Document-Level Relation Extraction with Relation Correlation Enhancement

Yusheng Huang, Zhouhan Lin*

Shanghai Jiao Tong University, Shanghai, China
huangyusheng@sjtu.edu.cn, lin.zhouhan@gmail.com

Abstract. Document-level relation extraction (DocRE) is a task that focuses on identifying relations between entities within a document. However, existing DocRE models often overlook the correlation between relations and lack a quantitative analysis of relation correlations. To address this limitation and effectively capture relation correlations in DocRE, we propose a relation graph method, which aims to explicitly exploit the interdependency among relations. Firstly, we construct a relation graph that models relation correlations using statistical co-occurrence information derived from prior relation knowledge. Secondly, we employ a re-weighting scheme to create an effective relation correlation matrix to guide the propagation of relation information. Furthermore, we leverage graph attention networks to aggregate relation embeddings. Importantly, our method can be seamlessly integrated as a plug-and-play module into existing models. Experimental results demonstrate that our approach can enhance the performance of multi-relation extraction, highlighting the effectiveness of considering relation correlations in DocRE. [1]

忽略了关系之间的相关性，缺乏对关系相关性的定量分析

利用从先验关系知识中得到的统计共现信息

采用重加权方案，建立一个有效的关系相关矩阵来指导关系信息的传播

Keywords: Document-level relation extraction · Relation correlation · Relation graph construction

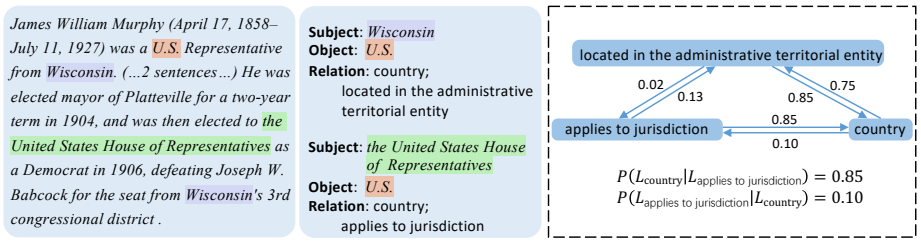
1 Introduction

Relation extraction (RE) plays a vital role in information extraction by identifying semantic relations between target entities in a given text. Previous research has primarily focused on sentence-level relation extraction, aiming to predict relations within a single sentence [7]. However, in real-world scenarios, valuable relational facts are often expressed through multiple mentions scattered across sentences, such as in Wikipedia articles [17]. Consequently, the extraction of relations from multiple sentences, known as document-level relation extraction, has attracted significant research attention in recent years.

Compared to sentence-level RE, document-level RE presents unique challenges in designing model structures. In sentence-level RE, a single relation type is associated with each entity pair, as observed in SemEval 2010 Task 8 [10] and TACRED [33]. However, in document-level RE, an entity pair can be associated

* Zhouhan Lin is the corresponding author.

[1] Codes are available at https://github.com/LUMIA-Group/LACE



(a) A document with entities. (b) Multi-label entity pairs. (c) Conditional probability between relations.

Fig. 1. Examples of relation correlation for multi-relation extraction. (a) presents a document containing multiple entities. (b) illustrates the multi-relation entity pairs. For instance, the subject entity *Wisconsin* and the object entity *U.S.* express the *country* and *located in the administrative territorial entity* relations. (c) demonstrates the conditional probabilities between three relations, which are derived from the DocRED dataset.

with multiple relations, making it more challenging than sentence-level RE. Figure 1(b) illustrates multi-relation examples extracted from the DocRED dataset [27], where each entity pair is associated with two distinct relations. Moreover, in document-level RE, the number of relation types to be classified can be large (e.g., 97 in the DocRED dataset), further increasing the difficulty of extracting multiple relations.

To address this challenge, previous studies have commonly approached it as a multi-label classification problem, where each relation is treated as a label. Binary cross-entropy loss is typically employed to handle this multi-label scenario [16,31]. During inference, a global threshold is applied to determine the relations. More recently, [14] utilize the asymmetric loss (ASL) [1] to mitigate the imbalance between positive and negative classes. Additionally, [32] propose to employ a balanced softmax method to mitigate the imbalanced relation distribution, where many entity pairs have no relation. [38] introduce the adaptive thresholding technique, which replaces the global threshold with a learnable threshold class. However, previous studies have rarely quantitatively analyzed the co-occurred relations and have not explicitly utilized this feature.

以往的研究以往的研究很少定量地分析共发生的关系，也没有明确地利用这一特征，也没有明确地利用这一特征

According to the statistics in DocRED dataset, we find that relations co-occur with priors. As illustrated in Figure 1(c), for entity pairs with multiple relations, the conditional probability of relation *country* appears given that relation *applies to jurisdiction* appears is 0.85, while the conditional probability of relation *applies to jurisdiction* appears given that relation *country* appears is 0.10. Besides, with great chance, relation *country* and relation *located in the administrative territorial entity* appear together. Considering the relations exhibit combinatorial characteristics, it is desirable to employ the relation correlations to ameliorate the model structure and boost the multi-relation extraction.

我们发现关系与先验相一致

考虑到这些关系的组合特征，我们希望采用关系相关性来改善模型结构，提高多关系提取能力

In this paper, we aim to tackle the challenge of multi-relation extraction in document-level RE by leveraging the correlation characteristics among relations. Specifically, we propose a relation graph method that leverages the prior knowledge of interdependency between relations to effectively guide the extraction

relation graph method

利用关系之间相互依赖的先验知识来有效地指导提取多个关系

为了模拟关系相关性，我们通过计算训练集中关系共现的频率来估计它

为了避免过拟合，我们过滤掉一定阈值以下的噪声边，并通过将 每个共现元素除以每个关系的出现数来创建一个条件概率矩阵

of multiple relations. To model the relation correlations, we estimate it by calculating the frequency of relation co-occurrences in the training set [24,3]. To avoid overfitting, we filter out noisy edges below a certain threshold and create a conditional probability matrix by dividing each co-occurrence element by the occurrence numbers of each relation. This matrix is then binarized to enhance the model's generalization capability, and the relation graph is constructed as a binary directed graph. Additionally, we employ a re-weighting scheme to construct an effective relation correlation matrix, which guides the propagation of relation information [5]. We employ Graph Attention Networks (GAT) [21] with the multi-head graph attention to aggregate relation embeddings. Based on the adaptive thresholding technique [38], the loss function in our method is also amended by emphasizing the multi-relation logits. Our method is easy for adoption as it could work as a plug-in for existing models. We conduct extensive experiments on the widely-used document-level RE dataset DocRED, which contains around 7% multi-relation entity pairs. Experimental results demonstrate the effectiveness of our method, achieving superior performance compared to baseline models. In summary, our contributions are as follows:

对该条件概率矩阵进行二值化，以增强模型的泛化能力，并将关系图构造为二值有向图。

我们采用一个重加权方案来构造一个有效的关系相关矩阵，以指导关系信息的传播

我们采用具有多头图注意的图注意网络（GAT）[21]来聚合关系嵌入。基于自适应阈值技术[38]，对损失函数进行了修正。

- We conduct comprehensive quantitative studies on relation correlations in document-level RE, providing insights for addressing the challenge of multi-relation extraction.
- We propose a relation graph method that explicitly leverages relation correlations, offering a plug-in solution for other effective models.
- We evaluate our method on a large-scale DocRE dataset, demonstrating its superior performance compared to baselines.

2 Related Work

Relation extraction, a crucial task in natural language processing, aims to predict the relations between two entities. It has widespread applications, including dialogue generation [9] and question answering [11]. Previous researches largely focus on sentence-level RE, where two entities are within a sentence. Many models have been proposed to tackle the sentence-level RE task, encompassing various blocks such as CNN [30,19], LSTM [37,2], attention mechanism [23,29], GNN [8,39], and transformer [26,4].

Recent researches work on document-level relation extraction since many real-world relations can only be extracted from multiple sentences [27]. From the perspective of techniques, various related approaches could be divided into the graph-based category and the transformer-based category. For the graph-based models that are advantageous to relational reasoning, [16] propose LSR that empowers the relational reasoning across multiple sentences through automatically inducing the latent document-level graph. [31] propose GAIN with two constructed graphs that captures complex interaction among mentions and entities. [35] propose GCGCN to model the complicated semantic interactions among multiple entities. [12] propose to characterize the complex interaction between multiple sentences and the possible relation instances via GEDA networks. [34]

introduce DHG for document-level RE to promote the multi-hop reasoning. For the transformer-based models that are capable of implicitly model long-distance dependencies, [22] discover that using pre-trained language models can improve the performance of this task. [20] propose HIN to make full use of the abundant information from entity level, sentence level and document level. [28] present CorefBERT to capture the coreferential relations in context. Recent works normally directly leverage the pre-trained language models such as BERT [6] or RoBERTa [15] as word embeddings.

3 Methodology

In this section, we provide a detailed explanation of our LAbel Correlation Enhanced (LACE) method, for document-level relation extraction. We begin by formulating the task in §3.1 and then introduce the overall architecture in §3.2. In §3.3, we discuss the encoder module for obtaining the feature vectors of entity pairs. The relation correlation module, outlined in §3.4, is designed to capture relation correlations. Finally, we present the classification module with multi-relation adaptive thresholding loss for model optimization in §3.5.

3.1 Task Formulation

Given an input document that consists of N entities $\mathcal{E} = \{e_i\}_{i=1}^{N}$, this task aims to identity a subset of relations from $\mathcal{R} \cup \{\text{NA}\}$ for each entity pair (e_s, e_o), where $s, o = 1, ..., N; s \neq o$. The first entity e_s is identified as the *subject* entity and the second entity e_o is identified as the *object* entity. $\mathcal{R}$ is a pre-defined relation type set, and NA denotes no relation expressed for the entity pair. Specifically, an entity e_i can contain multiple mentions with different surface names $e_i^k, k = 1, ..., m$. During testing, the trained model is supposed to predict labels of all the entity pairs $(e_s, e_o)_{s,o=1,...,N;s\neq o}$ within documents.

3.2 Overall Architecture

As illustrated in Figure 2, the overall architecture consists of three modules. The encoder module first yields the contextual embeddings of all the entity mentions, and then each entity embedding is obtained by integrating information from the corresponding entity mentions, i.e. surface name merging. Afterward, entity pair features are calculated to enhance the entity pair embedding. The relation correlation module generates relation feature vectors. The correlation matrix is built in a data-driven manner, which is based on the statistics of the provided training set. We employ the edge re-weighting scheme to create a weighted adjacency matrix, which is beneficial for deploying graph neural networks. GAT is applied to the correlation matrix and relation features to generate more informative relation feature vectors. In the classification module, a bi-linear layer is utilized for prediction. Besides, based on the adaptive thresholding technique, we resort to a refined loss function for better multi-label classification.

Encoder Module
Relation Correlation Module
Classification Mudule

correlation matrix以数据驱动的方式构建，它是基于所提供的训练集的统计数据

我们采用边缘重加权方案来创建一个加权邻接矩阵weighted adjacency matrix，这有利于图神经网络的部署。

将GAT应用于correlation matrix和关系特征，生成信息更丰富的关系特征向量

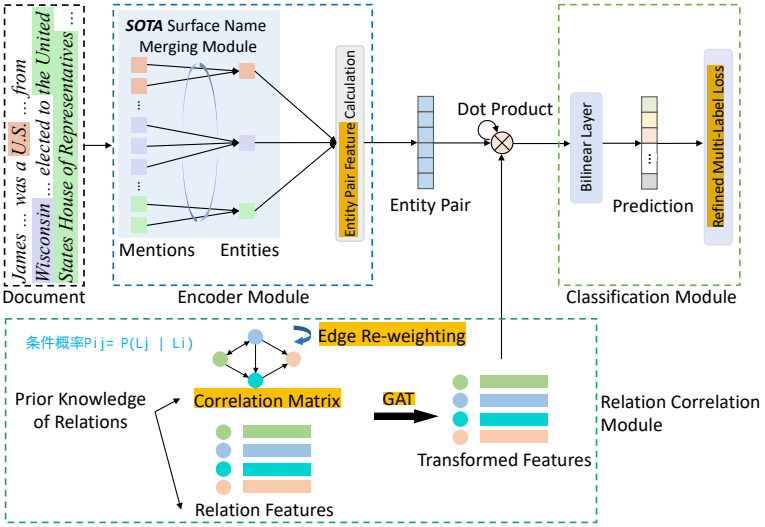


Fig. 2. Architecture of our LACE method, consisting of three modules: the encoder module, relation correlation module, and the classification module.

3.3 Encoder Module

For a document consisting of l words $[x_t]_{t=1}^{l}$ and N entities (each entity containing several mentions), we obtain the word embedding x_i^w and entity type embedding x_i^t. Then we concatenate them and feed them to BiLSTM layers to generate the contextualized input representations $\boldsymbol{h}_i$:

$$\boldsymbol{h}_i = \text{BiLSTM}([x_i^w; x_i^t]). \tag{1}$$

Mention representations m_i are obtained by conducting a max-pooling operation on the words, and entity representations $\boldsymbol{E}_i$ are generated by the log-sum-exp pooling over all the entity mention representations m_i:

$$\boldsymbol{E}_i = \log \sum_{i=1}^{j} \exp\left(m_i\right), \tag{2}$$

where j is the number of entity mentions.

In this way, we can generate the embeddings of the head entity and tail entity, denoted as $\boldsymbol{E}_s \in \mathbb{R}^{d_B}$ and $\boldsymbol{E}_o \in \mathbb{R}^{d_B}$, respectively. The entity pair features are obtained by concatenating these embeddings.

Many studies fall in this module, which focus on generating more contextual representations by leveraging Transformer [38] or document graphs [1]. These studies can be seamlessly integrated into our LACE method, or conversely, LACE can be incorporated as a plug-in for other models.

用条件概率计算Correlation Matrix Pij = P (Lj | Li)
由于有些关系和其他关系很少一起出现导致共发生关系的概率值很大；训练测试统计量偏差

使用一个阈值二值化条件矩阵，得到二值化相关矩阵B
由于图神经网络的过平滑问题，让邻居带有权重

从B中得到重新加权的关系相关矩阵 R，之后应用GAT

3.4 Relation Correlation Module

我们以条件概率的形式建立关系相关相互依赖关系

P(Lb | La) 表示关系类型 La 出现时关系类型 Lb 出现的概率。

考虑到对于条件概率 P(Lb | La) 与 P(La | Lb) 不相等，我们构建了一个基于训练集的关系先验知识的关系相关有向图用于建模，这意味着邻接矩阵是不对称的。

We model the relation correlation interdependency in the form of conditional probability, i.e., $P(L_b \mid L_a)$ means the probability of occurrence of relation type L_b when relation type L_a appears. Considering that for conditional probabilities $P(L_b \mid L_a) \neq P(L_a \mid L_b)$, we construct a relation-related directed graph based on the relation prior knowledge of the training set for modeling, which means that the adjacency matrix is asymmetric.

Correlation Matrix Construction：
1. 首先计算训练集中关系的共现情况，得到共现矩阵Cr×r - r是关系

2. 为了得到关系之间的条件概率，将共现矩阵Cr×r中的每个元素 除以 关系共现的总数，得Pij = P (Lj | Li)

Correlation Matrix Construction To construct the correlation matrix, we first count the co-occurrence of relations in the training set and obtain the co-occurrence matrix $\boldsymbol{C}^{r\times r}$, where r is the number of pre-defined relation types. To obtain the conditional probabilities between relations, each element in the co-occurrence matrix $\boldsymbol{C}^{r\times r}$ is divided by the total number of relation co-occurrences, i.e.,

conditional probabilities P

$$P_{ij} = C_{ij} / \sum_{j} C_{ij}, \tag{3}$$

where $P_{ij} = P(L_j \mid L_i)$ denotes the probability of relation type L_j when relation type L_i appears.

上述构造相关矩阵的缺点：

1. 有些关系很少与其他关系一起出现。这将导致共发生关系的概率值很大，这是不合理的。

2. 训练数据集的统计量与测试数据集的统计量之间可能存在偏差。

因此，为了缓解这些问题：

However, the above method for correlation matrix construction may suffer two drawbacks. Firstly, some relations rarely appear together with others. This will lead to a large probability value for the co-occurred relation, which is unreasonable. Secondly, there may be a deviation between the statistics of the training dataset and the statistics of the test dataset. Using the exact numbers tend to overfit the training dataset, which might hurt the generalization capacity. Therefore, to alleviate these issues, we set a threshold τ to filter these rare co-occurred relations. Then we binarize the conditional probability matrix $\boldsymbol{P}$ by

设置了一个阈值 τ 来过滤这些罕见的共发生关系
然后二值化条件概率矩阵P得到二值化相关矩阵B

二值化条件概率矩阵P

$$B_{ij} = \begin{cases} 0, & \text{if } P_{ij} < \delta \\ 1, & \text{if } P_{ij} \geq \delta \end{cases}, \tag{4}$$

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

使用二值化相关矩阵B 的问题：

在图神经网络中使用二值相关矩阵B的一个问题是过平滑问题[36]，即节点向量倾向于收敛于相似的值。关系特征与其相邻节点之间没有权重差别。

为了缓解这一问题，我们采用了以下重新加权方案
从B中得到重新加权的关系相关矩阵 R

Edge Re-weighting Scheme One concern for utilizing the binary correlation matrix $\boldsymbol{B}$ for graph neural networks is the over-smoothing issue [36] that the node attribute vectors tend to converge to similar values. There is no natural weight difference between the relation features and its neighbor nodes'. To mitigate this issue, we employ the following re-weighting scheme,

$$R_{ij} = \begin{cases} p / \sum_{\substack{j=1 \\ i\neq j}}^{r} B_{ij}, & \text{if } i \neq j \\ 1 - p, & \text{if } i = j \end{cases}, \tag{5}$$

where $\boldsymbol{R}$ is the re-weighted relation correlation matrix and p is a hyper-parameter. 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.

在训练过程中将关系特征以及其邻居的固定权重应用于模型中。这有助于缓解某些问题

使用邻接矩阵构图，之后进行GAT计算

关系特征是用与单词嵌入相同的方式得到的嵌入向量。

Relation features are the embedding vectors obtained in the same way as word embeddings. We then exploit GAT networks with a K-head attention mechanism to aggregate relation features for two reasons. First, GAT is suitable for directed graphs. Second, GAT maintains a stronger representation ability since the weights of each node can be different. The transformed features by GAT are denoted as $\boldsymbol{R} \in \mathbb{R}^{r \times d_B}$.

利用具有k-头注意机制的GAT网络来聚合关系特征，原因有两个：
1. GAT适用于有向图。
2. 由于每个节点的表示能力更强，GAT的权重可能不同。

将GAT变换后的特征记为R∈Rr×dB。

3.5 Classification Module

Given feature vectors $\boldsymbol{E}_s, \boldsymbol{E}_o \in \mathbb{R}^{d_B}$ of the entity pair (e_s, e_o) and the transformed relation features $\boldsymbol{R} \in \mathbb{R}^{r \times d_B}$, we map them to hidden representations $\boldsymbol{I_s}, \boldsymbol{I_o} \in \mathbb{R}^r$ followed by the layer normalization operation,

$$\boldsymbol{I_s} = \text{LayerNorm}\ (\boldsymbol{R} \cdot \boldsymbol{E}_s), \tag{6}$$

$$\boldsymbol{I_o} = \text{LayerNorm}\ (\boldsymbol{R} \cdot \boldsymbol{E}_o). \tag{7}$$

Then, we obtain the prediction probability of the relation r' via a bilinear layer,

关系的预测概率Pr'

$$\mathrm{P}_{r'} = \sigma\left((\boldsymbol{E}_s \oplus \boldsymbol{I_s})^\top W_r (\boldsymbol{E}_o \oplus \boldsymbol{I_o}) + b_r\right), \tag{8}$$

where σ is the sigmoid activation function. $W_r \in \mathbb{R}^{(d_B+r)\times(d_B+r)}$, $b_r \in \mathbb{R}$ are model parameters, and $\oplus$ denotes the concatenation operation.

Previous study [38] has shown the effectiveness of the Adaptive Thresholding loss (AT loss), where a threshold class is set such that logits of the positive classes are greater than the threshold class while the logits of the negative classes are less than the threshold class. However, their designed loss function does not quite match the multi-label problem, since they implicitly use the softmax function in the calculation of the positive-class loss function. Therefore, during each loss calculation, the AT loss is unable to extract multiple relations. The superposition of multiple calculations would result in a significant increase in time overhead. To mitigate this issue, we propose a novel loss function called Multi-relation Adaptive Thresholding loss (MAT loss), which is defined as follows,

AT Loss损失函数并不完全匹配多标签问题，因为他们在计算正类损失函数时隐式地使用了softmax函数。因此，在每次损失计算过程中，AT损失都无法提取出多重关系

多个计算的叠加将导致时间开销的显著增加。

为了缓解这一问题，我们提出了一种新的损失函数，称为多关系自适应阈值损失（MAT损失），其定义如下：

$$\mathcal{L}_+ = -\log(1 - \mathrm{P}(\mathtt{TH})) - \sum_{r' \in L_p} (y^{r'} \log \mathrm{P}(r') + \left(1 - y^{r'}\right) \log(1 - \mathrm{P}(r'))), \tag{9}$$

$$\mathcal{L}_- = -\log\left(\frac{\exp\left(\text{logit}_{\mathtt{TH}}\right)}{\sum_{r' \in L_o \cup \{\mathtt{TH}\}} \exp\left(\text{logit}_{r'}\right)}\right), \tag{10}$$

当处理多标签问题时，一个样本可能同时属于多个类别，因此多个类别应该具有较高的概率。使用softmax会强制这些概率相互抵消，使得只有一个类别获得高概率，而其他类别的概率会降低。这就是为什么在多标签问题中使用softmax的损失函数会有问题。

where the threshold class TH is the NA class. L_p and L_o denote the relations the exist and do not exist between the entity pair, respectively. logit means the number without σ in Equation 8. The final loss function is $\mathcal{L} = \alpha\mathcal{L}_+ + (1-\alpha)\mathcal{L}_-$, where α is a hyper-parameter. In this way, our MAT loss enables the extraction of multiple relations.

During inference, we assign labels to entity pairs whose prediction probabilities meet the following criteria,

$$\mathrm{P}\left(r' \mid e_s, e_o\right) \geq (1+\theta)\ \mathrm{P}(\mathtt{TH}), \tag{11}$$

where θ is a hyper-parameter that maximizes evaluation metrics.

θ 是一个可以最大化评估指标的超参数。

4 Experiments

4.1 Dataset

We evaluate our proposed approach on a large-scale human-annotated dataset for document-level relation extraction DocRED [27], which is constructed from Wikipedia articles. DocRED is larger than other existing counterpart datasets in aspects of the number of documents, relation types, and relation facts. Specifically, DocRED contains 3053 documents for the training set, 1000 documents for the development set, and 1000 documents for the test set, with 96 relation types and 56354 relational facts. For entity pairs with relations, around 7% of them express more than one relation type, and an entity pair can express up to 4 relations. [2]

4.2 Implementation Details

We employ GloVe [18] and BERT-based-cased [6] word embeddings in the encoder module, respectively. When employing GloVe word embeddings, we use Adam optimizer with learning rate being e^{-3}. When employing BERT-based-cased, we use AdamW with a linear warmup for the first 6% steps. 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 4. We set p to be 0.3 in Equation 5 and θ to be 0.85 in Equation 11. 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. α in the classification module is 0.4. All hyper-parameters are tuned on the development set.

4.3 Baseline Systems

We compare our approach with the following models, including three categories.

GloVe-based Models. These models report results using GloVe word embeddings and utilize various neural network architectures including CNN, BiLSTM, and Context-Aware [27], to encode the entire document, and then obtain the embeddings of entity pairs for relation classification. The recent mention-based-reasoning model MRN [1] also present the results with GloVe word embedding.

Transformer-based Models. These models directly exploit the pre-trained language model BERT for document encoding without document graph construction, including HIN-BERT [20], CorefBERT [28], and ATLOP-BERT [38]. We mainly compare our method LACE with ATLOP model that aims to mitigate the multi-relation problem.

[2] We conduct no experiments on the CDR [13] and GDA [25] datasets in the biomedical domain, because they do not suffer the multi-relation issue. Therefore, they do not match our scenario.

Graph-based Models. Homogeneous or heterogeneous graphs are constructed based on the document features for reasoning. Then, various graph-based models are leveraged to perform inference on entity pairs, including BiLSTM-AGGCN [8], LSR-BERT [16], GAIN-BERT [31]. The MRN-BERT [1] aims to capture the local and global interactions via multi-hop mention-level reasoning.

When compared to GloVe-based models and graph-based models, we integrate the MRL layer from MRN into the encoder module. When compared to Transformer-based models, we incorporate the localized context pooling technique from ATLOP into the encoder module.

Table 1. Results on the development set and test set of DocRED.

Model	**Dev**		**Test**	
	Ign F_1	F_1	Ign F_1	F_1
With GloVe				
CNN [27]	41.58	43.45	40.33	42.26
BiLSTM [27]	48.87	50.94	48.78	51.06
Context-Aware [27]	48.94	51.09	48.40	50.70
MRN [14]	56.62	58.59	56.19	58.46
LACE	**57.01**	**58.92**	**56.61**	**58.64**
With BERT+Transformer				
HIN-BERT [20]	54.29	56.31	53.70	55.60
CorefBERT [28]	55.32	57.51	54.54	56.96
ATLOP-BERT [38]	59.22	61.09	59.31	61.30
LACE-BERT	**59.58**	**61.43**	**59.40**	**61.50**
With BERT+Graph				
BiLSTM-AGGCN [8]	46.29	52.47	48.89	51.45
LSR-BERT [16]	52.43	59.00	56.97	59.05
GAIN-BERT [31]	59.14	61.22	59.00	61.24
MRN-BERT [14]	59.74	61.61	59.52	61.74
LACE-MRL-BERT	**59.98**	**61.75**	**59.85**	**61.90**

4.4 Quantitative Results

Table 1 shows the experimental results on the DocRED dataset. Following previous studies [27,38], we adopt the Ign F_1 and F_1 as the evaluation metrics, where Ign F1 is calculated by excluding the shared relation facts between the training set and development/test set.

For the GloVe-based models, our method LACE achieves 56.61% Ign F_1 and 58.64% F_1-score on the test set, outperforming all other methods. For the transformer-based models using BERT, our method LACE-BERT achieves 61.50% F1-score on the test set, which outperforms the ATLOP-BERT model.

These experimental results also show that the pre-trained language model can cooperate well with the LACE method. For the graph-based models, we chieve 61.90% F_1-score on the test set. The result demonstrates that capturing the mention-level contextual information is helpful and our proposed method could work well with the mention-based reasoning method. Overall, results demonstrate the effectiveness of leveraging the relation information.

4.5 Analysis of Relation Correlation Module

We investigate the effect of key components in the relation correlation module.

Matrix Construction Threshold. As shown in Table 2, we analyze the effect of probability filtering threshold δ in Equation 4. We obtain the highest F_1 score when δ equals 0.05 for all experiments. Besides, results indicates that $\delta = 0.03$ will lead to more performance degradation compared with $\delta = 0.07$. We believe that this is due to the smaller threshold value resulting in more noise edges.

Table 2. F1-score on the development set when tuning the probability filtering threshold δ.

Model	3%	5%	7%
LACE	58.74	**58.92**	58.82
LACE-BERT	61.38	**61.43**	61.40
LACE-MRL-BERT	61.67	**61.75**	61.71

Table 3. F1-score on the development set with different GAT layers. L denotes layer.

Model	1-L	2-L	3-L
LACE	58.80	**58.92**	58.64
LACE-BERT	61.40	**61.43**	61.23
LACE-MRL-BERT	61.69	**61.75**	61.63

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

Table 4. F1-score for multi-relation extraction on the development set. *Rel* denotes relations.

Model	2-*Rel*	3-*Rel*	Overall
ATLOP-BERT	40.13	29.59	39.62
LACE-BERT	**42.03**	**32.58**	**41.55**

4.6 Performance on Multi-Label Extraction

In order to evaluate the performance of multi-label extraction, we re-implement ATLOP-BERT model and report the experimental results of multi-relation extraction as shown in Table 4. As seen, our approach LACE-BERT gains 1.9% and 2.99% F1-score improvement on the 2-relation and 3-relation extraction, respectively, which demonstrates the effectiveness of leveraging the relation correlations. Overall, our approach achieves 1.93% F1-score improvements on multi-label extraction compared with ATLOP-BERT.

Table 5. F1-score on the development set for ablation study. RCM denotes the relation correlation module.

Model	2-*Relation*	3-*Relation*	Overall
LACE-BERT	**42.03**	**32.58**	**41.55**
- RCM	40.74	30.62	40.25
- $\mathcal{L}_{MAT}$	41.43	31.78	40.94

4.7 Ablation Study

We conduct ablation studies to verify the necessity of two critical modules in LACE-BERT for multi-relation extraction as depicted in Table 5. Results show that two modules contribute to the final improvements. Firstly, removing the relation correlation module causes more performance degradation, and we thus believe that leveraging the prior knowledge of relation interdependency is helpful for the multi-relation extraction. Secondly, we replace our multi-relation adaptive thresholding loss $\mathcal{L}_{MAT}$ with the adaptive thresholding loss [38] for comparison. We believe that the reason for the improvement is that the MAT loss enlarges the margin values between all the positive classes and the threshold class.

MAT损失增大了所有正类和阈值类之间的边际值

4.8 Case Study

Figure 3 shows a case study of our proposed approach LACE-BERT, in comparison with ATLOP-BERT baseline. We can observe that ATLOP-BERT can only identify the *P17* and *P131* relations for the entity pair *(Ontario, Canada)*,

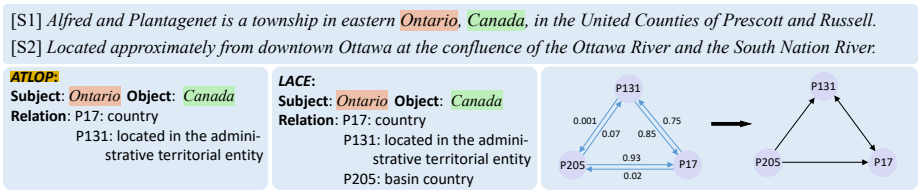


Fig. 3. Case study of a triple-relation entity pair from the development set of DocRED. We visualize the conditional probabilities among these relations and exhibit the constructed directed sub-graph.

where the two relations frequently appear together. However, ATLOP-BERT fails to identity the *P205* relation, while LACE-BERT deduces this relation. By introducing the label correlation matrix, this relation *P205* establishes connections with other relations with high conditional probabilities, which is advantageous for multi-relation extraction.

该关系P205与其他具有高条件概率的关系建立联系，有利于多关系提取

5 Conclusion

In this work, we propose our method LACE for document-level relation extraction. LACE includes a relation graph construction approach which explicitly leverages the statistical co-occurrence information of relations. Our method effectively captures the interdependency among relations, resulting in improved performance on multi-relation extraction. Experimental results demonstrate the superior performance of our proposed approach on a large-scale document-level relation extraction dataset.

Acknowledgements The authors would like to thank the support from the National Natural Science Foundation of China (NSFC) grant (No. 62106143), and Shanghai Pujiang Program (No. 21PJ1405700).

References

1. Baruch, E.B., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. CoRR (2020)
2. Cai, R., Zhang, X., Wang, H.: Bidirectional recurrent convolutional neural network for relation classification. In: Proc. of ACL (2016)
3. Che, X., Chen, D., Mi, J.: Label correlation in multi-label classification using local attribute reductions with fuzzy rough sets. FSS (2022)
4. Chen, M., Lan, G., Du, F., Lobanov, V.S.: Joint learning with pre-trained transformer on named entity recognition and relation extraction tasks for clinical analytics. In: ClinicalNLP@EMNLP 2020, Online, November 19, 2020 (2020)
5. Chen, Z., Wei, X., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: CVPR (2019)

6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proc. of ACL (2019)
7. Feng, J., Huang, M., Zhao, L., Yang, Y., Zhu, X.: Reinforcement learning for relation classification from noisy data. In: Proc. of AAAI (2018)
8. Guo, Z., Zhang, Y., Lu, W.: Attention guided graph convolutional networks for relation extraction. In: Proc. of ACL (2019)
9. He, H., Balakrishnan, A., Eric, M., Liang, P.: Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In: Proc. of ACL (2017)
10. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Séaghdha, D.Ó., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: Semeval-2010 task 8. In: SEW@NAACL-HLT 2009, Boulder, CO, USA, June 4, 2009 (2009)
11. Hixon, B., Clark, P., Hajishirzi, H.: Learning knowledge graphs for question answering through conversational dialog. In: ACL (2015)
12. Li, B., Ye, W., Sheng, Z., Xie, R., Xi, X., Zhang, S.: Graph enhanced dual attention network for document-level relation extraction. In: Proc. of COLING (2020)
13. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: Biocreative V CDR task corpus: a resource for chemical disease relation extraction. Database J. Biol. Databases Curation **2016** (2016)
14. Li, J., Xu, K., Li, F., Fei, H., Ren, Y., Ji, D.: MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In: Proc. of ACL (2021)
15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR (2019)
16. Nan, G., Guo, Z., Sekulic, I., Lu, W.: Reasoning with latent structure refinement for document-level relation extraction. In: ACL (2020)
17. Peng, N., Poon, H., Quirk, C., Toutanova, K., Yih, W.: Cross-sentence n-ary relation extraction with graph lstms. TACL (2017)
18. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543. ACL (2014)
19. dos Santos, C.N., Xiang, B., Zhou, B.: Classifying relations by ranking with convolutional neural networks. In: ACL (2015)
20. Tang, H., Cao, Y., Zhang, Z., Cao, J., Fang, F., Wang, S., Yin, P.: HIN: hierarchical inference network for document-level relation extraction. In: Proc. of KDD (2020)
21. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: ICLR (2018)
22. Wang, H., Focke, C., Sylvester, R., Mishra, N., Wang, W.Y.: Fine-tune bert for docred with two-step process. CoRR (2019)
23. Wang, L., Cao, Z., de Melo, G., Liu, Z.: Relation classification via multi-level attention cnns. In: Proc. of ACL (2016)
24. Wang, Y., He, D., Li, F., Long, X., Zhou, Z., Ma, J., Wen, S.: Multi-label classification with label graph superimposing. In: Proc. of AAAI (2020)
25. Wu, Y., Luo, R., Leung, H.C.M., Ting, H., Lam, T.W.: RENET: A deep learning approach for extracting gene-disease associations from literature. In: RECOMB 2019, Washington, DC, USA, May 5-8, 2019, Proceedings (2019)
26. Xiao, Y., Tan, C., Fan, Z., Xu, Q., Zhu, W.: Joint entity and relation extraction with a hybrid transformer and reinforcement learning based model. In: Proc. of AAAI (2020)

27. Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., Sun, M.: Docred: A large-scale document-level relation extraction dataset. In: ACL (2019)
28. Ye, D., Lin, Y., Du, J., Liu, Z., Li, P., Sun, M., Liu, Z.: Coreferential reasoning learning for language representation. In: Proc. of EMNLP (2020)
29. Ye, Z., Ling, Z.: Distant supervision relation extraction with intra-bag and inter-bag attentions. In: Proc. of ACL (2019)
30. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proc. of COLING (2014)
31. Zeng, S., Xu, R., Chang, B., Li, L.: Double graph based reasoning for document-level relation extraction. In: EMNLP (2020)
32. Zhang, N., Chen, X., Xie, X., Deng, S., Tan, C., Chen, M., Huang, F., Si, L., Chen, H.: Document-level relation extraction as semantic segmentation. In: Proc. of IJCAI (2021)
33. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware attention and supervised data improve slot filling. In: Proc. of EMNLP (2017)
34. Zhang, Z., Yu, B., Shu, X., Liu, T., Tang, H., Wang, Y., Guo, L.: Document-level relation extraction with dual-tier heterogeneous graph. In: Proc. of COLING (2020)
35. Zhou, H., Xu, Y., Yao, W., Liu, Z., Lang, C., Jiang, H.: Global context-enhanced graph convolutional networks for document-level relation extraction. In: COLING (2020)
36. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI Open (2020)
37. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proc. of ACL (2016)
38. Zhou, W., Huang, K., Ma, T., Huang, J.: Document-level relation extraction with adaptive thresholding and localized context pooling. In: Proc. of AAAI (2021)
39. Zhu, H., Lin, Y., Liu, Z., Fu, J., Chua, T., Sun, M.: Graph neural networks with generated parameters for relation extraction. In: Proc. of ACL (2019)