Attention Based Document-level Relation Extraction with None Class Ranking Loss

Xiaolong Xu[1,2] , **Chenbin Li**[1] , **Haolong Xiang**[1,2*] , **Lianyong Qi**[3] , **Xuyun Zhang**[4*] and **Wanchun Dou**[5]

[1]School of Software, Nanjing University of Information Science & Technology, China
[2]Jiangsu Province Engineering Research Center of Advanced Computing and Intelligent Services, China
[3]College of Computer Science and Technology, China University of Petroleum, China
[4]School of Computing, Macquarie University, Australia
[5]Department of Computer Science and Technology, Nanjing University, China
xlxu@ieee.org, 202212490318@nuist.edu.cn, hlx6700@gmail.com, lianyongqi@gmail.com, xuyun.zhang@mq.edu.au, douwc@nju.edu.cn

Abstract

Through document-level relation extraction (RE), the analysis of the global relation between entities in the text is feasible, and more comprehensive and accurate semantic information can be obtained. In document-level RE, the model needs to infer the implicit relations between two entities in different sentences. To obtain more semantic information, existing methods mainly focus on exploring entity representations. However, they ignore the correlations and indivisibility between relations, entities and contexts. Furthermore, current methods only independently estimate the cases of predefined relations, ignoring the case of "no relation", which results in poor prediction. To address the above issues, we propose a document-level RE method based on attention mechanisms, which considers the case of "no relation". Specifically, our approach leverages graph attention and multi-head attention networks to capture the correlations and indivisibility among relations, entities, and contexts, respectively. In addition, a novel multi-label loss function that promotes large margins in label confidence scores between each predefined class and the none class is employed to improve the prediction performance. Extensive experiments conducted on benchmarking datasets demonstrate that our proposed method outperforms the state-of-the-art baselines with higher accuracy.

1 Introduction

Relation extraction (RE) is an important subtask in information extraction (IE) and plays a key role in constructing large-scale knowledge graphs for many applications such as recommendation systems. Considerable effort has been made to extract all entity relation triples within a sentence, yielding some success in what is known as sentence-level RE [Zhang *et al.*, 2018; Soares *et al.*, 2019]. In contrast, there is much less work in document-level RE [Zeng *et al.*, 2020; Zhou *et al.*, 2021; Jiang *et al.*, 2022; Zhou and Lee, 2022]. Document-level RE is more challenging than sentence-level RE since documents often contain a lot of implicit relations that can only be identified with the help of relation reasoning [Xu *et al.*, 2023; Zhang *et al.*, 2023]. It is necessary to consider the information across sentences and establish the relation between entities, rather than just extract the relation from a single sentence. In addition, there may be cross-sentence, cross-paragraph or even cross-chapter relations in the document, which requires cross-text relation reasoning. For example, as shown in Figure 1, the subjects "Wisconsin", "the United States House of Representatives" and the object "US" are located in different sentences, which makes it hard to discover the relations between them. To be specific, document-level RE is facing two main challenges as follows:

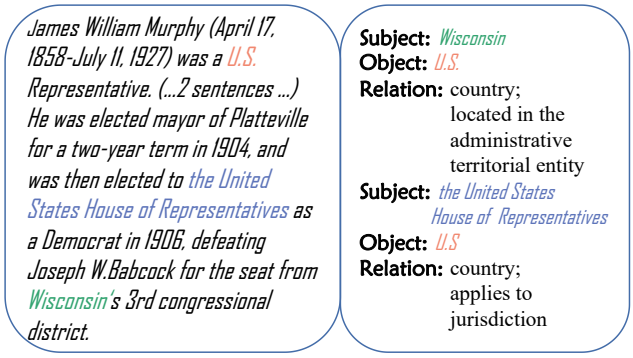

Figure 1: An example of document-level RE.

Challenge 1: Long-tail Problem. In document-level RE, there are numerous low-frequency or sparse relation instances, known as the long-tail problem. In the context of long-tail data, the model tends to prioritize high-frequency categories while struggling to recognize the tail ones, exacerbating the challenge in document-level RE. This imbalance makes it more difficult to train the model, as the model can not fully learn the context of the relations and capture their distinctive features.

Challenge 2: Multi-label Problem. In document-level RE, a single document can involve multiple relation types simultaneously, leading to multiple labels. For instance, a news report may encompass various aspects such as relations between individuals, location information, and event timing, which correspond to the labels. Consequently, the task of document-level RE becomes a multi-label classification problem where the model needs to predict all existing relation types within the document concurrently.

To address these challenges, most existing approaches apply Transformer [Xie *et al.*, 2022; Zhou *et al.*, 2021] and graph-based approaches [Nan *et al.*, 2020; Peng *et al.*, 2022; Xu *et al.*, 2021b] to solve the long-tail problem and multi-label classification problem of relation types. However, these approaches ignore that relation, as an important component in the triple, exhibits informative correlations together with entities and contexts. For example, the two relations "last name" and "inclusion" often co-occur between the same pair of entities, but these two relations occur at different rates cue with the specific contexts. Thus, the ability to model and fully exploit the rich interactions across relations, entities, and contexts is critical to the task. Although the researchers [Xu *et al.*, 2022] have proposed EmRel to fusion representations of relations, entities and contexts, they still ignore the correlations between relation types.

Moreover, the majority of existing methods for document-level RE rely on binary cross entropy (BCE) loss [Bishop, 1995], which solely considers the case of predefined relations independently while disregarding the case of "no relation". However, this loss overlooks the crucial insight that both the cases of none class and predefined classes are pivotal for multi-label prediction due to their high correlation (increased cases of the none class translate to decreased cases of all predefined classes). Without consideration of the case of "no relation", whether an entity pair has certain relations is determined via global thresholding (i.e., the labels, corresponding to the certain relations, have higher confidence scores than a defined threshold will be selected as final predictions). However, this approach fails to take the context of each entity pair into account and produces too many or too few labels, thereby yielding poor predictions.

To deal with the above issues, we propose a novel approach with none class ranking loss, which leverages embedded representations of relations. Specifically, we treat relation representations as embedding vectors and model the correlations among relation types and the indivisibility among relations, entities, and contexts via an attention-based fusion module. Then, we employ a novel multi-label loss named none class ranking loss to improve the prediction performance.

To demonstrate the validity and generality of our method, we conducted comprehensive experiments on public datasets, and the results showed that our method consistently outperformed all competing baselines. The main contributions of this paper are as follows:

- We model the correlations among relation types and indivisibility of relations, entities, and contexts through graph attention and multi-head attention networks.
- We use a novel multi-label loss named none class ranking loss to improve the prediction performance through a novel multi-label loss named none class ranking loss.
- Intensive experiments are conducted on the two commonly used datasets, which proves that capturing rich semantic information contained in relations, entities and contexts and taking the case of "no relation" into consideration help improve the accuracy of joint extraction of entity and relation. The sourcecode is available at https://github.com/sugarman490/Document-RE.

2 Related Work

Extracting multi-triples from document-level text has recently aroused increasing interest [Yao *et al.*, 2019]. Existing methods take the entity perspective that proposes various techniques to model entity interactions. LSR [Nan *et al.*, 2020] and GAIN [Zeng *et al.*, 2020] construct an entity graph, and perform graph-level reasoning to refine the entity node representations. SSAN [Xu *et al.*, 2021a] introduces entity structure as useful prior, and models such information within the transformer attention layer. DocuNet [Zhang *et al.*, 2021] utilizes a segmentation network to model the interdependency among entity pairs. DRE [Xu *et al.*, 2023] proposes a new evidence extraction model which explicitly extracts the precise evidence for each entity pair. [Zhang *et al.*, 2023] emphasizes the bidirectional semantic interaction between the document encoder and the graph encoder. In conclusion, inter-triple correlations are only captured at the entity level by these methods. However, the correlations and indivisibility among relations, entities and contexts are neglected.

Relation Embedding. There is one specific previous work [Chen and Badlani, 2020] that also considers modeling relation representations. CRE uses the sentence representation as relation embeddings, and scores them with the entity embeddings trained along with the knowledge base. EmRel [Xu *et al.*, 2022] jointly represents relations with entities. However, they still ignore the correlations among relation types which are important for improving the accuracy of extracting triples. This raises significant advantages with our method in design that our method models both the correlations among relation types and the inseparability between relations and entities or contexts. SRLR [Huang *et al.*, 2023] separates the entity pair to the mention-level that can capture more fine-grained relation representation and elicit logical reasoning based on evidence sentences to enhance the performance.

Multi-label Learning. Multi-label classification is a well-studied problem with rich literature. Here, we focus on loss functions for multi-label learning. BCE is the most popular multi-label loss, which reduces the multi-label problem into a number of independent binary (one-vs-all) classification tasks, one for each label. Another common multil-abel loss function is pairwise ranking loss, which transforms multi-label learning into a ranking problem via pairwise (one-vs-one) comparison. Many attempts have been made to improve these two types of multi-label loss functions [Yeh *et al.*, 2017; Li *et al.*, 2017; Durand *et al.*, 2019; Xu *et al.*, 2019; Wu *et al.*, 2019; Ridnik *et al.*, 2021]. However, as far as we

know, none of the existing loss functions as well as theoretical studies consider none class instances, which commonly appear in real-world multi-label applications.

Contrary to previous approaches, we consider the correlations between relation types and the indivisibility among relations, entities, and contexts by utilizing graph attention networks and multi-head attention networks. Furthermore, we employ a novel loss to improve the poor prediction caused by ignoring the case of "no relation". With these changes, our method achieves higher accuracy than the previous ones.

3 Methodology

In this section, we provide a detailed description of our method. As illustrated in Figure 2, our method comprises three components: the fusion module for representation, the alignment module, and the none class ranking loss. Firstly, within the fusion module for representation, interactive modeling is conducted between relations using graph attention networks (GAT) [Veličković *et al.*, 2018] to capture the correlations among relation types (Section 3.1), which is ignored by previous methods. Subsequently, we briefly describe a representation alignment module that aligns the triple structures (Section 3.2). Finally, we utilize the none class ranking loss to improve the poor prediction performance existing in current methods (Section 3.3). Many ablation experiments demonstrate the effectiveness of our designed modules.

3.1 Fusion Module for Representation

The pre-trained language model BERT [Kenton and Toutanova, 2019] is utilized to construct the entity representation and extract the context representation from the final layer at the corresponding position:

$$(h_1, h_2, ...h_n) = \text{BERT}(w_1, w_2, ...w_n), \tag{1}$$

where we denote it as $\mathbf{H} \in \mathbb{R}^{|\{w_i\}|\times d_h}$. We then construct each entity representation $\mathbf{e}_i \in \mathbb{R}^{d_e}$ by applying the pooling operation to its corresponding reference location, and further map it to the corresponding subject and object representations $\mathbf{e}_i^s, \mathbf{e}_i^o$. Thus, we represent all extracted entity representations as $\mathbf{E}^s, \mathbf{E}^o \in \mathbb{R}^{|E|\times d_e}$.

Relation Correlation. Using several relation labels in Figure 1 as an example, we can see correlations among relation types in Figure 3. The relation correlation interdependence is initially modeled using conditional probability, where $P(L_b \mid L_a)$ denotes the likelihood of relation type L_b occurring given the occurrence of relation type L_a. To account for the symmetry in conditional probability $P(L_b \mid L_a) \neq P(L_a \mid L_b)$, a relation-dependent digraph is constructed based on prior relation knowledge from the training set used for modeling, resulting in an asymmetric adjacency matrix.

To construct the correlation matrix, we calculate the co-occurrence of the relations and obtain the co-occurrence matrix $C^{r\times r}$, where r is the number of predefined relation types. To obtain the conditional probability among relations, we divide each element of the co-occurrence matrix $C^{r\times r}$ by the total number of co-occurrences of the relations:

$$P_{ij} = C_{ij} / \sum_j C_{ij}, \tag{2}$$

where $P_{ij} = P\left(L_j \mid L_i\right)$ represents the probability of relation type L_j when relation type L_i occurs.

However, the previous method of constructing a correlation matrix may present two drawbacks. Firstly, certain relations seldom occur in conjunction with others, leading to an inflated probability value of co-occurrence that is deemed unreasonable. Secondly, there may exist a statistical bias between the training dataset and test dataset. Utilizing precise numbers tends to overfit the training data, potentially compromising generalization. To mitigate these issues, we introduce a threshold value τ to filter out these infrequent co-occurrence relations and address the aforementioned concerns. Then we binarize the conditional probability matrix P by

$$B_{ij} = \begin{cases} 0, & \text{if} \quad P_{ij} < \delta \\ 1, & \text{if} \quad P_{ij} \geq \delta \end{cases}, \tag{3}$$

where B is the binarized correlation matrix. δ is the conditional probability threshold. Besides, We add the self-loop by setting $B_{ii} = 1, i = 1, ..., r$.

One challenge associated with utilizing a binary correlation matrix B in graph neural networks is the issue of oversmoothing [Zhou *et al.*, 2020], wherein node attribute vectors tend to converge towards similar values. This occurs due to the absence of inherent weight differentiation between relational features and those of neighboring nodes. To address this problem, we employ the following reweighting scheme:

$$R_{ij} = \begin{cases} p / \sum_{j=1}^{r} B_{ij}, & \text{if} \quad i \neq j \\ 1 - p, & \text{if} \quad i = j \end{cases}, \tag{4}$$

where R is the re-weighted relation correlation matrix and p is a hyperparameter. In this way, the fixed weights for the relation feature and its neighbors will be applied during training, which is beneficial for alleviating this issue.

Relation features are embeddings obtained similarly to word embeddings. We employ GAT networks with K-head attention mechanisms to aggregate these relation features for two primary reasons. Firstly, GAT is well-suited for directed graphs, making it an appropriate choice. Secondly, GAT offers enhanced expressive capabilities as the weights of each node can vary. The resulting feature after the GAT transformation is denoted as $R \in \mathbb{R}^{r\times d_B}$, where r represents the number of relation features and d_B denotes the dimensionality of the transformed feature.

Representation Fusion. To effectively capture the indivisibility between entities and relations within a shared knowledge representation space, we fuse them to facilitate mutual learning. To model the interactions between these components, we employ the attention network [Bahdanau *et al.*, 2015], which has demonstrated great success in capturing intricate relations within contexts [Yu *et al.*, 2018] and modalities [Lu *et al.*, 2016]. In particular, we adopt the widely-used multi-head Attention (MHA) network [Vaswani *et al.*, 2017] in our work. In the context of our task, where the target represents $\mathbf{X}_Q$ and the source represents $\mathbf{X}_S$, each head of the MHA network operates as follows:

$$\begin{aligned} \widehat{\mathbf{X}}_Q =& \texttt{Att}(\mathbf{X}_Q W^Q, \mathbf{X}_S W^K, \mathbf{X}_S W^V) \\ =& \text{softmax}(\frac{(\mathbf{X}_Q W^Q)(\mathbf{X}_S W^K)^T}{\sqrt{d_k}}) \mathbf{X}_S W^V, \end{aligned} \tag{5}$$

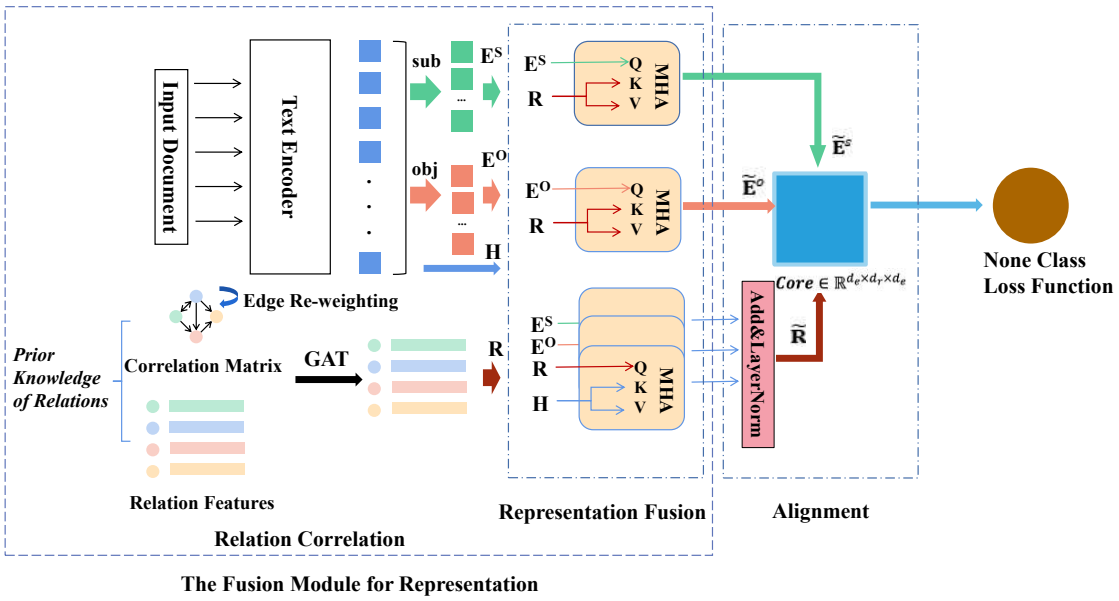

Figure 2: The framework of our method.

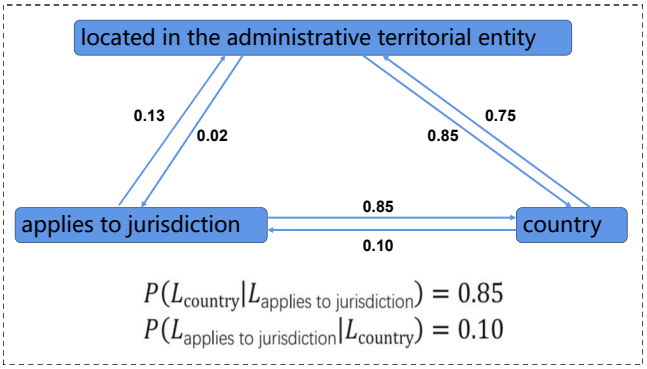

Figure 3: An example of the correlation calculation among the relations in Figure 1.

where $\widehat{\mathbf{X}}_Q$ represents the updated representation of $\mathbf{X}_Q$ with respect to $\mathbf{X}_S$, all heads function concurrently and their outputs are concatenated together. To take advantage of the full interaction among all components, we first build an entity/context aware representation of relations R:

$$\begin{aligned}\hat{\mathbf{R}}^s &= \mathbf{Att}_{s2r}(\mathbf{R}W^Q, \mathbf{E}^sW^K, \mathbf{E}^sW^V),\\ \hat{\mathbf{R}}^o &= \mathbf{Att}_{o2r}(\mathbf{R}W^Q, \mathbf{E}^oW^K, \mathbf{E}^oW^V),\\ \hat{\mathbf{R}}^c &= \mathbf{Att}_{c2r}(\mathbf{R}W^Q, \mathbf{H}W^K, \mathbf{H}W^V),\end{aligned} \tag{6}$$

which are then aggregated together using layer normalization:

$$\hat{\mathbf{R}} = \text{LayerNorm}(\hat{\mathbf{R}}^s + \hat{\mathbf{R}}^o + \hat{\mathbf{R}}^c). \tag{7}$$

We generate relation-aware entity representations based on the relation features acquired from the Relation Correlation Module as follows:

$$\begin{aligned}\widehat{\mathbf{E}}^s &= \mathbf{Att}_{r2s}(\mathbf{E}^sW^Q, \mathbf{R}W^K, \mathbf{R}W^V),\\ \widehat{\mathbf{E}}^o &= \mathbf{Att}_{r2o}(\mathbf{E}^oW^Q, \mathbf{R}W^K, \mathbf{R}W^V).\end{aligned} \tag{8}$$

The abbreviations "s", "o" and "c" stand for subject, object and context respectively. Each attention module is enveloped with residual connections, feedforward layers, layer normalization, and instantiated with distinct W^Q, W^K, and W^V parameters to capture diverse attention patterns. The fusion module produces comprehensive representations of relations, subjects and objects denoted as $\hat{\mathbf{R}}$,$\widehat{\mathbf{E}}^s$,$\widehat{\mathbf{E}}^o$.

3.2 Alignment Module

We extracts triples by aligning their ternary components R, $\widehat{\mathbf{E}}^s$, and $\widehat{\mathbf{E}}^o$. To fully leverage their expressive capabilities, we propose factory-based alignments utilizing Tucker factorization [Tucker and others, 1964]. In this approach, we introduce a core tensor $\mathcal{Z} \in \mathbb{R}^{d_e * d_B * d_e}$. The validity score for each $< s_i, r_k, o_j >$ is calculated as follows:

$$\phi(s_i, r_k, o_j) = \sigma(\mathcal{Z} \times_1 \hat{\mathbf{e}}_i^s \times_2 \hat{\mathbf{r}}_k \times_3 \hat{\mathbf{e}}_j^o + b_k), \tag{9}$$

where $\hat{\mathbf{e}}_i^s = \widehat{\mathbf{E}}^s_{i,:}$, $\hat{\mathbf{r}}_k = \mathbf{R}_{k,:}$,$\hat{\mathbf{e}}_i^o = \widehat{\mathbf{E}}^o_{j,:}$, and $\times n$ indicates tensor product along the n-th mode, σ denotes sigmoid function.

3.3 None Class Ranking Loss

The task of Document-level RE can be viewed as a multi-label classification problem, where each entity pair represents an instance and the association between them serves as a label sample. Let $\mathcal{X}$ denote the instance space and $\mathcal{Y} = \{0, 1\}^K$ represent the label space, with K being the number of predefined classes. An instance $x \in \mathcal{X}$ is associated with a subset of labels, represented by the binary

vector $y \in \mathcal{Y} = (y_1, \ldots, y_K)$, where $y_i = 1$ if the i-th label is positive for x; otherwise $y_i = 0$. To indicate that there are none class instances, all labels can be set to negative $(y_i = 0, i = 1, \ldots, K)$. For ease of notation and clarity in presenting the loss function, we introduce a special class called none and assign it the 0-th label y_0 without loss of generality. Consequently, the absence of any class instance can simply be indicated by setting $y_0 = 1$.

Performance Measure. In contrast to document-level RE, traditional multi-label datasets typically assume that each instance contains at least one positive label, ensuring that no instances lack a class for training. Consequently, existing multi-label performance measures primarily focus on the prediction of predefined labels, disregarding the prediction impact of the none class. However, the probability of the predefined class is closely associated with the probability of the none class. Neglecting this label correlation in performance measures can lead to the oversight of poor predictions that should be penalized during training. This oversight increases the risk of misclassification, particularly in the case of document-level RE, where a significant portion of entity pairs falls within the none class.

To address the need for improved evaluation of document-level RE performance, we transform the multi-label prediction of each instance into a label ranking, where positive predefined labels should be ranked higher than the none class label, while negative ones should be ranked below. Formally, we introduce a multi-label performance measure. The formulation is as follows:

$$\begin{aligned}\ell_{\mathrm{NA}}(\boldsymbol{y}, \boldsymbol{f}) = \sum_{i=1}^{K} \Big(&[\![y_i > 0]\!][\![f_i < f_0]\!] \\ &+ [\![y_i \le 0]\!][\![f_i > f_0]\!] + \frac{1}{2}[\![f_i = f_0]\!] \Big),\end{aligned} \tag{10}$$

where f_i is the confidence score of class i, and $[\![\,]\!]$ is a mapping that returns 1 if its content is true and 0 otherwise. The multi-label performance measure treats the none class label as an adaptive threshold that distinguishes positive from negative labels and counts the number of reverse-arranged label pairs between the none class and each predefined class, which takes into account errors when sorting both the none class and the predefined labels.

Loss Function. For practical multi-label learning, it is essential to minimize surrogate loss, which should be simple and amenable to optimization. Therefore, we employ none class ranking loss:

$$L_{\mathrm{NA}}(\boldsymbol{y}, \boldsymbol{f}) = -\sum_{i=1}^{K} \Big(y_i \cdot \log \sigma(m_i^+) + (1 - y_i) \cdot \log \sigma(m_i^-) \Big), \tag{11}$$

where $m_i^+ = f_i - f_0$ and $m_i^- = f_0 - f_i$ represent positive and negative label margins, respectively, and $\sigma(x)$ is the sigmod function that converts label margins to the probability output in [0,1]. Intuitively, minimizing encourages positive (negative) labels to score as high (low) as possible over none class labels. When testing, we can return a label with a score higher than f_0 as a prediction label, or NA if no such label exists.

Margin Regularization. In document-level RE datasets, the number of positive samples per relation is much larger than the number of negative samples. This positive and negative sample imbalance allows the loss to push a large number of negative samples away from the threshold f_0 to achieve low training loss. Therefore, training with unbalanced data tends to make negative predictions for predefined labels with rare positive samples. To solve this problem, we propose to control the average margin and design the following regularization term:

$$L_{\mathrm{NA}_0}(\boldsymbol{y}, \boldsymbol{f}) = -y_0 \cdot \log \sigma(m_0^+) - (1 - y_0) \cdot \log \sigma(m_0^-). \tag{12}$$

The average positive and negative margins, denoted as $m_0^+ = f_0 - \frac{1}{K}\sum_{i=1}^{K} f_i$ and $m_0^- = \frac{1}{K}\sum_{i=1}^{K} f_i - f_0$ respectively, are utilized in the margin regularization $L_{\mathrm{NA}_0}(\boldsymbol{y}, \boldsymbol{f})$ for instances with $y_0 = 0$ to penalize the average score overall pre-defined labels if it is smaller than f_0. This approach avoids over-suppressing scores of negative samples. For none class instances with $y_0 = 1$, a threshold value of f_0 larger than the largest score among all pre-defined labels is required. However, directly maximizing the margin $f_0 - \max_{i \neq 0} f_i$ can lead to unstable results due to changes in the pre-defined label with $\max_{i \neq 0} f_i$ during training. Instead, $L_{\mathrm{NA}_0}(\boldsymbol{y}, \boldsymbol{f})$ maximizes the upper-bound of $f_0 - \frac{1}{K}\sum_{i=1}^{K} f_i$ which is given by the average margin value of $f_0 - \max_{i \neq 0} f_i$, encouraging that value to be at top position in label ranking.

Margin Shifting. Another practical problem in document-level RE is the problem of labeling errors. Because documents can describe a variety of things, it is difficult for human annotators to accurately and broadly annotate all the relations of each entity pair. As a result, document-level RE datasets often contain considerable noise in markup. Among mislabeled data, false-negative samples are particularly common in practice.

To obtain stronger robustness to mislabeled samples, we suggest shifting the transformed negative margin $\sigma(m_i^-)$ in formula (10) as follows:

$$p_i^- = \max(\sigma(m_i^-) + \gamma, 1), i = 0, 1, \ldots, K, \tag{13}$$

where $0 < \gamma < 1$ is a hyper-parameter. This margin shifting approach increases the probability of all negative samples by a factor γ, thereby attenuating the impact of very hard negative samples with small $\sigma(m_i^-) = \mathbb{P}(y_i = 0|\boldsymbol{x})$ during training, which are suspected to be mislabeled. Additionally, very easy samples with $\sigma(m_i^-) > 1 - \gamma$ get a zero loss and are ignored, enabling the loss to focus more on harder samples.

Despite the existence of similar shifting methods [Lin *et al.*, 2017; Ridnik *et al.*, 2021; Wu *et al.*, 2020], they fail to consider none class labels and shift the score f_i instead of the margin m_i^-. Since the none class score f_0 aligns the margins m_i^- of pre-defined labels in the same range, margin shifting is more suitable than score shifting in adjusting the prediction probability for different training samples. Incorporating the margin regularization and shifting, our final loss function is defined as follows:

$$L_{\mathrm{NA}}(\boldsymbol{y}, \boldsymbol{f}) = -\sum_{i=0}^{K} \left(y_i \cdot \log \sigma(m_i^+) + (1 - y_i) \cdot \log p_i^- \right). \tag{14}$$

Model	DocRED				Re-DocRED			
	Dev		Test		Dev		Test	
	Ign F1	F1	IgnF1	F1	Ign F1	F1	Ign F1	F1
GEDA-BERT [Li *et al.*, 2020]	54.52	56.16	53.71	55.74	-	-	-	-
LSR-BERT [Nan *et al.*, 2020]	52.43	59.00	56.97	59.05	-	-	-	-
GLRE-BERT [Wang *et al.*, 2020]	-	-	55.40	57.40	-	-	-	-
GAIN-BERT [Zeng *et al.*, 2020]	59.14	61.22	59.00	61.24	71.99*	73.49*	71.88*	73.44*
HeterGSAN-BERT [Xu *et al.*, 2021c]	58.13	60.18	57.12	59.45	-	-	-	-
SSAN-BERT [Xu *et al.*, 2021a]	56.68	58.95	56.06	58.41	-	-	-	-
BERT-TS [Wang *et al.*,]	-	54.42	-	53.92	-	-	-	-
HIN-BERT [Tang *et al.*, 2020]	54.29	56.31	53.70	55.60	-	-	-	-
CorefBERT [Ye *et al.*, 2020]	55.32	57.51	54.54	56.96	-	-	-	-
ATLOP-BERT [Zhou *et al.*, 2021]	59.22	61.09	59.31	61.30	73.35*	74.22*	73.22*	74.02*
SIRE-BERT [Zeng *et al.*, 2021]	59.82	61.60	60.18	62.05	-	-	-	-
DocuNet-BERT [Zhang *et al.*, 2021]	59.86	61.83	59.93	61.86	73.60†	74.62†	73.53†	74.48†
KD-DocRE-BERT [Tan *et al.*, 2022a]	60.08	62.03	60.04	62.08	73.68†	74.66†	73.64†	74.55†
KMGRE-BERT [Jiang *et al.*, 2022]	-	-	-	-	73.33*	74.44*	73.39*	74.46*
SRLR-BERT [Huang *et al.*, 2023]	60.02	62.14	59.74	61.88	-	-	-	-
Ours-BERT	**61.55±0.20**	**63.49±0.15**	**61.51**	**63.40**	**75.54**	**76.26**	**75.70**	**76.30**

Table 1: Experimental results on the datasets DocRED and Re-DocRED. We report the mean and standard deviation on the development set by conducting five experiments with different random seeds. Besides, we report the test scores of the best checkpoint on the development set. * indicates that scores are reported in [Jiang *et al.*, 2022]. Results with † are obtained by our reproduction. The best results are in **bold**.

4 Experiments

4.1 Datasets and Settings

Datasets. The experiments are conducted on two widely used datasets, shown as follows:

- **DocRED** [Yao *et al.*, 2019] is a common document-level RE dataset used to evaluate and promote research on document-level RE tasks. The goal of this dataset is to extract relation triples from a given document. The documentation for the DocRED dataset is sourced from Wikipedia and news articles, covering multiple fields and topics. Compared with traditional sentence-level RE tasks, DocRED datasets examine more complex RE scenarios that require cross-sentence, cross-paragraph and even cross-discourse relation inference and reasoning.
- **Re-DocRED** [Tan *et al.*, 2022b] is another revised version of DocRED. Since DocRED contains a considerable number of false-negative samples, we still conduct experiments on this revised version.

Settings. We use BERT-Base-Cased [Kenton and Toutanova, 2019] as the text encoder. The dimension of embedded relation representation is set as 768. The learning rate for BERT parameters is $5e^{-5}$ and e^{-4} for other layers. In the relation correlation module, we set the threshold τ to be 10 for filtering noisy co-occurred relations, and δ is set to be 0.05 in equation (3). We set p to be 0.3 in equation (4). We employ 2-layer GAT networks with k = 2 attention heads computing 500 hidden features per head. We utilize the exponential linear unit (ELU) as the activation function between GAT layers. The number of attention heads in the fusion module is simply set as 4.

4.2 Baselines and Evaluation Metrics

Baselines. We conduct a comparison between our method and the Transformer-based and Graph-based methods.

- **Transformer-based methods** directly employ PLMs to enhance the acquisition of entity representations for document-level RE. Examples of such methods include BERT-TS [Wang *et al.*,], HINBERT [Tang *et al.*, 2020], CorefBERT [Ye *et al.*, 2020], ATLOP-BERT [Zhou *et al.*, 2021] and SRLR-BERT [Huang *et al.*, 2023].
- **Graph-based methods** leverage GNNs to improve the reasoning capability of document-level RE models. Examples of such methods consist of EoG [Christopoulou *et al.*, 2019], DHG [Zhang *et al.*, 2020], GEDA [Li *et al.*, 2020], LSR [Nan *et al.*, 2020], HeterGSAN [Xu *et al.*, 2021c], SIRE [Zeng *et al.*, 2021], and SSAN [Xu *et al.*, 2021a]. Furthermore, for comparison purposes, we also include several recent competitive methods, namely DocuNet [Zhang *et al.*, 2021], KD-DocRE [Tan *et al.*, 2022a], and KMGRE [Jiang *et al.*, 2022].

Evaluation Metrics. Consistent with prior researches conducted [Yao *et al.*, 2019; Zhou *et al.*, 2021; Tan *et al.*, 2022b], we employ micro F1 and micro Ign F1 as our evaluation metrics. Ign F1 excludes the relational facts that are shared between the training and the development/test sets.

4.3 Main Results

In this part, we carry out a lot of comparative experiments with existing methods in terms of accuracy on F1 and Ign F1 evaluation metrics.

The results, presented in Table 1, indicate that our method universally outperforms its baselines on the two datasets. On

Model	Dev		Test	
	Ign F1	F1	Ign F1	F1
Ours	61.55±0.20	63.49±0.15	61.51	63.40
-Fusion	61.28±0.25	63.31±0.24	60.88	62.66
-Alignment	60.73±0.52	62.78±0.40	60.02	61.91
-Loss	58.88±0.14	60.97±0.15	59.95	61.82

Table 2: Ablation results on different modules in our method.

Model	DocRED			Re-DocRED		
	3%	5%	7%	3%	5%	7%
our method	63.40	**63.49**	63.45	76.18	**76.26**	76.22

Table 3: F1-score on the development set when tuning the probability filtering threshold δ.

Model	DocRED			Re-DocRED		
	1-L	2-L	3-L	1-L	2-L	3-L
our method	63.37	**63.49**	63.43	76.15	**76.26**	76.20

Table 4: F1-score on the development set with different GAT layers. L denotes layer.

DocRED, our method improves the baseline by +1.46 Dev F1, +1.32 Test F1, and also outperforms several previous studies including BERT-TS [Wang *et al.*,], CorefBERT [Ye *et al.*, 2020], LSR [Nan *et al.*, 2020], and SSAN [Xu *et al.*, 2021a]. On Re-DocRED, our method obtains improvements of 1.61 Dev F1 and 1.79 Test F1 points. These results demonstrate that our method can better capture the indivisibility among relations, entities, and contexts, and well handle the presence of the none class between entity pairs, thus improving the inference ability of the model.

Furthermore, our method surpasses recent state-of-the-art models such as DocuNet-BERT and KD-DocRE-BERT which incorporate dependencies between entity pairs to enhance their reasoning capabilities. This further attests to the exceptional reasoning prowess exhibited by our method.

4.4 Ablation Studies

Effect of Modules. We validate the design of the modules in our method. Table 2 shows that the fusion module for representation, the alignment module, and the none class ranking loss all contribute to the improvement. Furthermore, it can be observed that our method exhibits higher performance consistency across multiple iterations, which is influenced by the effectiveness of our alignment function. Notably, upon its removal, the standard deviation increases from ±0.25 to ±0.52.

Matrix Construction Threshold. We investigate the effect of key components in the relation correlation module. As shown in Table 3, we analyze the effect of probability filtering threshold δ in equation (3). It can be observed that the highest F1 score when $\delta = 0.05$ for all experiments. Besides, the results indicate that $\delta = 0.03$ will lead to more performance degradation compared with $\delta = 0.07$. The above phenomenon is due to the smaller threshold value resulting in more noise edges.

Model	DocRED				Re-DocRED			
	Dev		Test		Dev		Test	
	Ign F1	F1	Ign F1	F1	Ign F1	F1	Ign F1	F1
BCE	60.27	62.28	60.14	62.19	74.28	74.97	74.12	74.88
ATL	60.43	62.42	60.40	62.38	74.53	75.20	74.67	75.29
NoneClass	**61.55**	**63.49**	**61.51**	**63.40**	**75.54**	**76.26**	**75.70**	**76.30**

Table 5: Accuracy comparison of different losses in our method.

Number of GAT Layers. We report the results of different GAT layers with two heads in Table 4. Results demonstrate that 1-layer and 2-layer GAT networks achieve relatively similar results, while 3-layer GAT networks lead to greater performance degradation. The probable reason for the performance degradation is the over-smoothing issue that the node feature vectors are inclined to converge to comparable values.

Different Losses. To study how none class instances affect the performance of the none class ranking loss, we conduct experiments on our method with different losses. We still use the BERT-base for text encoding. Table 5 shows the accuracy comparison results. As can be seen, ATL and the none class ranking loss are better than BCE. On the other hand, due to the tuned threshold, the none class ranking loss obtains better accuracy than ATL when facing none class instances. Different from sentence-level datasets, document-level RE datasets contain many none class instances, which can be utilized by the none class ranking loss for better classification performance.

5 Conclusion and Future Work

In this paper, we propose a document-level RE model consisting of a fusion module for representation, an alignment module and a none class ranking loss. In particular, the fusion module for representation includes two parts: 1) the relation correlation part captures the correlations among relations; 2) the representation fusion part explicitly merges the representation of relations, entities, and contexts. Finally, the none class ranking loss maximizes label margins between none class and each predefined class, captures classless label dependencies, and enables adaptive thresholds for label predictions. Experimental results on two commonly used datasets show that our model outperforms all existing competing baselines. In the future, we consider applying the modules to other document-level tasks to verify their generality.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 42050102 and Dou Wanchun Expert Workstation of Yunnan Province (No. 202105AF150013). Dr Xuyun Zhang is the recipient of an ARC DECRA (project No. DE210101458) funded by the Australian Government.

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[Bishop, 1995] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[Chen and Badlani, 2020] Xiaoyu Chen and Rohan Badlani. Relation extraction with contextualized relation embedding (cre). In *Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–19, 2020.

[Christopoulou *et al.*, 2019] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4925–4936, 2019.

[Durand *et al.*, 2019] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019.

[Huang *et al.*, 2023] Heyan Huang, Changsen Yuan, Qian Liu, and Yixin Cao. Document-level relation extraction via separate relation representation and logical reasoning. *ACM Transactions on Information Systems*, 42(1):1–24, 2023.

[Jiang *et al.*, 2022] Feng Jiang, Jianwei Niu, Shasha Mo, and Shengda Fan. Key mention pairs guided document-level relation extraction. In *International Conference on Computational Linguistics*, pages 1904–1914, 2022.

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

[Li *et al.*, 2017] Yuncheng Li, Yale Song, and Jiebo Luo. Improving pairwise ranking for multi-label image classification. In *IEEE conference on computer vision and pattern recognition*, pages 3617–3625, 2017.

[Li *et al.*, 2020] Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. Graph enhanced dual attention network for document-level relation extraction. In *International conference on computational linguistics*, pages 1551–1560, 2020.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Lu *et al.*, 2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

[Nan *et al.*, 2020] Guoshun Nan, Zhijiang Guo, Ivan Sekulić, and Wei Lu. Reasoning with latent structure refinement for document-level relation extraction. In *Association for Computational Linguistics*, pages 1546–1557, 2020.

[Peng *et al.*, 2022] Xingyu Peng, Chong Zhang, and Ke Xu. Document-level relation extraction via subgraph reasoning. In *International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization*, pages 4331–4337, 2022.

[Ridnik *et al.*, 2021] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021.

[Soares *et al.*, 2019] Livio Baldini Soares, Nicholas Fitzgerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Association for Computational Linguistics*, pages 2895–2905, 2019.

[Tan *et al.*, 2022a] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Association for Computational Linguistics: ACL 2022*, pages 1672–1681, 2022.

[Tan *et al.*, 2022b] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting docred-addressing the false negative problem in relation extraction. In *Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, 2022.

[Tang *et al.*, 2020] Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. Hin: Hierarchical inference network for document-level relation extraction. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*, pages 197–209. Springer, 2020.

[Tucker and others, 1964] Ledyard R Tucker et al. The extension of factor analysis to three-dimensional matrices. *Contributions to mathematical psychology*, 110119, 1964.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[Wang *et al.*,] Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. Fine-tune bert for docred with two-step process.

[Wang *et al.*, 2020] Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. Global-to-local neural networks for

document-level relation extraction. In *Conference on Empirical Methods in Natural Language Processing*, pages 3711–3721, 2020.

[Wu *et al.*, 2019] Jiawei Wu, Wenhan Xiong, and William Yang Wang. Learning to learn and predict: A meta-learning approach for multi-label classification. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4354–4364, 2019.

[Wu *et al.*, 2020] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 162–178. Springer, 2020.

[Xie *et al.*, 2022] Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Association for Computational Linguistics*, pages 257–268, 2022.

[Xu *et al.*, 2019] Miao Xu, Yu-Feng Li, and Zhi-Hua Zhou. Robust multi-label learning with pro loss. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1610–1624, 2019.

[Xu *et al.*, 2021a] Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *AAAI conference on artificial intelligence*, volume 35, pages 14149–14157, 2021.

[Xu *et al.*, 2021b] Wang Xu, Kehai Chen, and Tiejun Zhao. Discriminative reasoning for document-level relation extraction. In *Association for Computational Linguistics*, pages 1653–1663, 2021.

[Xu *et al.*, 2021c] Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with reconstruction. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 14167–14175, 2021.

[Xu *et al.*, 2022] Benfeng Xu, Quan Wang, Yajuan Lyu, Yabing Shi, Yong Zhu, Jie Gao, and Zhendong Mao. Emrel: Joint representation of entities and embedded relations for multi-triple extraction. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 659–665, 2022.

[Xu *et al.*, 2023] Tianyu Xu, Jianfeng Qu, Wen Hua, Zhixu Li, Jiajie Xu, An Liu, Lei Zhao, and Xiaofang Zhou. Evidence reasoning and curriculum learning for document-level relation extraction. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[Yao *et al.*, 2019] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. Docred: A large-scale document-level relation extraction dataset. In *Association for Computational Linguistics*, pages 764–777, 2019.

[Ye *et al.*, 2020] Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. Coreferential reasoning learning for language representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 7170–7186, 2020.

[Yeh *et al.*, 2017] Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. Learning deep latent space for multi-label classification. In *AAAI conference on artificial intelligence*, volume 31, 2017.

[Yu *et al.*, 2018] Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *International conference on learning representations*, volume 2, 2018.

[Zeng *et al.*, 2020] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. Double graph based reasoning for document-level relation extraction. In *Conference on Empirical Methods in Natural Language Processing*, pages 1630–1640, 2020.

[Zeng *et al.*, 2021] Shuang Zeng, Yuting Wu, and Baobao Chang. Sire: Separate intra-and inter-sentential reasoning for document-level relation extraction. In *Association for Computational Linguistics*, pages 524–534, 2021.

[Zhang *et al.*, 2018] Yuhao Zhang, Peng Qi, and Christopher D Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, 2018.

[Zhang *et al.*, 2020] Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. Document-level relation extraction with dual-tier heterogeneous graph. In *International Conference on Computational Linguistics*, pages 1630–1641, 2020.

[Zhang *et al.*, 2021] Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. Document-level relation extraction as semantic segmentation. *arXiv preprint arXiv:2106.03618*, 2021.

[Zhang *et al.*, 2023] Liang Zhang, Zijun Min, Jinsong Su, Pei Yu, Ante Wang, and Yidong Chen. Exploring effective inter-encoder semantic interaction for document-level relation extraction. In *International Joint Conference on Artificial Intelligence*, pages 5278–5286, 2023.

[Zhou and Lee, 2022] Yang Zhou and Wee Sun Lee. None class ranking loss for document-level relation extraction. *arXiv preprint arXiv:2205.00476*, 2022.

[Zhou *et al.*, 2020] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

[Zhou *et al.*, 2021] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. In *AAAI conference on artificial intelligence*, volume 35, pages 14612–14620, 2021.