I2DRE: Intra- and Inter-Pair Interactions based Document-Level Relation Extraction with Binary Hill Loss

1st Shiao Meng
School of Software
Tsinghua University
Beijing, China
msa21@mails.tsinghua.edu.cn

2nd Lijie Wen*
School of Software
Tsinghua University
Beijing, China
wenlj@tsinghua.edu.cn

***Abstract*—Document-level relation extraction (DocRE) aims to identify semantic relations between entity pairs within a document and plays a crucial role in various knowledge-centric applications. Though have achieved certain progress, existing studies still expose two critical vulnerabilities: 1) For a given entity, they typically apply heuristic pooling methods to aggregate all its mention representations into a globally unique entity representation, ignoring that the importance of each mention varies when the entity is paired with different entities. This may lead to wrong predictions due to the disturbance of irrelevant mentions. 2) Most approaches use the features of entity pair itself for relation prediction, without fully utilizing the information from other entity pairs in the document. This limits the prediction of relation instances that rely on logical reasoning. To tackle these problems, we propose a novel intra- and inter-pair interactions based DocRE model (I2DRE). We first introduce an interaction mechanism within each entity pair. By allowing the two entities in each pair to guide each other's aggregation of mention representations, we derive pair-aware entity representations which can focus on key mentions for the entity pair. We further devise a gated biased attention module to selectively propagate information across entity pairs. This inter-pair interaction enables the model to effectively capture global reasoning clues. Moreover, we design a binary hill loss to tackle the missing relation problem prevalent in real-world DocRE datasets, where numerous relation instances are missing in annotations. The loss re-weights the gradients for negative class in traditional binary cross-entropy loss as hill-shaped to alleviate the adverse effect of missing relations. Experimental results on two benchmark datasets demonstrate that our method consistently outperforms existing methods.**

***Index Terms*—Information Extraction, Document-Level Relation Extraction, Intra-Pair Interaction, Inter-Pair Interaction, Noise-Robust Training**

I. Introduction

As a fundamental task of information extraction, relation extraction aims at extracting structured relational facts from unstructured text, which plays a crucial role in a variety of knowledge-based applications. Early studies mostly focus on sentence-level relation extraction which predicts the relation between a pair of entities in a single sentence [1], [2].

*Corresponding author.

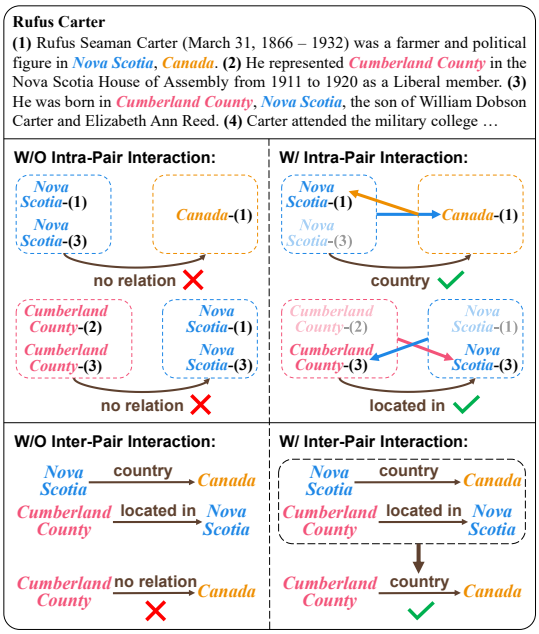

Fig. 1. An example from DocRED dataset [3] to illustrate the significance of intra-pair and inter-pair interactions in DocRE. Sentence numbers are marked in bold and enclosed by brackets. For brevity, only the mentions of the three entities involved are colored in the document.

Nevertheless, sentence-level relation extraction suffers from an inevitable limitation that large amounts of relational facts are expressed and can only be extracted in multiple sentences [3], [4]. Therefore, it is more practical and covers more real-world scenarios to extend relation extraction to document level.

Compared to sentence-level relation extraction, document-level relation extraction (DocRE) is more complex and challenging. One document contains multiple entities and each entity may occur multiple times (each occurrence is called a mention of the entity). Hence, we are required to predict the relations of multiple entity pairs whose quantity is quadratic with the number of entities, and it evolves into a multi-label classification task since an entity pair can express multiple

relations by different mention pairs. More complicated context in DocRE also puts forward higher requirements for reading, memorizing and reasoning across multiple sentences.

To tackle these challenges, some works design graph-based models [5]–[7], which abstract the document by graph structures and apply graph neural networks for information aggregation. Also, another line of sequence-based studies use transformer-only architectures [8]–[10], achieving state-of-the-art performance with the strength of pre-trained language models. Despite the promising results, however, existing methods still expose two critical vulnerabilities, which may limit their capability on relation extraction. First, to derive the representation of a given entity, they typically apply heuristic pooling methods to aggregate all its mention representations, e.g., mean pooling, max pooling and more usually logsumexp pooling. Such methods treat each mention equally and only derive a globally unique entity representation, ignoring that different mentions have different importance when predicting the relation between the entity and a specific paired entity, and also, the importance of each mention varies when the entity is paired with different entities. As shown in Fig. 1, for the entity *Nova Scotia*, the mention in sentence 1 is key for predicting its relation with *Canada*, while the mention in sentence 3 is key for the prediction with *Cumberland County*. Existing approaches may fail on the two pairs due to the disturbance of irrelevant mentions. Second, most existing methods use the local features of entity pair itself for relation prediction, without fully utilizing the information from other entity pairs in the document. As a result, they may fail on the entity pair like (*Cumberland County*, *Canada*) in Fig. 1, since there is no explicit pattern indicating their relation. Actually, we can infer the triple (*Cumberland County*, country, *Canada*) by a two-hop reasoning from (*Cumberland County*, located in, *Nova Scotia*) and (*Nova Scotia*, country, *Canada*).

To overcome the limitations of existing studies, we propose a novel intra- and inter-pair interactions based DocRE model (I2DRE). To address the first issue, we introduce an interaction mechanism within each entity pair, which allows the two entities to guide each other's aggregation of mention representations. Specifically, for a given entity pair, we exploit the inherent attention patterns in pre-trained language models to capture how one entity attend to the mentions of the other. Using the attention signals as aggregation weights over mention representations, we derive pair-aware representations for the two entities, which are conducive to relation prediction since they focus on informative mentions for the entity pair. To tackle the second issue, we devise a gated biased attention module to better exploit inter-pair interaction. In this module, each pair representation attends to the representations of those pairs having overlapping entities in order to capture potential reasoning clues. Additionally, we equip this attention module with two vital control mechanisms, a bias term added to the standard attention logits and a gating operator upon completion of attention. By adaptively regulating the input and output information in attention flow, the two mechanisms collectively facilitate more sufficient utilization of information from *related* pairs (i.e., the pairs that express relations), thus boosting the reasoning ability of model.

Moreover, since the entity number dramatically increases with text length, and correspondingly, the entity pair number surges quadratically, there is an unprecedented difficulty in annotating relation instances for DocRE. Several recent studies [11], [12] analyse the annotation quality of real-world DocRE datasets and point out that a considerable amount of relational facts are missed out in annotations. The numerous missing relations bring noise to the training procedure and lead into an erroneous optimization direction to the learning of the model. For this reason, we further design a novel binary hill loss. Specifically, since DocRE is a multi-label classification problem, it is commonly reduced into multiple binary classification tasks and uses binary cross-entropy loss for training. Our loss is inspired by an observation in multi-label classification scenario [13]: if the model outputs a very large probability for a negative class (i.e., the class is labeled as negative in binary classification) during training, it is quite possible that the class is a missing class and mislabeled as negative. Therefore, we propose to reduce the loss gradients for negative class if the computed probability is very large. Since the gradients for negative class in binary cross-entropy loss increase monotonously with probability, we re-weight the gradients as hill-shaped, i.e., the gradients increase first and decrease after a certain large probability, so that the model is less misled by potential missing relations during training. The code is available at https://github.com/THU-BPM/I2DRE.

In summary, our main contributions are as follows:

- We propose I2DRE, a novel intra- and inter-pair interactions based model for document-level relation extraction. We incorporate intra-pair interaction and allow two entities in each pair guide each other's aggregation of mention representations, in order to better focus on key mentions for relation prediction. We also devise a gated biased attention module to exploit inter-pair interaction, which better captures the global reasoning clues with more effective information propagation among entity pairs.
- To tackle the missing relation issue prevalent in real-world DocRE datasets, we further design a binary hill loss, which re-weights the gradients for negative class in binary cross-entropy loss as hill-shaped to alleviate the adverse impact of missing relations.
- Comprehensive experiments on two benchmark datasets demonstrate that our proposed method consistently and significantly outperforms existing approaches. Further analyses validate the effectiveness of each module.

II. Problem Formulation

Given a document $\mathcal{D} = \{x_i\}_{i=1}^{N_t}$ which consists of N_t tokens and contains a set of entities $\mathcal{E} = \{e_i\}_{i=1}^{N_e}$, the task of document-level relation extraction is to predict the set of all possible relations between each entity pair $(e_s, e_o) \in \{(e_i, e_j) \mid i, j = 1, \ldots, N_e; i \neq j\}$ from a pre-defined relation type set $\mathcal{R}$, where the subscripts of e_s and e_o refer

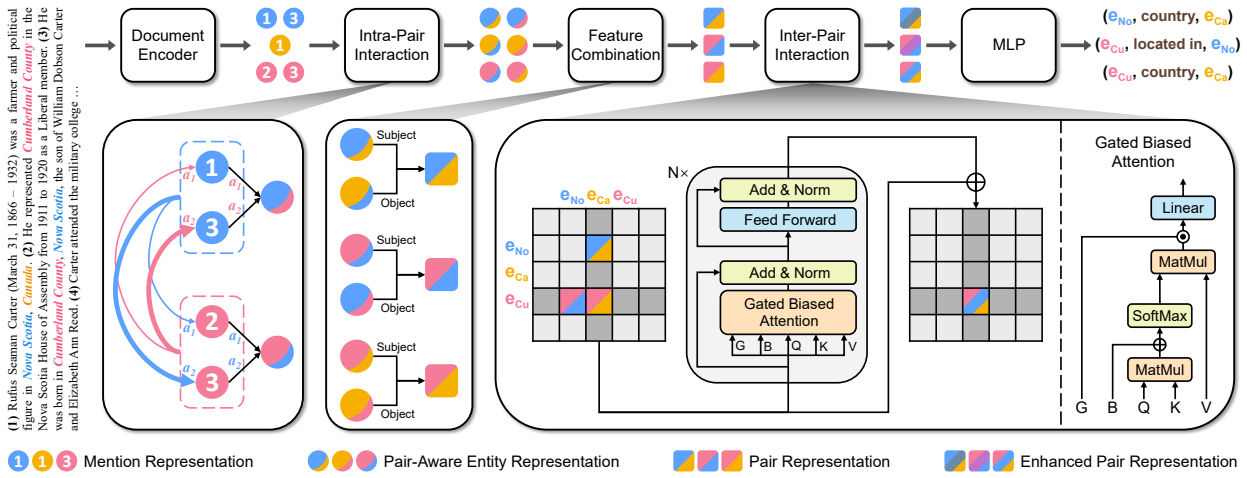

Fig. 2. The overall architecture of I2DRE. Different colors correspond to different entities. For the sake of brevity, only the mentions of the three entities involved are colored in the document. The number in the circular mention representation denotes the order number of sentence that the mention occurs in.

to the subject entity and object entity in a pair. An entity e_i can occur multiple times in the document via N_{e_i} mentions $\mathcal{M}_{e_i} = \{m_j^i\}_{j=1}^{N_{e_i}}$, where the mention m_j^i refers to the token span of e_i's j-th occurrence in the document.

III. Methodology

The overall architecture of our proposed I2DRE is illustrated in Fig. 2. We will elaborate on each module of I2DRE in the following subsections.

A. Document Encoder

Given the promising performance of transformer-based pre-trained language models (PLM) in various tasks [14], [15], we leverage PLM as the document encoder. Specifically, for a document, we first insert a special token "$*$" at the start and end of each entity mention to indicate the mention position. Then we feed the document into a pre-trained language model to obtain the contextualized token embeddings $\boldsymbol{H}$ and the cross token attention $\boldsymbol{A}$:

$$\boldsymbol{H}, \boldsymbol{A} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{N_t}], [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{N_t}] = \text{PLM}([x_1, \ldots, x_{N_t}]), \tag{1}$$

where $\boldsymbol{H} \in \mathbb{R}^{N_t \times d}$, d is the output dimension of the PLM, and $\boldsymbol{A} \in \mathbb{R}^{N_t \times N_t}$ is the average of attention heads in the last layer of PLM. If the document length after tokenization exceeds the maximum input length of the PLM, we will encode the document with a dynamic window whose size is equal to the maximum input length. The final representations of those overlapping tokens are the average of their embeddings in different windows. For each mention m_j^i, we take the embedding of "$*$" at the start of the mention as the corresponding mention representation $\boldsymbol{h}_{m_j^i}$.

B. Intra-Pair Interaction Modeling

Since an entity may have multiple mentions in a document, we are required to aggregate all mention representations of an entity to derive the corresponding entity representation. Existing studies typically apply different heuristic pooling methods to obtain a globally unique representation for each entity. However, all these heuristic strategies do not distinguish the importance of each mention to relation prediction, thus are vulnerable to the disturbance of irrelevant mentions.

Therefore, we propose to incorporate intra-pair interaction into the derivation of entity representations, which allow two entities in each pair guide each other's aggregation of mention representations. In each pair, we endow a pair-specific aggregation weight on each mention representation of one entity, depending on how much another entity focuses on each mention. Considering that the cross token attention in pre-trained language models naturally captures the mention-level dependencies, we propose to directly transfer the pre-trained transformer attention for the modeling of intra-pair interaction, instead of introducing an extra interactive module. Specifically, to derive the representation of entity e_s in pair (e_s, e_o), we take the token-level attention in $\boldsymbol{A}$ from "$*$" symbol before e_o's j-th mention m_j^o to "$*$" symbol before e_s's i-th mention m_i^s as the mention-level attention from m_j^o to m_i^s, which we denote as $a_{m_j^o \to m_i^s}$. Then we average the attention to m_i^s over N_{e_o} mentions of e_o to capture the entity-level attention $a_{e_o \to m_i^s}$ from e_o to m_i^s:

$$a_{e_o \to m_i^s} = \frac{1}{N_{e_o}} \sum_{j=1}^{N_{e_o}} a_{m_j^o \to m_i^s}. \tag{2}$$

The attention from e_o to all N_{e_s} mentions of e_s collectively form an attention vector $\boldsymbol{a}_{e_o \to \mathcal{M}_{e_s}} = [a_{e_o \to m_1^s}, \ldots, a_{e_o \to m_{N_{e_s}}^s}] \in \mathbb{R}^{N_{e_s}}$. To convert $\boldsymbol{a}_{e_o \to \mathcal{M}_{e_s}}$ into a weight distribution over the mention representations of e_s, and meanwhile, keep it consistent with the relative magnitude of attention scores in pre-trained transformer encoder, we further linearly normalize $\boldsymbol{a}_{e_o \to \mathcal{M}_{e_s}}$ as:

$$\boldsymbol{a}_{e_o \to \mathcal{M}_{e_s}} = \boldsymbol{a}_{e_o \to \mathcal{M}_{e_s}} / (\mathbf{1}^\top \boldsymbol{a}_{e_o \to \mathcal{M}_{e_s}}). \tag{3}$$

Then, for the collection of e_s's mention representations $\boldsymbol{H}_{e_s} = [\boldsymbol{h}_{m_1^s}, \ldots, \boldsymbol{h}_{m_{N_{e_s}}^s}] \in \mathbb{R}^{N_{e_s} \times d}$, we perform aggregation accord-

ing to the corresponding weight on each mention representation:

$$\boldsymbol{h}_{e_s}^{(s,o)} = \boldsymbol{H}_{e_s}^{\mathsf{T}} \boldsymbol{a}_{e_o \to \mathcal{M}_{e_s}}, \tag{4}$$

where $\boldsymbol{h}_{e_s}^{(s,o)} \in \mathbb{R}^d$ is the pair-aware representation of entity e_s when e_s is paired with e_o. Similarly, we can derive the representation of entity e_o in pair (e_s, e_o). Compared to a globally unique one for an entity, such pair-aware entity representations based on intra-pair interaction are semantically more accurate and meaningful, helping to focus more on key mentions for relation prediction.

C. Feature Combination

Given the pair-aware representations of each entity, we further combine two entity representations in each pair into the pair representation, so that we are able to model the inter-pair interaction. Following Zhou et al. [8], we first adopt localized context pooling to enhance the entity representation with contextual information. Given an entity pair (e_s, e_o), we denote the attention in $\boldsymbol{A}$ from "$*$" symbol before e_s's i-th mention m_i^s to all tokens in the document as $\boldsymbol{a}_{m_i^s} \in \mathbb{R}^{N_t}$, and average over N_{e_s} mentions of e_s to obtain the entity-level attention $\boldsymbol{a}_{e_s} = \frac{1}{N_{e_s}} \sum_{i=1}^{N_{e_s}} \boldsymbol{a}_{m_i^s} \in \mathbb{R}^{N_t}$, likewise for $\boldsymbol{a}_{e_o}$. Then we compute a context embedding $\boldsymbol{c}^{(s,o)}$ by:

$$\boldsymbol{c}^{(s,o)} = \boldsymbol{H}^{\mathsf{T}} \frac{\boldsymbol{a}_{e_s} \odot \boldsymbol{a}_{e_o}}{\boldsymbol{a}_{e_s}^{\mathsf{T}} \boldsymbol{a}_{e_o}}, \tag{5}$$

where $\odot$ is Hadamard product and $\boldsymbol{c}^{(s,o)} \in \mathbb{R}^d$ is the embedding of context relevant to (e_s, e_o). This localized context embedding is then fused into two pair-aware entity representations of (e_s, e_o):

$$\boldsymbol{z}_s^{(s,o)} = \tanh(\boldsymbol{W}_s \boldsymbol{h}_{e_s}^{(s,o)} + \boldsymbol{W}_{c_s} \boldsymbol{c}^{(s,o)}), \tag{6}$$

$$\boldsymbol{z}_o^{(s,o)} = \tanh(\boldsymbol{W}_o \boldsymbol{h}_{e_o}^{(s,o)} + \boldsymbol{W}_{c_o} \boldsymbol{c}^{(s,o)}), \tag{7}$$

where $\boldsymbol{W}_s$, $\boldsymbol{W}_{c_s}$, $\boldsymbol{W}_o$, $\boldsymbol{W}_{c_o} \in \mathbb{R}^{d \times d}$ are learnable parameters, $\boldsymbol{z}_s^{(s,o)}$, $\boldsymbol{z}_o^{(s,o)} \in \mathbb{R}^d$ are context-enhanced pair-aware representations of e_s and e_o. Then we use a grouped bilinear function for feature combination. We split $\boldsymbol{z}_s^{(s,o)}$ into k equal-sized groups as $\boldsymbol{z}_s^{(s,o)} = [\boldsymbol{z}_{s,1}^{(s,o)}; \ldots; \boldsymbol{z}_{s,k}^{(s,o)}]$ and likewise for $\boldsymbol{z}_o^{(s,o)}$. The value $p_i^{(s,o)}$ at i-th dimension of our pair representation is obtained by:

$$p_i^{(s,o)} = \sum_{j=1}^{k} (\boldsymbol{z}_{s,j}^{(s,o)\mathsf{T}} \boldsymbol{W}_{p_i}^j \boldsymbol{z}_{o,j}^{(s,o)}) + b_i, \tag{8}$$

where $\boldsymbol{W}_{p_i}^j \in \mathbb{R}^{d/k \times d/k}$ for $j = 1, \ldots, k$, $b_i \in \mathbb{R}$ are learnable parameters specific to i-th dimension. The derived pair representation of (e_s, e_o) is d-dimensional and denoted as $\boldsymbol{p}^{(s,o)} = [p_1^{(s,o)}, \ldots, p_d^{(s,o)}]$.

D. Inter-Pair Interaction Modeling

As stated in Section I, the global interaction among entity pairs in a document is vital to the extraction of multi-hop relations. Therefore, in this section, we devise a gated biased attention module to explicitly model the inter-pair interaction, which effectively captures the global reasoning clues for multi-hop relation prediction.

Concretely, we organize all pair representations in a document obtained in Section III-C into a $N_e \times N_e$ pair representation matrix $\boldsymbol{M} = \{\boldsymbol{p}^{(i,j)} \mid i, j = 1, \ldots, N_e\}$, where the row and column correspond to the subject and object entity in a pair, and the diagonal elements are ignored. Since the two-hop reasoning process of an entity pair only involves those pairs sharing the same subject or object entity, each pair representation in the matrix can attend to all other elements in the same row or column to encode potential reasoning information, which we formulate as:

$$\begin{aligned} \boldsymbol{u}^{(s,o)} = &\sum_{n=1}^{N_e} \mathrm{softmax}_n(\boldsymbol{q}^{(s,o)\mathsf{T}} \boldsymbol{k}^{(s,n)}) \boldsymbol{v}^{(s,n)} \\ &+ \sum_{n=1}^{N_e} \mathrm{softmax}_n(\boldsymbol{q}^{(s,o)\mathsf{T}} \boldsymbol{k}^{(n,o)}) \boldsymbol{v}^{(n,o)} \end{aligned}, \tag{9}$$

$$\bar{\boldsymbol{p}}^{(s,o)} = \boldsymbol{W}_O \boldsymbol{u}^{(s,o)} + \boldsymbol{b}_O, \tag{10}$$

where $\boldsymbol{q}^{(i,j)}, \boldsymbol{k}^{(i,j)}, \boldsymbol{v}^{(i,j)} = \boldsymbol{W}_Q \boldsymbol{p}^{(i,j)}, \boldsymbol{W}_K \boldsymbol{p}^{(i,j)}, \boldsymbol{W}_V \boldsymbol{p}^{(i,j)} \in \mathbb{R}^d$ are respectively the projected query, key and value embedding of each pair representation $\boldsymbol{p}^{(i,j)}$ in $\boldsymbol{M}$, $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b}_O \in \mathbb{R}^d$ are learnable parameters, $\bar{\boldsymbol{p}}^{(s,o)} \in \mathbb{R}^d$ is the updated pair representation of (e_s, e_o). Considering that a multi-hop reasoning process can be decomposed into a series of two-hop reasoning process, we follow the general structure of encoder stacks in Transformer [16], as shown in Fig. 2, for the enhancement of reasoning and a deeper mining of potential multi-hop relations.

Furthermore, although it is possible to model the inter-pair interaction with current design, we argue that the information propagation among pairs is not effective enough due to the severe positive-negative imbalance in DocRE [3], [12]. The dominant *unrelated* pairs actually contain no valid clues and bring noise to the pair representation matrix, leading to insufficient utilization of information from minority *related* pairs. Therefore, we further propose the gated biased attention, which equips the vanilla attention with two vital control mechanisms. First, we introduce a bias term $b^{(i,j)}$:

$$b^{(i,j)} = \boldsymbol{W}_B^{\mathsf{T}} \boldsymbol{p}^{(i,j)}, \tag{11}$$

where $\boldsymbol{W}_B \in \mathbb{R}^d$ is a learnable parameter, $b^{(i,j)} \in \mathbb{R}$ is a bias scalar computed for each pair representation $\boldsymbol{p}^{(i,j)}$ in $\boldsymbol{M}$ and added to the standard attention logits when (e_i, e_j) serves as the key pair. This bias term conveys the prior hypothesis regarding the probability of (e_i, e_j) being *related*, making the attention score depend not only on the dot-product similarity between the query pair and key pair, but also on whether the key pair itself has possible relations. As a result, the attention distribution is revised during learning towards more focus on *related* pairs and less focus on *unrelated* pairs. Second, we introduce an output gate $\boldsymbol{g}^{(i,j)}$:

$$\boldsymbol{g}^{(i,j)} = \mathrm{sigmoid}(\boldsymbol{W}_G \boldsymbol{p}^{(i,j)} + \boldsymbol{b}_G), \tag{12}$$

where $\boldsymbol{W}_G \in \mathbb{R}^{d\times d}$ and $\boldsymbol{b}_G \in \mathbb{R}^d$ are learnable parameters, $\boldsymbol{g}^{(i,j)} \in \mathbb{R}^d$ is a gate embedding computed for each pair representation $\boldsymbol{p}^{(i,j)}$ in $\boldsymbol{M}$. The intermediate representation $\boldsymbol{u}^{(i,j)}$ before output projection in attention is element-wise multiplied by $\boldsymbol{g}^{(i,j)}$. Similar with the bias term, this gating operator also learns a signal about each pair representation itself. The difference is the gate acts on the query pair instead of the key pair, which aims to control the output information of each updated pair representation. In consequence, the information flow between layers is regulated to discard more noise from *unrelated* pairs during learning.

Equipped with above two mechanisms, our proposed gated biased attention is formulated as:

$$\boldsymbol{u}^{(s,o)} = \boldsymbol{g}^{(s,o)} \odot \left(\sum_{n=1}^{N_e} \text{softmax}_n(\boldsymbol{q}^{(s,o)\mathsf{T}}\boldsymbol{k}^{(s,n)} + b^{(s,n)})\boldsymbol{v}^{(s,n)} + \sum_{n=1}^{N_e} \text{softmax}_n(\boldsymbol{q}^{(s,o)\mathsf{T}}\boldsymbol{k}^{(n,o)} + b^{(n,o)})\boldsymbol{v}^{(n,o)} \right). \tag{13}$$

By adaptively regulating the input and output information in attention flow, the two mechanisms collectively reinforce the focus on *related* pairs for inter-pair interaction, thus boosting the reasoning ability of model. Notably, we find the improvement is slight to only use one of the two in experiments (Section IV-D, IV-F), which implies that the dual mechanisms are collaborative and indispensible.

E. Binary Hill Loss

Finally, given the representation of pair (e_s, e_o) enhanced by inter-pair interaction, which we denote as $\tilde{\boldsymbol{p}}^{(s,o)}$, we employ a residual connection around $\tilde{\boldsymbol{p}}^{(s,o)}$ and $\boldsymbol{p}^{(s,o)}$ to retain the local features, then use a linear layer to predict the relations between (e_s, e_o):

$$\boldsymbol{l}^{(s,o)} = \boldsymbol{W}_l(\tilde{\boldsymbol{p}}^{(s,o)} + \boldsymbol{p}^{(s,o)}) + \boldsymbol{b}_l, \tag{14}$$

where $\boldsymbol{W}_l \in \mathbb{R}^{N_r\times d}$ and $\boldsymbol{b}_l \in \mathbb{R}^{N_r}$ are learnable parameters, $\boldsymbol{l}^{(s,o)} \in \mathbb{R}^{N_r}$ is the output logits for all relations, $N_r = |\mathcal{R}|$ is the number of pre-defined relation types. Since an entity pair can have multiple relations, we map the logit of each relation into a probability by sigmoid function and use binary cross-entropy loss for training. During inference, we apply and tune a threshold to convert the probability into predicted relations. Considering a global threshold may not be optimal for all pairs, we follow Zhou et al. [8] to introduce a special threshold class TH and use the logit of TH class as a learnable threshold for each pair. Accordingly, given the pair (e_s, e_o) and a relation r_i, the probability of r_i is computed as:

$$P_{r_i}^{(s,o)} = \frac{\exp(l_{r_i}^{(s,o)})}{\exp(l_{r_i}^{(s,o)}) + \exp(l_{TH}^{(s,o)})}, \tag{15}$$

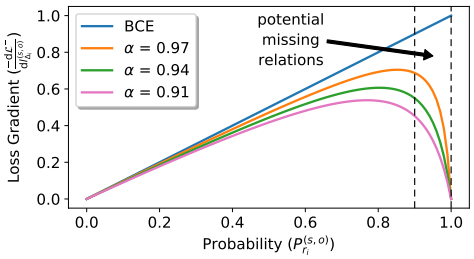

Fig. 3. Analysis of gradients for negative class in Binary Hill Loss under different probability scaling ratios α, in comparison with Binary Cross-Entropy Loss (BCE).

where $l_{r_i}^{(s,o)}, l_{TH}^{(s,o)} \in \mathbb{R}$ are the logit of relation r_i and threshold class TH. Then the loss function of pair (e_s, e_o) is formulated as:

$$\mathcal{L} = -\sum_{r_i\in\mathcal{R}} \left(y_i^{(s,o)} \log P_{r_i}^{(s,o)} + (1 - y_i^{(s,o)}) \log(1 - P_{r_i}^{(s,o)}) \right), \tag{16}$$

where $y_i^{(s,o)} = 1$ if r_i exists between (e_s, e_o), otherwise $y_i^{(s,o)} = 0$.

However, as mentioned in Section I, the missing relation problem is prevalent in real-world DocRE datasets and misleads the learning of model. Therefore, we further propose a novel binary hill loss. The design of our loss is inspired by an observation in multi-label classification scenario [13]: if the model computes a very large probability for a negative class ($y_i^{(s,o)} = 0$) during training, it is possible that the class is mislabeled and its correct label should be positive. Hence, we can alleviate the adverse effect of missing relations by reducing the loss gradients for those negative relations with a very large probability. Since the gradients for negative class in binary cross-entropy loss increase monotonously with probability, we re-weight the gradients as hill-shaped, i.e., the gradients increase first and decrease after a certain large probability. Specifically, we propose to perform re-weighting by probability scaling. Let $\mathcal{L}^- = \log(1 - P_{r_i}^{(s,o)})$ denote the negative loss part for r_i, we introduce a probability scaling ratio α and redefine $\mathcal{L}^-$ as:

$$\mathcal{L}^- = \log(1 - \alpha P_{r_i}^{(s,o)}), \tag{17}$$

where α is a tunable hyperparameter. With probability scaling, the loss gradients for negative class are revised as:

$$\frac{\mathrm{d}\mathcal{L}^-}{\mathrm{d}l_{\Delta_i}^{(s,o)}} = \frac{\mathrm{d}\mathcal{L}^-}{\mathrm{d}P_{r_i}^{(s,o)}} \frac{\mathrm{d}P_{r_i}^{(s,o)}}{\mathrm{d}l_{\Delta_i}^{(s,o)}} = \frac{\alpha P_{r_i}^{(s,o)}(1 - P_{r_i}^{(s,o)})}{1 - \alpha P_{r_i}^{(s,o)}}, \tag{18}$$

where $l_{\Delta_i}^{(s,o)} = l_{r_i}^{(s,o)} - l_{TH}^{(s,o)}$. As illustrated in Fig. 3, with a probability scaling ratio $\alpha < 1$, the gradients for potential missing relations (with a large computed probability) are diminished significantly. Different α lead to different turning points as well as varying degrees of slope. When $\alpha = 1$, the loss degenerates into the binary cross-entropy loss. By re-weighting the gradients for negative relations, the proposed binary hill loss effectively attenuates the adverse effect of

TABLE I

RESULTS ON THE DEVELOPMENT AND TEST SET OF DOCRED AND RE-DOCRED. THE BEST PERFORMANCE AND SECOND BEST PERFORMANCE METHODS ARE DENOTED IN BOLD AND UNDERLINED. "∗" REPRESENTS THE RESULTS ON RE-DOCRED ARE BASED ON OUR IMPLEMENTATION. "†" REPRESENTS THE RESULTS ON BOTH DATASETS ARE BASED ON OUR IMPLEMENTATION.

Model	DocRED				Re-DocRED			
	Dev		Test		Dev		Test	
	Ign F1	F1	Ign F1	F1	Ign F1	F1	Ign F1	F1
GAIN-BERT$_{base}$*	59.14	61.22	59.00	61.24	72.88	73.57	72.53	73.45
ATLOP-BERT$_{base}$*	59.22	61.09	59.31	61.30	73.09	73.86	72.81	73.59
DocuNet-BERT$_{base}$*	59.86	61.83	59.93	61.86	73.52	74.30	73.17	74.04
SagDRE-BERT$_{base}$*	60.32	62.11	60.11	62.32	73.71	74.49	73.62	74.52
KDDocRE-BERT$_{base}$*	60.08	62.03	60.04	62.08	73.86	74.68	73.63	74.45
I2DRE-BERT$_{base}$	**60.71 ± 0.12**	**62.66 ± 0.14**	**60.65**	**62.75**	**74.73 ± 0.11**	**75.58 ± 0.13**	**74.48 ± 0.11**	**75.27 ± 0.12**
SSAN-RoBERTa$_{large}$*	60.25	62.08	59.47	61.42	75.42	76.12	75.61	76.25
ATLOP-RoBERTa$_{large}$	61.32	63.18	61.39	63.40	76.79	77.46	76.82	77.56
DocuNet-RoBERTa$_{large}$	62.23	64.12	62.39	64.55	77.49	78.14	77.26	77.87
SagDRE-RoBERTa$_{large}$†	62.05	64.01	62.08	64.13	77.61	78.49	77.46	78.03
KDDocRE-RoBERTa$_{large}$	62.16	64.19	62.57	64.28	77.85	78.51	77.60	78.28
I2DRE-RoBERTa$_{large}$	**62.76 ± 0.14**	**64.79 ± 0.16**	**63.02**	**64.97**	**78.80 ± 0.13**	**79.59 ± 0.14**	**78.53 ± 0.09**	**79.30 ± 0.10**

missing relations on the training of model, thus yielding a better performance in the presence of missing relations.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We conduct experiments on two widely-used public benchmark datasets:

- DocRED [3] is one of the largest and most popular public benchmark datasets for document-level relation extraction. DocRED is constructed from Wikipedia and Wikidata and is manually annotated in a recommend-revise scheme. It has 96 pre-defined relation types, along with 3053, 1000 and 1000 documents for training, development and test. Each document in DocRED has 19.5 entities and 12.5 relation triples on average.
- Re-DocRED [12] is a revised version of DocRED, which resolves the missing relation issue in DocRED. It adds missed triples in DocRED's training and development set. The 3053 revised training documents contain 28.1 triples on average and 1000 revised development documents (split into 500/500 dev/test documents) have more complete annotation with 34.7 triples on average.

For evaluation metrics, following most DocRE studies, we use Ign F1 and F1 score to evaluate the overall performance. Ign F1 measures the F1 score excluding those relational facts shared by the training and dev/test sets.

B. Baselines

We compare our method with the following competitive DocRE baselines including sequence-based and graph-based approaches: (1) SSAN [17] incorporates entity structure information into the encoding network. (2) GAIN [5] proposes a graph aggregation-and-inference network with mention-level and entity-level graphs. (3) ATLOP [8] tackles multi-label and multi-entity problems with adaptive thresholding and localized context pooling. (4) DocuNet [9] formulates and approaches DocRE as a semantic segmentation task. (5) SagDRE [7] incorporates sentence-level and token-level sequential information in the graph-based model. (6) KDDocRE [10] proposes a DocRE framework enhanced by two-hop relation mining and adaptive focal loss. Part of baselines above (ATLOP, DocuNet and KDDocRE) have publicly reported results on Re-DocRED and we reimplement other baselines with official codes for more comprehensive comparison. For a fair comparison, we use BERT$_{base}$ [14] or RoBERTa$_{large}$ [15] as the encoder to keep consistent with baselines. Note that some models [18]–[20] are not compared since they leverage supporting evidence information or perform additional preprocessing.

C. Main Results

Our main results on DocRED and Re-DocRED are shown in Table I. In general, the proposed I2DRE consistently outperforms all baselines on two datasets in terms of both Ign F1 and F1. It demonstrates the effectiveness and superiority of our method. For DocRED, Table I shows that our best model I2DRE-RoBERTa$_{large}$ outperforms previous best methods with same encoder by 0.85%, 0.93% and 0.72%, 0.65% on the development and test set in terms of Ign F1 and F1. Compared to DocRED, Re-DocRED is more reliable for revising a large number of missing relations in DocRED. As shown in Table I, our method yields more significant improvements over baselines on Re-DocRED. When using BERT$_{base}$ as encoder, I2DRE achieves an increase of 1.18%, 1.21% and 1.15%, 1.01% over the best baselines on the development and test set in terms of Ign F1 and F1. And the corresponding performance increase with RoBERTa$_{large}$ as encoder is 1.22%, 1.38% and 1.20%, 1.30%. Additionally, the performance gain of I2DRE over baselines when using RoBERTa$_{large}$ as encoder is mostly larger than using BERT$_{base}$ on both datasets, indicating that our method can better utilize a stronger encoding model.

D. Ablation Study

To evaluate the influence of each module in our method, we conduct ablation study on the development set of DocRED and Re-DocRED by disabling one component at a time. As

TABLE II
ABLATION STUDY RESULTS.

Model	DocRED Dev		Re-DocRED Dev	
	Ign F1	F1	Ign F1	F1
I2DRE-RoBERTa$_{\text{large}}$	**62.76**	**64.79**	**78.80**	**79.59**
w/o intra-pair interaction	62.18	64.15	78.04	78.84
w/o inter-pair interaction	61.77	63.86	77.59	78.41
– bias of inter-pair interaction	62.47	64.54	78.41	79.26
– gate of inter-pair interaction	62.41	64.42	78.47	79.18
– bias & gate of inter-pair interaction	62.30	64.35	78.25	79.01
w/o binary hill loss	62.09	64.14	78.42	79.18

illustrated in Table II, all components contribute to the overall model performance. Specifically, we can observe that: (1) The modeling of intra-pair interaction is effective for document-level relation extraction. Compared to a globally unique one, the pair-aware entity representations derived by intra-pair interaction can better focus on key mentions for relation prediction. (2) Removing the gated biased attention module leads to a drop of 0.99/0.93 and 1.21/1.18 Ign F1/F1 on DocRED and Re-DocRED, which demonstrates the significance of our modeling of inter-pair interaction. (3) The proposed bias term and gating operator play a vital role in modeling inter-pair interaction. The gated biased attention leads to a raise of 0.46/0.44 and 0.55/0.58 Ign F1/F1 on two datasets compared to the vanilla attention. We attribute the gains to the positive regulatory effect of both mechanisms in attention flow. Moreover, only adopting either of the two mechanisms brings slight effect, indicating the synergy between the two mechanisms. (4) Replacing the binary hill loss with binary cross-entropy loss leads to a drop of 0.67/0.65 and 0.38/0.41 Ign F1/F1 on two datasets. It suggests that re-weighting the loss gradients for negative relations as hill-shaped helps mitigate the adverse effect of missing relations.

E. Analysis of Intra-Pair Interaction

In this section, we further explore the impact of intra-pair interaction on the performance improvements. Since our modeling of intra-pair interaction mainly aims to improve the prediction on those pairs containing multi-mention entities, we divide the entity pairs of all documents in DocRED's development set into disjoint subsets according to the total mention number, i.e., the sum of the mention number of subject and object entity in a pair. We obtain four pair subsets, three of which corresponding to the total mention number of 2, 3 and 4, respectively, and all pairs with total mention number equal to or more than 5 form the last subset. Then we evaluate models trained with or without intra-pair interaction on each subset. Experiment results are shown in Fig. 4. We can observe that the model w/ intra-pair interaction consistently outperforms the model w/o intra-pair interaction on each subset. And more importantly, the improvements get larger as the total mention number increases. It demonstrates that by incorporating intra-pair interaction, the derived fine-grained pair-aware entity representations conduce to capture key mentions for relation prediction in each pair, especially those entity pairs with more mentions involved.

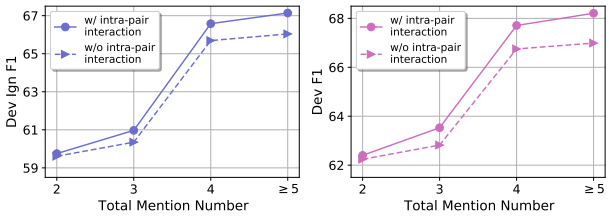

Fig. 4. Dev Ign F1 (left) and Dev F1 (right) score of entity pairs with different total mention numbers on DocRED.

TABLE III
INFER-F1 RESULTS ON THE DEVELOPMENT SET OF DOCRED. RESULTS OF THE MODEL WITH * ARE BASED ON OUR IMPLEMENTATION.

Model	Infer-P	Infer-R	Infer-F1
ATLOP-RoBERTa$_{\text{large}}$ *	40.63	60.74	48.69
DocuNet-RoBERTa$_{\text{large}}$ *	41.37	61.25	49.38
SagDRE-RoBERTa$_{\text{large}}$ *	41.73	61.32	49.66
KDDocRE-RoBERTa$_{\text{large}}$	42.15	61.56	50.04
I2DRE-RoBERTa$_{\text{large}}$	**42.91**	**62.04**	**50.73**
w/o inter-pair interaction	40.80	60.81	48.83
– bias of inter-pair interaction	42.37	61.72	50.25
– gate of inter-pair interaction	42.21	61.75	50.14
– bias & gate of inter-pair interaction	41.96	61.58	49.91

F. Analysis of Inter-Pair Interaction

Considering our modeling of inter-pair interaction mainly aims to tackle with multi-hop relation reasoning, we further analyse the model performance on potential multi-hop relational facts. Following Zeng et al. [5], we use Infer-F1 as the metric that only considers relations engaged in multi-hop reasoning process, i.e., there exist $e_s \xrightarrow{r_1} e_i \xrightarrow{r_2} \dots e_j \xrightarrow{r_k} e_o$ and $e_s \xrightarrow{r} e_o$ in the golden relational facts. Experiment results on the development set of DocRED are shown in Table III. We observe that I2DRE is significantly superior to previous methods in Infer-F1, achieving an increase of 1.38% over the best baseline model. Moreover, removing the gated biased attention module leads to a drop of 1.90 Infer-F1, which is more than double the performance drop of F1 on the overall development set. These results demonstrate that our model can better handle the multi-hop relation reasoning, and the improvements greatly benefit from the devised gated biased attention module. In particular, the proposed bias term and gating operator significantly boost the reasoning ability, without which it leads to a drop of 0.82 Infer-F1. It suggests that by adaptively regulating the input and output information in attention flow, the model is better able to capture reasoning clues from *related* pairs. Moreover, the increase of Infer-F1 with only either mechanism is limited, which further validates the synergy between the two mechanisms.

V. RELATED WORK

Relation extraction is a pivot task for information extraction [21], [22]. Early studies mainly focus on predicting the relation between two entities within a sentence. However, it has been shown that a considerable amount of relational facts are expressed and can only be extracted in multiple sentences, leading to a surge of research interest in document-level

relation extraction [3], [23]. Based on the way of modeling relational information from context, existing studies can be categorized into graph-based and sequence-based approaches. Graph-based methods [5]–[7] typically abstract a document by graph structures based on specific rules or dependencies, and then perform inference with graph neural networks on the graph. For example, Xu et al. [6] use graph attention network to encode a heterogeneous document graph, then enhance the graph representation with path reconstruction. On the other hand, sequence-based methods [8]–[10] encode the long-distance contextual dependencies with transformer-only architectures [16], achieving state-of-the-art results with the strength of pre-trained language models. Particularly, Zhou et al. [8] propose a localized context pooling mechanism to enrich the entity representation for relation prediction.

VI. Conclusion

In this paper, we propose I2DRE, a novel document-level relation extraction framework enhanced by both intra-pair and inter-pair interactions. We incorporate intra-pair interaction, which break the inherent heuristic methods and allow two entities in each pair guide each other's aggregation of mention representations, conducing to focus on key mentions for relation prediction. And we devise a gated biased attention module to exploit inter-pair interaction, which better captures the multi-hop reasoning clues with more effective information propagation among pairs. Moreover, to tackle the missing relation problem prevalent in real-world DocRE datasets, we design a novel binary hill loss to alleviate the adverse effect of potential missing relations. Experiment results and further analyses demonstrate both the superiority of our method and the effectiveness of each component.

Acknowledgment

This work is supported by the National Key Research and Development Program of China (No.2024YFB3309702), the National Nature Science Foundation of China (No.62021002), Tsinghua BNRist and Beijing Key Laboratory of Industrial Big Data System and Application.

References

[1] L. Baldini Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski, "Matching the blanks: Distributional similarity for relation learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2895–2905.

[2] X. Hu, J. Chen, S. Meng, L. Wen, and P. S. Yu, "Selflre: Self-refining representation learning for low-resource relation extraction," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, p. 2364–2368.

[3] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, and M. Sun, "DocRED: A large-scale document-level relation extraction dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 764–777.

[4] S. Meng, X. Hu, A. Liu, F. Ma, Y. Yang, S. Li, and L. Wen, "On the robustness of document-level relation extraction models to entity name variations," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 16 362–16 374.

[5] S. Zeng, R. Xu, B. Chang, and L. Li, "Double graph based reasoning for document-level relation extraction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1630–1640.

[6] W. Xu, K. Chen, and T. Zhao, "Document-level relation extraction with reconstruction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14 167–14 175, 2021.

[7] Y. Wei and Q. Li, "Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, p. 2000–2008.

[8] W. Zhou, K. Huang, T. Ma, and J. Huang, "Document-level relation extraction with adaptive thresholding and localized context pooling," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14 612–14 620, 2021.

[9] N. Zhang, X. Chen, X. Xie, S. Deng, C. Tan, M. Chen, F. Huang, L. Si, and H. Chen, "Document-level relation extraction as semantic segmentation," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2021, pp. 3999–4006.

[10] Q. Tan, R. He, L. Bing, and H. T. Ng, "Document-level relation extraction with adaptive focal loss and knowledge distillation," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1672–1681.

[11] Q. Huang, S. Hao, Y. Ye, S. Zhu, Y. Feng, and D. Zhao, "Does recommend-revise produce reliable annotations? an analysis on missing instances in DocRED," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6241–6252.

[12] Q. Tan, L. Xu, L. Bing, H. T. Ng, and S. M. Aljunied, "Revisiting DocRED - addressing the false negative problem in relation extraction," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 8472–8487.

[13] Y. Zhang, Y. Cheng, X. Huang, F. Wen, R. Feng, Y. Li, and Y. Guo, "Simple and robust loss design for multi-label learning with missing labels," 2021.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[17] B. Xu, Q. Wang, Y. Lyu, Y. Zhu, and Z. Mao, "Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14 149–14 157, 2021.

[18] Y. Xie, J. Shen, S. Li, Y. Mao, and J. Han, "Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 257–268.

[19] Y. Xiao, Z. Zhang, Y. Mao, C. Yang, and J. Han, "SAIS: Supervising and augmenting intermediate steps for document-level relation extraction," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 2395–2409.

[20] Y. Ma, A. Wang, and N. Okazaki, "DREEAM: Guiding attention with evidence for improving document-level relation extraction," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 1971–1983.

[21] X. Hu, J. Chen, A. Liu, S. Meng, L. Wen, and P. S. Yu, "Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, p. 5185–5194.

[22] X. Hu, Z. Hong, C. Zhang, A. Liu, S. Meng, L. Wen, I. King, and P. S. Yu, "Reading broadly to open your mind: Improving open relation extraction with search documents under self-supervisions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 5, pp. 2026–2040, 2024.

[23] S. Meng, X. Hu, A. Liu, S. Li, F. Ma, Y. Yang, and L. Wen, "RAPL: A relation-aware prototype learning approach for few-shot document-level relation extraction," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 5208–5226.