4}$, the embedding dimension, and the hidden dimension to 768. The GNNs have 3 layers and the hidden size of node embedding is 768. All hyper-parameters are tuned based on the development set. Other parameters in the network are all obtained by random orthogonal initialization (Saxe et al., 2014) and updated during training. All the experiments are trained with an NVIDIA RTX 3090 GPU.

Following Yao et al. (2019) and previous works, we use the micro F_1 and micro Ign F_1 as the evaluation metrics for DocRED. Ign F_1 denotes the result after excluding the common relational facts that appear in both training set and development/test sets. For CDR and GDA, in addition to using micro F_1, we also report the Intra F_1 and Inter F_1 metrics to evaluate the model's performance on intra-sentential relations and inter-sentential relations on the dev set, since they strictly annotate these two types of facts but DocRED does not. In our experiments, a triplet is taken as correct when the two corresponding entities and the relation type are all correct and we exclude all triplets with relation of "None".

4.3 Results on DocRED

We conduct comprehensive and comparable experiments on DocRED dataset. The results are shown in Table 2.

We compare our CGM2IR model with lots of methods from two categories. The first one is graph-based methods, including LSR (Nan et al., 2020), GEDA (Li et al., 2020), GCGCN-BERT (Zhou et al., 2020), GLRE (Wang et al., 2020), GAIN (Zeng et al., 2020), HeterGSAN (Xu et al., 2021c), SIRE (Zeng et al., 2021) and DRE (Xu et al., 2021b). The second one is non-graph-based methods including BERT (Wang et al., 2019), HIN-BERT (Tang et al., 2020), CorefBERT (Ye et al., 2020), SSAN (Xu et al., 2021a), ATLOP (Zhou et al., 2021), MRN (Li et al., 2021) and DocuNet (Zhang et al., 2021). The baselines we selected all use BERT as their encoder.

As shown in Table 2, we observe that CGM2IR outperforms all baseline methods on both development and test sets. Compared with the models in these two categories, both F_1 and Ign F_1 of our model are significantly improved. Among the various amounts of baselines, ATLOP (Zhou et al., 2021) and DocuNet (Zhang et al., 2021) are the most two relevant to our approach. Compared to ATLOP-BERT$_{base}$, the performance of CGM2IR-BERT$_{base}$ improves roughly about 0.8% for Ign F_1 and 0.92% for F_1. CGM2IR-BERT$_{base}$ also brings about 0.2% ign F_1 enhancement compared to DocuNet-BERT$_{base}$, which verifies the effectiveness of our proposed method. Furthermore, CGM2IR-RoBERTa$_{large}$ obtains better results than baselines with BERT-large or RoBERTa-large as well. For example, CGM2IR-RoBERTa$_{large}$ achieves 0.71% Ign F_1/0.77% F_1 gain compared

Model	Dev		Test	
	Ign F_1	F_1	Ign F_1	F_1
LSR-BERT$_{base}$ (Nan et al., 2020)	52.43	59.00	56.97	59.05
GEDA-BERT$_{base}$ (Li et al., 2020)	54.52	56.16	53.71	55.74
GCGCN-BERT$_{base}$ (Zhou et al., 2020)	55.43	57.35	54.53	56.67
GLRE-BERT$_{base}$ (Wang et al., 2020)	-	-	55.40	57.40
HeterGSAN-BERT$_{base}$ (Xu et al., 2021c)	58.13	60.18	57.12	59.45
GAIN-BERT$_{base}$ (Zeng et al., 2020)	59.14	61.22	59.00	61.24
DRE-BERT$_{base}$ (Xu et al., 2021b)	59.33	61.39	59.15	61.37
SIRE-BERT$_{base}$ (Zeng et al., 2021)	59.82	61.60	60.18	62.05
BERT$_{base}$ (Wang et al., 2019)	-	54.16	-	53.20
HIN-BERT$_{base}$ (Tang et al., 2020)	54.29	56.31	53.70	55.60
CorefBERT$_{base}$ (Ye et al., 2020)	55.32	57.51	54.54	56.96
SSAN-BERT$_{base}$ (Xu et al., 2021a)	57.03	59.19	55.84	58.16
ATLOP-BERT$_{base}$ (Zhou et al., 2021)	59.22	61.09	59.31	61.30
MRN-BERT$_{base}$ (Li et al., 2021)	59.74	61.61	59.52	61.74
DocuNet-BERT$_{base}$(Zhang et al., 2021)	59.86	61.83	59.93	61.86
CGM2IR-BERT$_{base}$	**60.02**	**62.01**	**60.24**	**62.06**
BERT$_{large}$ (Wang et al., 2019)	56.67	58.83	56.47	58.69
CorefRoBERTa$_{large}$ (Ye et al., 2020)	57.84	59.93	57.68	59.91
SSAN-RoBERTa$_{large}$ (Xu et al., 2021a)	60.25	62.08	59.47	61.42
GAIN-BERT$_{large}$ (Zeng et al., 2020)	60.87	63.09	60.31	62.76
ATLOP-RoBERTa$_{large}$ (Zhou et al., 2021)	61.32	63.18	61.39	63.40
CGM2IR-RoBERTa$_{large}$	**62.03**	**63.95**	**61.96**	**63.89**

Table 2: Results on the development and test set of DocRED. We separate graph-based and non-graph-based methods into two groups. The results of baselines are from their related papers.

Model	F_1	intra-F_1	inter-F_1
• **CDR** Dataset			
EoG	63.6	68.2	50.9
LSR	64.8	68.9	53.1
DHG-BERT$_{base}$	65.9	70.1	54.6
MRN	65.9	70.4	54.2
ATLOP-SciBERT$_{base}$	69.2	74.2	52.6
CGM2IR-SciBERT$_{base}$	**73.8**	**79.2**	**55.1**
• **GDA** Dataset			
EoG	81.5	85.2	50.0
LSR	82.2	85.4	51.1
MRN	82.9	86.1	53.5
DHG-BERT$_{base}$	83.1	85.6	58.8
ATLOP-SciBERT$_{base}$	83.9	87.3	52.9
CGM2IR-SciBERT$_{base}$	**84.7**	**88.3**	**59.0**

Table 3: Results on CDR and GDA datasets.

to ATLOP-RoBERTa$_{large}$ on the development set. In general, these results demonstrate both the effectiveness of context guided mention integration and the usefulness of inter-pair reasoning.

4.4 Results on CDR and GDA

Table 3 depicts the comparisons with state-of-the-art models on CDR and GDA. We compare our CGM2IR model with five baselines, including EoG (Christopoulou et al., 2019), DHG (Zhang et al., 2020), LSR (Nan et al., 2020), MRN (Li et al., 2021), ATLOP (Zhou et al., 2021). Our model adopts SciBERT$_{base}$ for its superiority when dealing with biomedical domain texts.

It can be observed that CGM2IR achieves the new state-of-the-art F_1 score on these two datasets in the biomedical domain. On CDR test set, CGM2IR obtains +4.6 F_1 gain, which significantly outperforms all other approaches. On GDA test set, similar improvements can also be observed. These results demonstrate the effectiveness and generality of our approach.

Model	Ign F_1	F_1
CGM2IR-BERT$_{base}$	**60.02**	**62.01**
w/o mention integration module	59.64	61.63
w/o inter-pair reasoning module	59.87	61.74
w/o both module	59.12	60.89

Table 4: Ablation study of CGM2IR on the development set of DocRED, where "w/o" indicates without.

4.5 Ablation Study

We also conduct a thorough ablation study as shown in Table 4 to study the contribution of two key modules: context guided mention integration module and inter-pair reasoning module. From Table 4, we can observe that:

(1) When the context guided mention integration module is discarded and replaced with the logsum-exp pooling layer, the performance of our model on the DocRED dev set drops by 0.38% in both F_1 and

Ign F_1 score. Similarly, removal of the inter-pair reasoning module results in a 0.27% drop in F_1 and 0.14% in Ign F_1. This phenomenon indicates the effectiveness of context guided mention integration module and inter-pair reasoning module.

(2) Removal of both modules leads to a more considerable decrease. The F_1 score decreases from 62.01% to 60.89% and the Ign F_1 score decreases from 60.02% to 59.12%. This study demonstrates that all components work together in synergy with the final relation classification.

4.6 Intra- and Inter-sentence Triplet Extraction

To further evaluate the performance, we report the results of intra- and inter-sentence relation extraction on CDR and GDA, since they explicitly annotate these two types of facts. The experimental results are listed in Table 3, from which we can find that CGM2IR outperforms the current best models on these two datasets in regard to both intra- and inter-F_1. For example, Our model obtains +5.0 intra-F_1/+2.5 inter-F_1 and +1.0 intra-F_1/+6.1 inter-F_1 gain compared with ATLOP on the test set of these two datasets. The improvements indicate that our model can effectively capture the complex interactions among entity pairs across the document. The intra-sentence relations contained in local text can be well considered, as well as the long-distance dependent inter-sentence relations.

4.7 Effect Analysis for Context Guided Cross-Attention

To assess the effectiveness of context guided cross-attention in modeling entity representations, we compare five different strategies for generating entity representations including global mean pooling, global max pooling, global attention pooling, global logsumexp pooling, and our context guided cross-attention. For simplicity, after encoding the document, we directly concat the representations of the head entity and the tail entity then send them to the final classifier. The results on the development set of DocRED are illustrated in Table 5, from which we can observe that the context guided cross-attention is absolutely superior to the global strategies. This result indicates that context guided cross-attention is reasonable and effective, which drives the head and tail entities together to dynamically determine their respective representations.

Method	Ign F_1	F_1
global mean pooling	57.24	58.23
global max pooling	57.41	58.54
global attention pooling	58.17	59.00
global logsumexp pooling	58.23	59.12
context guided cross-attention	**59.34**	**60.76**

Table 5: Results of different strategies for generating entity representations on DocRED.

Model	Infer F_1	P	R
BERT-RE$^*_{base}$	39.62	34.12	47.23
GAIN-GloVe†	40.82	32.76	54.14
RoBERTa-RE$^*_{base}$	41.78	37.97	46.45
SIRE-GloVe†	42.72	34.83	55.22
GAIN-BERT$^*_{base}$	46.89	38.71	59.45
CGM2IR-BERT$_{base}$	**48.04**	**39.54**	**61.21**

Table 6: Infer-F_1 results on dev set of DocRED. Results with * are reported in Zeng et al. (2020), † are reported in Zeng et al. (2021).

4.8 Effect Analysis for Inter-pair Reasoning

In addition, we evaluate the reasoning ability of our model on the development set of DocRED in Table 6. Following Zeng et al. (2021), we use infer-F_1 as a metric that only considers instances of the two-hop positive relations in the development set of DocRED. More specifically, we only evaluate the golden relational facts r_1, r_2 and r_3 when there exists $e_h \xrightarrow{r_1} e_o \xrightarrow{r_2} e_t$ and $e_h \xrightarrow{r_3} e_t$.

As illustrated in Table 6, CGM2IR outperforms all the baselines in infer-F_1. Specifically, CGM2IR-BERT$_{base}$ improves roughly about 1.15% for infer-F_1 score compared with GAIN-BERT$_{base}$. This reveals that the inter-pair reasoning module plays an important role in capturing intrinsic clues and performing logic reasoning on entities chains.

5 Conclusion

In this paper, we propose CGM2IR that incorporates context guided mention integration and inter-pair reasoning to improve DRE. Instead of simply synthesizing multiple coreferential mentions at once, CGM2IR dynamically generates fine-grained entity representations for each entity pair. Moreover, we construct a homogeneous entity pair graph and employ GNNs to capture intrinsic clues and perform reasoning among entity pairs. Experimental results on three widely used DRE datasets demonstrate that our CGM2IR model is effective and outperforms previous state-of-the-art models.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765, Berlin, Germany. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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 28th International Conference on Computational Linguistics*, pages 1551–1560, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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, Online. Association for Computational Linguistics.

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, Valencia, Spain. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

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. Springer.

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, Online. Association for Computational Linguistics.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Yang Wang. 2019. Fine-tune bert for docred with two-step process. *CoRR*, abs/1909.11898.

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 of *Lecture Notes in Computer Science*, pages 272–284. Springer.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021a. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14149–14157. AAAI Press.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. Discriminative reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1653–1663, Online. Association for Computational Linguistics.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021c. Document-level relation extraction with reconstruction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14167–14175. AAAI Press.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 524–534, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3999–4006. ijcai.org.

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 28th International Conference on Computational Linguistics*, pages 1630–1641, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. 2019. Learning deep bilinear transformation for fine-grained image representation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4279–4288.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. PRGC: Potential relation and global correspondence based joint relational triple extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6225–6235, Online. Association for Computational Linguistics.

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 28th International Conference on Computational Linguistics*, pages 5259–5270, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14612–14620. AAAI Press.