Document-Level Relation Extraction With Context Guided Mention Integration and Inter-Pair Reasoning

Daojian Zeng, Chao Zhao, Chao Jiang, Jianling Zhu, and Jianhua Dai

Abstract—Document-level Relation Extraction (DRE) aims to recognize the relations between two entities. The entity may correspond to multiple mentions that span beyond the 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 the entity pairs. In this article, we propose two novel techniques, Context Guided Mention Integration and Inter-pair Reasoning (CGM2IR), to improve 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 interdependence 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 strong baselines and achieves remarkable effectiveness.

Index Terms—Document-level relation extraction, inter-pair reasoning, context guided cross-attention, entity pair graph, graph neural network.

I. 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 and has a wide range of practical applications [1], [2]. Recent progress in neural relation extraction has achieved great success [3], [4], 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 boundaries1. Many recent works [5], [6] pay emphasis on document-level scene that requires a larger context to identify relations, making it a more practical but also more challenging task.

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 Fig. 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 Fig. 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 ⟨Kate: royalty_of: Britain⟩.

Many previous works have tried to fulfill DRE and tackled 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.) [7], heuristics rules2,

1 According to Yao et al. [5], at least 40.7% of relations can only be identified from multiple sentences.
2 For example, EoG [8] builds a graph network with sentences, entities, and entity mentions as nodes and edges connected between different nodes according to heuristical rules.

Fig. 1. Example of DRE. Note that mentions of the same entity are marked with identical color.

Manuscript received 17 July 2022; revised 9 July 2023; accepted 3 September 2023. Date of publication 18 September 2023; date of current version 20 October 2023. This work was supported in part by the Youth Cultivation Project of Research Institute of Languages and Cultures, Hunan Normal University under Grant 2020QNP05, in part by the Hunan Provincial Natural Science Fund of China under Grant 2020J34624, in part by the National Social Science Fund of China under Grant 20&ZD047, and in part by the National Natural Science Fund of China under Grant 62276095. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yun-Nung Chen.

(Corresponding author: Jianhua Dai.)

Daojian Zeng is with the Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410081, China (e-mail: zengdj@hunnu.edu.cn).

Chao Zhao is with the Changsha Dworld AI Tech Company Ltd., Changsha 410081, China (e-mail: zhaochao@mail.chancein.cn).

Chao Jiang, Jianling Zhu, and Jianhua Dai are with the Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410081, China (e-mail: jiangchao@hunnu.edu.cn; zjulj@hunnu.edu.cn; jhdai@hunnu.edu.cn).

Digital Object Identifier 10.1109/TASLP.2023.3316454

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
or structured attention [9]. 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 the multi-head attention as the neighbourhood aggregation function. It implicitly models long-distance dependencies. There are also some works [6], [10] 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 article, instead of simply synthesizing multiple coreferential mentions, we propose a novel context guided attention mechanism for mention integration. Similarly to Zhou et al. [6], 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 [8]. Otherwise, a special reasoning network is designed for relation inference [11], [12]. 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. The global interactions between pairs of entities contain valuable information and patterns that significantly enhance the model’s reasoning about the relations between these entity pairs. Moreover, the inherent dependencies and correlations existing between different entity pairs further amplify the model’s reasoning capabilities. By taking into account the context-specific relationships and interactions among entity pairs, the model can conduct a more precise and comprehensive analysis.

In this article, 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 Graph Neural Networks (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 (Context Guided Mention Integration and Inter-pair Reasoning), 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. Experimental results reveal that our CGM2IR model exhibits a notable performance improvement compared to previous strong baselines. 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 the 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 that our CGM2IR model surpasses previous strong baselines and achieves remarkable performance.

II. 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 [13], [14] have been proposed to effectively fulfill Sentence-level Relation Extraction (SRE), especially the pre-training-then-fine-tuning paradigm of PTMs [15]. SRE faces an inevitable restriction in practice, where many relation facts can only be extracted from multiple sentences. Recently, researchers gradually push SRE forward to DRE [5].

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 [6]. To fulfill this task, most current approaches [8] 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 [16] 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. [8] build a document graph with heterogeneous types of nodes and edges. Nan et al. [9] proposed a latent structure induction to induce the dependency tree in the document dynamically. Wang et al. [17], Zeng et al. [11], Li et al. [18], Zhang et al. [19] integrated similar structural dependencies to model documents. Afterwards, graph based algorithm was employed to pass messages and conduct reasoning in an implicit way [8]. Otherwise, a special reasoning network was designed for relation inference [11], [12], [20], [21], [22]. 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 [23], [24] that attempt to use PTMs directly for DRE without involving graph structure. Xu et al. [10] incorporated
entity structure dependencies within Transformers encoding part and throughout the overall system. Zhou et al. [6] proposed an adaptive-thresholding loss and a localized context pooling to improve the performance. Then, Ru et al. [25] proposed LogiRE, a novel probabilistic model that learns logic rules. LogiRE treats logic rules as latent variables and consists of two modules: a rule generator and a relation extractor. Additionally, Zhang et al. [26] break down the DRE task into two sub-problems: relation detection and argument resolution. This modular approach is designed to reduce noise and enhance the reliability of the learning process. 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. [6] and Zhang et al. [27] are the most two relevant to our approach. Zhou et al. [6] also considered context to enhance the entity representation. Zhang et al. [27] 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.

III. 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 Fig. 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.

A. Document Encoding Module

To model the semantics of input document better, CGM2IR adopts BERT [28] 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 $x = \{t_i\}^l_1$ and a set of annotated entities $e_i = \{m_j\}^t_i$ where entity $e_i$ may have multiple mentions that scatter across the document. Borrowing the idea of entity marker [4], 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 $H = \{h_1, h_2, \ldots, h_n\}$.

$$H = PTMs(\{x_1, x_2, \ldots, x_n\})$$  \hspace{1cm} (1)

where $n$ is the length of tokens with all markers. For each mention, we take the embedding of the start marker as the mention embeddings. This approach allows us to avoid the need for considering the number of tokens in the mention. Limited by the input length of BERT, we use a dynamic window [6] to sequentially encode the whole documents when $n > 512$.

B. Context Guided Mention Integration Module

As argued in Section I, an entity may be mentioned under the same phrase or alias in multiple sentences throughout the document. To obtain the 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. [6], we explicitly use the token-level attention score \( A \) in the last encoder block of BERT to compute the pair-specific context embedding \( c_{h,t} \) for entity pair \((e_h, e_t)\) as follows:

\[
c_{(h,t)} = H a_{(h,t)}
\]

\[
a_{(h,t)} = \frac{A_h \odot A_t}{1^t (A_h \odot A_t)}
\]  

(2)

where \( A_h \) and \( A_t \) means the attentions of the head and tail entities, respectively. Specifically, we first take the attention of the special token "**" at the start of the mention as its attention. Then, we aggregate the attentions for all mentions belonging to head entity \( e_h \) to get entity-level attention \( A_h = \text{avg}_{m_i \in e_h} (A_{m_i}) \), where \( A_{m_i} \) is the attention matrix for \( i \)-th mention of head entity \( e_h \) to all tokens in the document. A similar operation yields \( A_t \) 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 \( h_{m_1}, h_{m_2}, \ldots, h_{m_p} \) of the head or tail entity, where \( h_{m_i} \), \( R^d \), is the embedding of the start marker "**" of the mention \( m_i \). Guided by context feature \( c_{(h,t)} \in R^d \) for this pair, the head entity \( e_{h,t}^h \) is computed as follows:

\[
e_{h,t}^h = \sum_{i=1}^{p} \alpha_{i(t)}^h h_{m_i}
\]

\[
a_{i(t)}^h = \frac{W_Q c_{(h,t)} W_K h_{m_i}}{\sqrt{d}}
\]

\[
\alpha_{i(t)}^h = \frac{\exp\left(\alpha_{i(t)}^h\right)}{\sum_{j=1}^{p} \exp\left(\alpha_{j(t)}^h\right)}
\]  

(3)

where \( W_Q \in \mathbb{R}^{d \times d}, W_K \in \mathbb{R}^{d \times d} \) denotes the query and key transformation matrices, \( d \) is the dimension of hidden states. In a similar way, we can obtain the representation of the tail entity \( e_{h,t}^t \). 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.

### C. 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 \( G = (\mathcal{V}, \mathcal{E}) \), 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 \( (e_h, e_t) \) and its context features \( c_{(h,t)} \), we first compute the entity embeddings with their context embedding, and then map them to hidden representations \( z_{(h,t)}^h \) and \( z_{(h,t)}^t \) respectively with a feedforward neural network. Finally, the entity pair embedding, \( p_{(h,t)} \), is calculated through a group bilinear\(^3\) function as follows:

\[
z_{(h,t)}^h = \tanh(W_h e_{(h,t)}^h + W_c c_{(h,t)})
\]

\[
z_{(h,t)}^t = \tanh(W_t e_{(h,t)}^t + W_c c_{(h,t)})
\]

\[
p_{(h,t)} = \sigma\left( \sum_{i=1}^{k} z_{(h,t)}^{h+} W_p z_{(h,t)}^{i+} \right)
\]  

(4)

where \( W_h \in \mathbb{R}^{d \times d}, W_t \in \mathbb{R}^{d \times d}, W_c \in \mathbb{R}^{d \times d}, W_{c_1} \in \mathbb{R}^{d \times d}, W_{c_2} \in \mathbb{R}^{d \times d} \) and \( W_{p_1} \in \mathbb{R}^{d \times k \times d} \), \( k \) is learnable parameters. Furthermore, \( p^h \) and \( p^t \) are the coreference embeddings of the head entity and tail entity corresponding to the node respectively. The coreference embedding is obtained by mapping entity ids (determined by their order of first appearance in the document) to vectors using an embedding matrix. When we initialize the homogeneous entity pair graph, we use two prior features to help the GNNs for message propagation. On the one hand, we only 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 [21, 22]. On the other hand, following Yao et al. [5], we explicitly concatenate each node with the coreference embedding of the head and tail entities it belongs to, since the representation of each node is obtained by weighted pooling of representations of different mentions of the head and tail entities:

\[
P^{0}_{(h,t)} = [p^h, p_{(h,t)}; p^t]
\]  

(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. [21] and Zeng et al. [22]. 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:

\[
P^{l+1}_u = W_p \left( W_r \sum_{v \in N_u} \alpha_{(u,v)} P^l_u \right) + b_p
\]

\(^3\)Group bilinear [29] splits the embedding dimensions into \( k \) equal-sized groups and applies bilinear within the groups.
\[
\alpha_{(u,v)} = \frac{\exp(QP^u_v(KP^v_u)^\top)}{\sum_{v' \in \mathcal{N}(u)} \exp(QP^v_{v'}(KP^{v'}_v)^\top)}
\]

where \(W_p \in \mathbb{R}^{d \times d}, W_r \in \mathbb{R}^{d \times d}, b_p \in \mathbb{R}^{d \times d}, Q \in \mathbb{R}^{d \times d}, K \in \mathbb{R}^{d \times d}\) are learnable weight matrices. The vertex \(u\) denotes a node in GNNs, which represents an entity pair \((e_h, e_t)\). \(\mathcal{N}(u)\) is the set of neighbor nodes to the vertex \(u\). \(P^u_v\) refers to the embedding representation of the vertex \(u\) of the \(l\)-th layer of GNNs. In addition, we employ residual connection between two layers and perform layer normalization.

### D. 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.

\[
r_{(h,t)} = [e_h^{l}; e_t^{l}; P_{(h,t)}]
\]

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

\[
P(r|e_h, e_t) = \text{sigmoid}(W_a r_{(h,t)} + b_a + b_b) \tag{8}
\]

where \(W_a \in \mathbb{R}^{3d \times d}, W_b \in \mathbb{R}^{d \times r}, b_b, b_b\) are learnable parameters, \(\sigma\) is an element-wise activation function (e.g., tanh).

To address the multi-label and sample imbalance problem more effectively, we adopt an adaptive-thresholding loss \([6]\) as the classification loss to train our model in an end-to-end way. Specifically, it introduces an additional threshold relation category \(T_h\), and optimizes the loss by increasing the logits of the positive relations \(P_T\) higher than the \(T_h\) relation and decreasing the logits of the negative relations \(N_T\) lower than the \(T_h\) relation.

\[
\mathcal{L} = - \sum_{r \in P_T} \log \left( \frac{\exp(\text{logit}_r)}{\sum_{r' \in P_T \cup \{T_h\}} \exp(\text{logit}_{r'})} \right) - \log \left( \frac{\exp(\text{logit}_{T_h})}{\sum_{r' \in N_T \cup \{T_h\}} \exp(\text{logit}_{r'})} \right) \tag{9}
\]

where logit is the output in the last layer before sigmoid function.

### IV. EXPERIMENTS

#### A. Datasets

We evaluate the effectiveness of our CGM2IR model on three public DRE datasets: DocRED, CDR and GDA. The dataset statistics are shown in Table I.

**DocRED** is a large-scale human-annotated dataset for DRE proposed by Yao et al. \([5]\). 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-Induced-Reactions) \([30]\) and **GDA** (Gene-Disease Associations) \([31]\) 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 Zhou et al. \([6]\) and Zeng et al. \([11]\).

#### B. Experiment Settings and Evaluation Metrics

In our CCM2IR implementation, we use cased BERT-base \([28]\) or RoBERTa-large \([34]\) the encoder on DocRED and based SciBERT-base \([35]\) on CDR and GDA. AdamW \([36]\] 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 \([37]\) and updated during training. All the experiments are trained with an NVIDIA RTX 3090 GPU.

Following Yao et al. \([5]\) and previous works, we use the micro \(F_1\) and macro \(F_1\) as the evaluation metrics for DocRED. \(F_1\) 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”.

#### C. Results on DocRED

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

We compare our CCM2IR model with lots of methods from two categories. The first one is graph-based methods, including LSR \([9]\), GEDA \([18]\), GGCN-BERT \([32]\), GLRE \([17]\), GAIN \([11]\), HeterGSAN \([20]\), SIRE \([22]\) and DRE \([21]\). The second one is non-graph-based methods including BERT \([23]\), HIN-BERT \([33]\), CorefBERT \([24]\), SSAN \([10]\), ATLOP \([6]\), MrN \([12]\) and DocuNet \([27]\). The baselines we selected all use BERT as their encoder.
TABLE II
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                  | Dev | Test |
|------------------------|-----|------|
| LS-K-BERT<sub>base</sub> [9] | 52.43 | 79.00 | 56.97 | 59.05 |
| GEDAI-BERT<sub>base</sub> [18] | 54.52 | 56.16 | 53.71 | 55.74 |
| GGGCN-BERT<sub>base</sub> [32] | 55.43 | 57.35 | 54.53 | 56.67 |
| GLRE-BERT<sub>base</sub> [17] | - | - | 55.40 | 57.40 |
| HeterGSAN-BERT<sub>base</sub> [20] | 58.13 | 60.18 | 57.12 | 59.45 |
| DA-BERT<sub>base</sub> [11] | 59.14 | 61.22 | 59.00 | 61.24 |
| DRE-BERT<sub>base</sub> [21] | 59.53 | 61.39 | 59.15 | 61.37 |
| SIRE-BERT<sub>base</sub> [22] | 59.82 | 61.60 | 60.18 | 62.05 |
| BERT<sub>base</sub> [23] | - | 54.16 | - | - |
| HIN-BERT<sub>base</sub> [33] | 54.29 | 56.31 | 53.70 | 55.60 |
| CorefBERT<sub>base</sub> [24] | 55.32 | 57.51 | 54.54 | 56.96 |
| SSAN-BERT<sub>base</sub> [10] | 57.03 | 59.19 | 55.84 | 58.16 |
| ATLOP-BERT<sub>base</sub> [6] | 59.22 | 61.09 | 59.31 | 61.30 |
| ATLOP-LogIRE-BERT<sub>base</sub> [25] | - | - | 59.48 | 61.45 |
| MRN-BERT<sub>base</sub> [12] | 59.74 | 61.61 | 59.52 | 61.74 |
| DocuNet-BERT<sub>base</sub> [27] | 59.85 | 61.89 | 59.93 | 61.86 |
| ATLOP-SciBERT<sub>base</sub> [15] | 60.02 | 62.01 | 60.34 | 62.06 |
| BERT<sub>large</sub> [23] | 56.67 | 58.83 | 56.47 | 58.69 |
| CorefRoBERTA<sub>large</sub> [24] | 57.84 | 59.93 | 57.68 | 59.91 |
| SSAN-RoBERTA<sub>large</sub> [10] | 60.25 | 62.08 | 59.47 | 61.42 |
| GAIN-RoBERTA<sub>large</sub> [11] | 60.87 | 63.09 | 60.31 | 62.76 |
| ATLOP-RoBERTA<sub>large</sub> [6] | 61.32 | 63.18 | 61.39 | 63.40 |
| CGM2IR-RoBERTA<sub>large</sub> [34] | 62.03 | 63.95 | 61.96 | 65.89 |

As shown in Table II, we observe that CGM2IR outperforms all baseline methods on both development and test sets. In the table, bold values represent the maximum values. Compared with the models in these two categories, both F<sub>1</sub> and Ign F<sub>1</sub> of our model are significantly improved. Among the various amounts of baselines, ATLOP [6] and DocuNet [27] are the most two relevant to our approach. Compared to ATLOP-BERT<sub>base</sub>, the performance of CGM2IR-BERT<sub>base</sub> improves roughly about 0.8% for Ign F<sub>1</sub> and 0.92% for F<sub>1</sub>. CGM2IR-BERT<sub>base</sub> also brings about 0.2% ign F<sub>1</sub> enhancement compared to DocuNet-BERT<sub>base</sub>, which verifies the effectiveness of our proposed method. Furthermore, CGM2IR-RoBERTA<sub>large</sub> obtains better results than baselines with BERT-large or RoBERTa-large as well. For example, CGM2IR-RoBERTA<sub>large</sub> achieves 0.71% Ign F<sub>1</sub>/0.77% F<sub>1</sub> gain compared to ATLOP-RoBERTA<sub>large</sub> on the development set. In general, these results demonstrate both the effectiveness of context guided mention integration and the usefulness of inter-pair reasoning.

D. Results on CDR and GDA

Table III depicts the comparisons with state-of-the-art models on CDR and GDA. We compare our CGM2IR model with five baselines, including EoG [8], DHG [19], LSR [9], MRN [12], ATLOP [6]. Our model adopts SciBERT<sub>base</sub> for its superiority when dealing with biomedical domain texts. In the table, bold values represent the maximum values.

It can be observed that CGM2IR achieves the new state-of-the-art F<sub>1</sub> score on these two datasets in the biomedical domain. On CDR test set, CGM2IR obtains +4.6 F<sub>1</sub> 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.

E. Ablation Study

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

1) When the context guided mention integration module is discarded and replaced with the logsumexp pooling layer, the performance of our model on the DocRED dev set drops by 0.38% in both F<sub>1</sub> and Ign F<sub>1</sub> score. Similarly, removal of the inter-pair reasoning module results in a 0.27% drop in F<sub>1</sub> and 0.14% in Ign F<sub>1</sub>. 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<sub>1</sub> score decreases from 62.01% to 60.89% and the Ign F<sub>1</sub> score decreases from 60.02% to 59.12%. This study demonstrates that all components work together in synergy with the final relation classification.

F. 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 III, 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.

G. 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 V, from which we can observe that the context guided cross-attention is absolutely superior to the global strategies. In the table, bold values represent the maximum values. 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.

H. Effect Analysis for Inter-Pair Reasoning

In addition, we evaluate the reasoning ability of our model on the development set of DocRED in Table VI. In the table, bold values represent the maximum values. Following Zeng et al. [22], 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 $e_h \rightarrow r_1 \rightarrow e_o \rightarrow r_2 \rightarrow e_t$ and $e_h \rightarrow r_3 \rightarrow e_t$.

As illustrated in Table VI, 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.

V. Conclusion

In this article, we propose a model called CGM2IR (Context Guided Mention Integration and Inter-pair Reasoning) 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 competitive models.

REFERENCES

[1] B. Salillberg, N. Grbic, and I. Claesson, “Complex-valued independent component analysis for online blind speech extraction,” IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 8, pp. 1624–1632, Nov. 2008.
[2] D. Yu, G. Hinton, N. Morgan, J.-T. Chien, and S. Sagayama, “Introduction to the special section on deep learning for speech and language processing,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 4–6, Jan. 2012.
[3] D. Zeng, K. Liu, Y. Chen, and J. Zhao, “Distant supervision for relation extraction via piecewise convolutional neural networks,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2015, pp. 1753–1762.
[4] L. B. Soares, N. FritzGerald, J. Ling, and T. Kwiatkowski, “Matching the blanks: Distributional similarity for relation learning,” in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 2885–2905.
[5] Y. Yao et al., “DocRED: A large-scale document-level relation extraction dataset,” in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 764–777.
[6] W. Zhou, K. Huang, T. Ma, and J. Huang, “Document-level relation extraction with adaptive thresholding and localized context pooling,” in Proc. 34th AAAI Conf. Artif. Intell., 2020, pp. 14612–14620.
[7] S. K. Sahu, F. Christopoulou, M. Miwa, and S. Ananiadou, “Inter-sentence relation extraction with document-level graph convolutional neural network,” in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 4309–4316.
[8] F. Christopoulou, M. Miwa, and S. Ananiadou, “Connecting the dots: Document-level neural relation extraction with edge-oriented graphs,” in Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 4925–4936.
[9] G. Nan, Z. Guo, I. Sekulic, and W. Lu, “Reasoning with latent structure refinement for document-level relation extraction,” in Proc. 38th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 1546–1557.
[10] B. Xu, Q. Wang, Y. Lyu, Y. Zhu, and Z. Mao, “Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 14149–14157.
[11] S. Zeng, R. Xu, B. Chang, and L. Li, “Double graph based reasoning for document-level relation extraction,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 1630–1640.
[12] J. Li, K. Xu, F. Li, H. Fei, Y. Ren, and D. Ji, “MRN: A locally and globally mention-based reasoning network for document-level relation extraction,” in Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP, 2021, pp. 1359–1370.
[13] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, “Relation classification via convolutional deep neural network,” in Proc. COLING, 25th Int. Conf. Comput. Linguistics: Tech. Papers, 2014, pp. 2335–2344.
[14] R. Cai, X. Zhang, and H. Wang, “Bidirectional recurrent convolutional neural network for relation classification,” in Proc. 54th Annu. Meeting Assoc. Comput. Linguistics, 2016, pp. 756–765.

[15] H. Zheng et al., “PRGC: Potential relation and global correspondence based joint relational triple extraction,” in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., 2021, pp. 6225–6235.

[16] C. Quirk and H. Poon, “Distant supervision for relation extraction beyond the sentence boundary,” in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, 2017, pp. 1171–1182.

[17] W. Xu, K. Chen, and T. Zhao, “Discriminative reasoning for document-level relation extraction,” in Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP, 2021, pp. 1653–1663.

[18] S. Zeng, Y. Wu, and B. Chang, “SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction,” in Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP, 2021, pp. 524–534.

[19] H. Wang, C. Focke, R. Sylvester, N. Mishra, and W. Y. Wang, “Fine-tune BERT for docred with two-step process,” 2019, arXiv:1909.11898.

[20] W. Xu, K. Chen, and T. Zhao, “Document-level relation extraction with reconstruction,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 14167–14175.

[21] W. Xu, K. Chen, and T. Zhao, “Discriminative reasoning for document-level relation extraction,” in Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP, 2021, pp. 1653–1663.

[22] S. Zhang, J. Bu, E. Cao, and W. Sun, “Global-to-local neural networks for document-level relation extraction,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 3711–3721.

[23] H. Zhou, Y. Xu, W. Yao, Z. Liu, C. Lang, and H. Jiang, “Global context-enhanced graph convolutional networks for document-level relation extraction,” in Proc. 28th Int. Conf. Comput. Linguistics, 2020, pp. 6225–6235.

[24] Y. Liu et al., “Roberta: A robustly optimized BERT pretraining approach,” 2020, arXiv:1907.11692.

[25] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 3615–3620.

[26] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. 7th Int. Conf. Learn. Representations, 2019.

[27] N. Zhang et al., “Document-level relation extraction as semantic segmentation,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 1551–1560.

[28] S. Zhang, C. Wong, N. Usuyama, S. Jain, T. Naumann, and H. Poon, “Coreferential reasoning learning for language representation,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 1630–1641.

[29] Z. Zhang et al., “Document-level relation extraction with dual-tier heterogeneous graph,” in Proc. 28th Int. Conf. Comput. Linguistics, 2020, pp. 7170–7186.

[30] D. Ye et al., “Conversational reasoning learning for language representation,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2020, pp. 1239–1250.

[31] D. Quirk and H. Poon, “Distant supervision for relation extraction beyond the sentence boundary,” in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, 2017, pp. 1171–1182.

[32] H. Zhou, Y. Xu, W. Yao, Z. Liu, C. Lang, and H. Jiang, “Global context-enhanced graph convolutional networks for document-level relation extraction,” in Proc. 28th Int. Conf. Comput. Linguistics, 2020, pp. 5259–5270.

[33] H. Zheng, J. Fu, Z. Zha, and J. Luo, “Learning deep bilinear transformation for fine-grained image representation,” in Proc. Annu. Conf. Neural Inf. Process. Syst., 2019, pp. 4279–4288.

[34] J. Li et al., “BioCreative V CDR task corpus: A resource for chemical disease relation extraction,” Database, vol. 2016, 2016, Art. no. baw068.

[35] Y. Liu, H. C. M. Leung, H. Ting, and T. W. Lam, “RENET: A deep learning approach for extracting gene-disease associations from literature,” in Proc. Res. Comput. Mol. Biol. - 23rd Annu. Int. Conf., 2019, pp. 272–284.

[36] H. Zhou, Y. Xu, W. Yao, Z. Liu, C. Lang, and H. Jiang, “Global context-enhanced graph convolutional networks for document-level relation extraction,” in Proc. 28th Int. Conf. Comput. Linguistics, 2020, pp. 5259–5270.

[37] H. Zheng, J. Fu, Z. Zha, and J. Luo, “Learning deep bilinear transformation for fine-grained image representation,” in Proc. Annu. Conf. Neural Inf. Process. Syst., 2019, pp. 4279–4288.

[38] J. Li et al., “BioCreative V CDR task corpus: A resource for chemical disease relation extraction,” Database, vol. 2016, 2016, Art. no. baw068.

[39] Y. Wu, R. Luo, H. C. M. Leung, H. Ting, and T. W. Lam, “RENET: A deep learning approach for extracting gene-disease associations from literature,” in Proc. Res. Comput. Mol. Biol. - 23rd Annu. Int. Conf., 2019, pp. 272–284.

[40] H. Zhou, Y. Xu, W. Yao, Z. Liu, C. Lang, and H. Jiang, “Global context-enhanced graph convolutional networks for document-level relation extraction,” in Proc. 28th Int. Conf. Comput. Linguistics, 2020, pp. 5259–5270.

[41] H. Tang et al., “HIN: Hierarchical inference network for document-level relation extraction,” in Proc. Adv. Knowl. Discov. Data Mining - 24th Pacific-Asia Conf., 2020, pp. 197–209.

[42] Y. Liu et al., “Roberta: A robustly optimized BERT pretraining approach,” 2019, arXiv:1907.11692.

[43] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 3615–3620.

[44] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. 7th Int. Conf. Learn. Representations, 2019.

[45] A. M. Saxe, J. L. McClelland, and S. Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks,” in Proc. Int. Conf. Learn. Representations, 2014.