

Document-Level Relation Extraction with Global Relations and Entity Pair Reasoning

Fu Zhang^{†*}, Yi Yan[†], Jingwei Cheng

School of Computer Science and Engineering, Northeastern University, China
zhangfu@mail.neu.edu.cn; yanyi_neu@163.com

Abstract

Document-level relation extraction (DocRE) aims to extract structured relational triples from unstructured text based on given entities. Existing methods are mainly categorized into transformer-based models and graph-based models. While transformer-based models capture global contextual information, they typically focus on individual entity pairs, making it challenging to capture complex interactions between multiple entity pairs. Graph-based models build document graphs using entities or sentences as nodes for reasoning but often lack explicit mechanisms to model fine-grained interactions between entity pairs, limiting their ability to handle complex relational reasoning tasks. Additionally, previous research has not considered predicting all possible relations in advance to assist with DocRE tasks. To address these issues, we propose a new framework namely **GREP** (global relations and entity pair reasoning) for DocRE tasks. GREP leverages the global interdependencies between entity pairs to capture fine-grained interactions and perform multi reasoning at the entity pair level. In addition, GREP for the first time proposes an auxiliary task that predicts all possible relations in advance that exist in a document, which enables the model to filter out the most unlikely relations. Experimental results on widely-used datasets demonstrate that our model achieves state-of-the-art performance¹.

1 Introduction

Document-level relation extraction (DocRE) aims to extract structured relation triples from unstructured text based on given entities. Early studies primarily focused on sentence-level relation extraction (Zhang et al., 2018; Zhu et al., 2019; Sun et al., 2020), which predicts relations between entities within a single sentence. However, DocRE is more


[1]Elizabeth II was Queen of Mauritius from 1968 to 1992 . [3]The Queen was also the monarch of the United Kingdom and the other Commonwealth realms . [8]The Queen and her husband Prince Philip , Duke of Edinburgh , visited Mauritius for three days ...		
Reasoning triples in simple scene (United Kingdom, member of, Commonwealth) (Commonwealth, chairperson, Elizabeth II) (Philip, country of citizenship, United Kingdom) ...		
Entity pair graph 	Reasoning results in complex scene (United Kingdom, ?, Elizabeth II) relation: head of state (Philip, ?, Elizabeth II) relation: spouse ...	Global relations spouse (✓) member of (✓) head of state (✓) mother (✗) ...

Figure 1: An example of DocRE. The reasoning in complex multi-hop scenes can be achieved based on our proposed entity pair graph and global relations.

challenging, as many relations span multiple sentences and require multi-step reasoning.

Currently, two main approaches are employed to tackle DocRE: transformer-based models and graph-based models (Delaunay et al., 2023). Transformer-based models leverage self-attention mechanisms to capture global contextual information and are adept at handling long-distance dependencies (Zhou et al., 2021; Xie et al., 2022). Graph-based models construct document graphs by treating entities, mentions, or sentences as nodes (Zeng et al., 2020; Lu et al., 2023) and employ graph convolutional network (GCN) (Scarselli et al., 2008) for reasoning on the graphs. However, while these methods implicitly model interactions between entities, they often ignore direct interactions between entity pairs and lack an explicit mechanism to model such interactions, which introduces limitations in complex relational reasoning tasks.

As a multi-label classification task, identifying the relations between entities in DocRE usually requires one-hop or multi-hop reasoning on entity pairs, as shown in Figure 1. For instance, inferring the relation between *Philip* and the *Elizabeth II* is complex since the document does not explicitly ex-

¹Our code: <https://github.com/yanyi74/GREP>.

[†]Equal contribution. *Corresponding author.

press their relationship. The model needs to reason over the multi-hop relation paths between (*United Kingdom, member of, Commonwealth*) and (*Commonwealth, chairperson, Elizabeth II*) to identify the relation (*United Kingdom, head of state, Elizabeth II*). Further, by combining the other triples in Figure 1, the model can finally deduce that the relation between *Philip* and *Elizabeth II* is “*spouse*”. By constructing a graph based on entity pairs in Figure 1, the reasoning paths between entities can be explicitly captured, thereby enhancing the model’s ability to perform multi-step reasoning.

In addition, in a document, an entity pair may involve multiple different relations, which are often interrelated rather than independent. Therefore, we argue that if all relations can be predicted in advance for this document (e.g., the global relations as shown in Figure 1), it could provide a more comprehensive reasoning context for the model. This enables the model to more fully cover all possible relations when predicting the relations of an entity pair, which is crucial for DocRE multi-label classification tasks. Nevertheless, previous work has never considered predicting all relations in advance to assist with document-level extraction tasks.

Based on these observations, we propose a new framework namely **GREP** (global relations and entity pair reasoning) for document-level relation extraction. *First*, to address the first issue of better modeling the interactions between entity pairs, we utilize the global interdependencies between entity pairs to perform multi-step reasoning at the entity pair level, where entity pairs in a document are constructed into a graph that contains all reasoning paths between entity pairs. This approach captures deeper semantic associations and fine-grained interaction information. *Second*, to address the second issue, we for the first time propose a new auxiliary task: predicting all possible relations that exist in a document. By analyzing the relation types that may occur, this task enables the model to focus on the relations present in a document. Through the above two strategies, our model not only accurately identifies relations for entity pairs with multi-hop reasoning paths but also effectively reduces interference from non-existent relations, optimizing the final relation prediction performance. Moreover, building on previous works (Xiao et al., 2022; Xie et al., 2022; Ma et al., 2023), which commonly incorporate evidence retrieval in DocRE, we also introduce evidence retrieval as an auxiliary task to focus on key sentences in documents. In summary,

the main contributions of this paper are as follows:

- We propose a novel framework that leverages the global interdependencies between entity pairs to capture fine-grained interactions and perform multi-step reasoning at entity pair level.
- We introduce, for the first time, a simple yet effective auxiliary task that predicts all possible relations within a document. This task enables the model to filter out the most unlikely relations, thereby improving the overall performance of relation extraction.
- Experimental results on two DocRE datasets demonstrate that our approach achieves state-of-the-art (SOTA) performance while maintains high efficiency with low computational overhead. Furthermore, the proposed global relation prediction task serves as a versatile plugin, consistently improving the prediction performance of other DocRE models.

2 Related Work

DocRE methods can be mainly divided into two categories: graph-based methods and transformer-based methods.

2.1 Transformer-based Models

With the advances of the transformer (Vaswani et al., 2017), ATLOP (Zhou et al., 2021) addresses the multi-label classification task in DocRE through adaptive thresholds and localized context pooling. Furthermore, models like Eider (Xie et al., 2022) and DREEAM (Ma et al., 2023) extend ATLOP by integrating evidence extraction tasks. SRF (Zhang et al., 2024) leverages mention fusion, evidence extraction, and secondary reasoning to enhance prediction in DocRE. TTM-RE (Gao et al., 2024) introduces a memory-augmented model by incorporating pseudo entities and fine-tuning on a large distantly-labeled training dataset.

2.2 Graph-based Models

Graph-based approaches construct document graphs, treating entities and their mentions as nodes, and learn associations between entities through information propagation (Christopoulou et al., 2019; Nan et al., 2020; Wang et al., 2020; Xu et al., 2021). DocuNet (Zhang et al., 2021) captures local and global information by predicting an entity-level relation matrix, similar to semantic

segmentation in computer vision. AA (Lu et al., 2023) integrates graph-based and transformer-based methods, effectively capturing fine-grained interactions between entities. Descriptions of other graph/transformer-based methods can be found in the review work (Delaunay et al., 2023).

However, these methods mainly focus on mention-level and entity-level modeling, often neglecting the direct interactions between entity pairs and the effective integration of global relational features. In contrast, our work combines entity pair reasoning with an auxiliary task of predicting relations in documents. This integration enhances the global reasoning ability of entity pairs and captures relational features more accurately, significantly improving the overall performance of the model.

3 Problem Definition

In the DocRE task, given a document D , it contains tokens $W_D = \{w_i\}_{i=1}^{|W_D|}$, sentences $X_D = \{x_i\}_{i=1}^{|X_D|}$, and entities $E_D = \{e_i\}_{i=1}^{|E_D|}$. Each entity $e \in E_D$ is represented by its mentions $M_e = \{m_i\}_{i=1}^{|M_e|}$ and each mention $m \in M_e$ is a phrase in the document. The goal is to predict the relations between all entity pairs (e_s, e_o) from a predefined set of relations $\mathcal{R} \cup \{NA\}$, where NA signifies the absence of relation for an entity pair. For each entity pair (e_s, e_o) with a non-NA relation, we define its evidence $V_{s,o} = \{x_{v_k}\}_{k=1}^K$ as the subset of sentences that is sufficient for human annotators to infer the relation. Evidence annotations may be provided during training, depending on the dataset, but are not available during inference.

4 Methodology

Our GREP model consists of three main modules as shown in Figure 2: **Entity Pair Reasoning Module**, which leverages global dependencies between entity pairs to construct an entity pair graph for effective reasoning; **Evidence Extraction Module**, which guides the model’s focus to key sentences in the document; and **Global Relation Prediction Module**, which predicts all potential relations in the document to filter out irrelevant ones and enhance overall relation extraction performance.

4.1 Document Encoding

For a document D containing τ tokens, $W_D = \{w_i\}_{i=1}^\tau$, and multiple entities, each with multiple mentions, we first insert “*” before and after each mention (Zhang et al., 2017), then use an encoder

to obtain the token embedding matrix $H \in \mathbb{R}^{\tau \times d}$, where d is the dimension for a pre-trained language model (PLM), and the inter-token attention matrix $A \in \mathbb{R}^{\tau \times \tau}$:

$$H, A = \text{PLM}([w_1, w_2, \dots, w_\tau]) \quad (1)$$

For each entity e , let its mention set be $\{m_i\}_{i=1}^{|M_e|}$, h_{m_i} be the embedding of mention m_i at the entity’s “*” position. We apply LogSumExp pooling (Jia et al., 2019) to the embeddings of all mentions to obtain the embedding representation of the entity:

$$h_e = \log \sum_{i=1}^{|M_e|} \exp(h_{m_i}) \quad (2)$$

For each entity pair (e_s, e_o) , we obtain its attention weight $q^{(s,o)}$ and context representation $c^{(s,o)}$:

$$q^{(s,o)} = \frac{a_s \circ a_o}{a_s^\top a_o} \quad (3)$$

$$c^{(s,o)} = H^\top q^{(s,o)} \quad (4)$$

where \circ denotes the Hadamard product, and a_s and a_o denote the attention to all tokens for entities e_s and e_o , respectively.

4.2 Entity Pair Reasoning Module

To better capture interactions between entity pairs in a document, we first generate an initial embedding for an entity pair. Then, we propose a method for constructing an entity pair graph for the document, which is used to update and obtain the final embedding of the entity pair, and subsequently perform inference based on this updated embedding.

First, for each entity pair (e_s, e_o) , we take their embeddings h_{e_s}, h_{e_o} and context feature $c^{(s,o)}$, and combine the head and tail entity embeddings with the context embedding. We then map them into hidden representations $z_s^{(s,o)}$ and $z_o^{(s,o)}$. The *initial entity pair embedding* $f^{(s,o)}$ is computed using a group bilinear function as follows:

$$z_s^{(s,o)} = \tanh(W_s[h_{e_s} || c^{(s,o)}] + b_s) \quad (5)$$

$$z_o^{(s,o)} = \tanh(W_o[h_{e_o} || c^{(s,o)}] + b_o) \quad (6)$$

$$f^{(s,o)} = z_s^{(s,o)\top} W_p z_o^{(s,o)} \quad (7)$$

where $W_s, W_o \in \mathbb{R}^{2d \times d}$, $W_p \in \mathbb{R}^{d \times d}$ are learnable parameters.

In our model, we propose an approach for *constructing an entity pair graph for a document* based on associations between entity pairs. Each entity

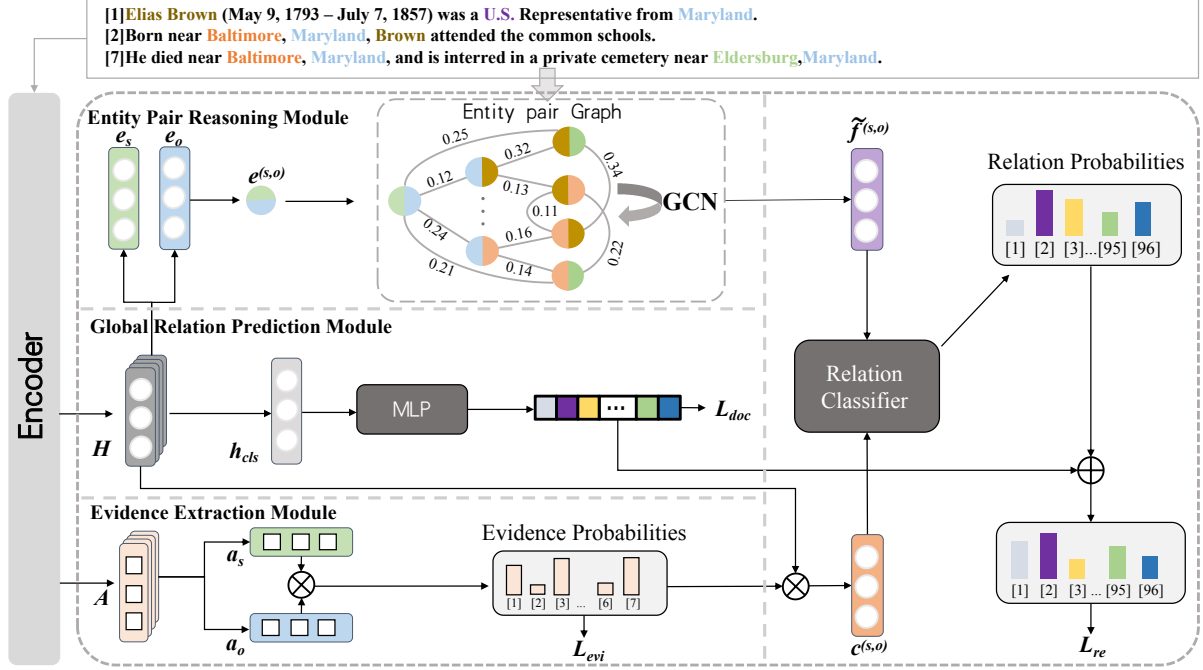


Figure 2: The overview of our GREP framework.

pair is treated as a node in the graph, and edges between nodes are added based on whether the tail entity of one entity pair matches the head entity of another. For a document D with n entities, we construct a graph $G = (V, N)$ where $n \times (n - 1)$ entity pairs form the graph’s nodes. Further, we propose to use an attention mechanism to assign different weights to neighboring nodes, modeling the interactions between entity pairs more effectively. Specifically, the attention weight $\alpha_{(j,k)}$ between nodes j and k is computed as follows:

$$\alpha_{(j,k)} = \frac{\exp[Q f_k^l (K f_j^l)^\top]}{\sum_{k' \in \mathcal{N}(j)} \exp[Q f_{k'}^l (K f_j^l)^\top]} \quad (8)$$

$$f_j^l = \sigma \left(\sum_{k=1}^n \alpha_{(j,k)} W^l f_j^{l-1} + b^l \right) + f_j^{l-1} \quad (9)$$

where Q and K are learnable parameter matrices used to map node features to queries and keys, f_k^l and f_j^l represent the features of nodes j and k at layer l , and $W^l \in \mathbb{R}^{d \times d}$ is a learnable parameter. Here, after obtaining the multi-head attention scores, we use a GCN (Kipf and Welling, 2017) to pass messages between entity pair nodes. By normalizing the attention weights over all nodes in the neighborhood $\mathcal{N}(j)$, we obtain the attention coefficient from node j to node k .

Through multiple convolutions over the entity pair graph, we iteratively update the representation of node j , which corresponds to the entity

pair (e_s, e_o) , resulting in the updated representation $f_{\text{update}}^{(s,o)}$ for the entity pair. Subsequently, we fuse this updated representation with the features of the individual entities to derive the final *enhanced entity pair representation*:

$$\tilde{f}_s^{(s,o)} = [h_{e_s}; f_{\text{update}}^{(s,o)}; c^{(s,o)}] \quad (10)$$

$$\tilde{f}_o^{(s,o)} = [h_{e_o}; f_{\text{update}}^{(s,o)}; c^{(s,o)}] \quad (11)$$

Finally, we employ a group bilinear function to predict the probabilities of relations for the entity pair (e_s, e_o) to achieve relation classification.

$$\tilde{z}_s^{(s,o)} = \tanh(\tilde{W}_s \cdot \tilde{f}_s^{(s,o)} + \tilde{b}_s) \quad (12)$$

$$\tilde{z}_o^{(s,o)} = \tanh(\tilde{W}_o \cdot \tilde{f}_o^{(s,o)} + \tilde{b}_o) \quad (13)$$

$$p^{(s,o)} = \tilde{z}_s^{\top(s,o)} \tilde{W}_p \tilde{z}_o^{(s,o)} \quad (14)$$

where $\tilde{W}_p \in \mathbb{R}^{d \times d}$, $\tilde{W}_s, \tilde{W}_o \in \mathbb{R}^{3d \times d}$ are learnable parameters.

4.3 Evidence Extraction Module

We introduce evidence retrieval as an auxiliary task to focus the model on key information in the document through evidence prediction. For an entity pair (e_s, e_o) in the document, we compute the context attention weight $q^{(s,o)}$ as described in Eq. (3). Then, for a sentence x_i starting at token $t_{\text{START}(x_i)}$ and ending at token $t_{\text{END}(x_i)}$, the attention score

$u_i^{(s,o)}$ of the entity pair for the sentence x_i is computed by summing the attention scores of all tokens in the sentence:

$$u_i^{(s,o)} = \sum_{t=\text{START}(x_i)}^{\text{END}(x_i)} q_t^{(s,o)} \quad (15)$$

Let $u^{(s,o)} \in D^{|X_D|}$ be the weight distribution of all sentences in the document, following the approach of (Ma et al., 2023), we minimize the Kullback-Leibler (KL) divergence between the extracted importance distribution $u^{(s,o)}$ and the evidence distribution obtained from the true evidence labels $v^{(s,o)} \in D^{|X_D|}$:

$$L_{evi} = \sum_{s \neq o} v^{(s,o)} \left(\log u^{(s,o)} - \log v^{(s,o)} \right) \quad (16)$$

4.4 Global Relation Prediction Module

In this module, we for the first time propose to introduce a new auxiliary task, global relation prediction task, which aims at predicting all possible relations that exist in a document. This enables the model to more fully cover all possible relations when predicting the relations of an entity pair, which is crucial for DocRE multi-label classification tasks.

Specifically, we pass the [CLS] token representation of the document, encoded by BERT (Devlin et al., 2019), through a linear layer for relation classification. Let h_{cls} be the [CLS] token representation of the document, W_r be the weight matrix of the linear layer, the predicted relation distribution p^{doc} is calculated as follows:

$$p^{doc} = \sigma(W_r h_{cls} + b_r) \quad (17)$$

To train the global relation prediction task, we minimize the binary cross-entropy loss function between the predicted relation distribution p^{doc} and the true relation labels y^{rel} :

$$L_{doc} = - \sum_{i=1}^n \left[y_i^{rel} \log p_i^{doc} + (1 - y_i^{rel}) \log (1 - p_i^{doc}) \right] \quad (18)$$

Finally, we enhance the model’s prediction capability by fusing the entity pair logits with the global relation prediction logits. This fusion uses document-level information to complement entity pair-level information, improving the model’s prediction accuracy in complex scenarios. Specifically, we combine the entity pair logits $p^{(s,o)}$ obtained

from Eq. (14) with the global relation prediction logits p^{doc} obtained from Eq. (17) to generate the final relation prediction logits $\tilde{p}^{(s,o)}$ as follows:

$$\tilde{p}^{(s,o)} = p^{(s,o)} + p^{doc} \quad (19)$$

4.5 Loss Function

To more effectively address the multi-label classification tasks, we also adopt the adaptive threshold loss method (Zhou et al., 2021) as the classification loss to train our model. Specifically, it introduces an additional threshold relation class TH , and optimizes the loss by increasing the logits of positive relations P_T above the threshold and decreasing the logits of negative relations N_T below the threshold. The loss function for relation classification can be formalized as:

$$L_{re} = - \sum_{r \in P_T} \log \left(\frac{\exp(\tilde{p}_r)}{\sum_{r' \in P_T \cup TH} \exp(\tilde{p}_{r'})} \right) - \log \left(\frac{\exp(\tilde{p}_{TH})}{\sum_{r' \in N_T \cup TH} \exp(\tilde{p}_{r'})} \right) \quad (20)$$

During training, we integrate the relation classification loss, evidence extraction loss, and global relation prediction loss, using coefficients α and β to balance these components. The overall training loss function for the model can be formalized as:

$$L = L_{re} + \alpha \times L_{doc} + \beta \times L_{evi} \quad (21)$$

4.6 Inference Fusion Phase

During the prediction phase, to prevent the model from focusing solely on the evidence sentences due to the introduction of the evidence module and neglecting the global information of the document, we also introduce an inference fusion strategy. Based on AA (Lu et al., 2023), the model trained with evidence loss in Eq. (21) infers the relations r of the entity pair (e_s, e_o) within a document D , abbreviated as $I(r|e_s, e_o; D)$. The supporting evidence sentences for this inference are then used to generate a pseudo-document D' . Subsequently, this pseudo-document is fed into the model trained without evidence loss to infer $I(r|e_s, e_o; D')$. The two results are fused, and the final fused prediction $I(r|e_s, e_o)$ is expressed as:

$$I(r|e_s, e_o) = I(r|e_s, e_o; D) + I(r|e_s, e_o; D') - \gamma \quad (22)$$

where γ is a hyperparameter tuned on the dev set.

Model	PLM	Dev				Test	
		Ign-F1	F1	Intra-F1	Inter-F1	Ign-F1	F1
ATLOP (Zhou et al., 2021)	BERT_base	59.22	61.09	67.26	53.20	59.31	61.30
GAIN (Zeng et al., 2020)	BERT_base	59.14	61.22	67.10	53.90	59.00	61.24
DocuNet (Zhang et al., 2021)	BERT_base	59.86	61.83	-	-	59.93	61.86
KD-DocRE (Tan et al., 2022a)	BERT_base	60.08	62.03	-	-	60.04	62.08
SAIS (Xiao et al., 2022)	BERT_base	59.98	62.96	-	-	<u>60.96</u>	62.77
Eider (Xie et al., 2022)	BERT_base	60.51	62.48	68.47	55.21	60.42	62.47
SRF (Zhang et al., 2024)	BERT_base	60.46	62.50	-	-	59.84	62.11
DREEAM (Ma et al., 2023)	BERT_base	60.51	62.55	-	-	60.03	62.49
AA (Lu et al., 2023)	BERT_base	<u>61.31</u>	<u>63.38</u>	<u>69.41</u>	<u>55.92</u>	60.84	<u>63.10</u>
Ours	BERT_base	62.10±0.06	64.10±0.12	69.71±0.11	57.12±0.18	61.36	63.55
ATLOP (Zhou et al., 2021)	RoBERTa_large	61.32	63.18	69.60	55.01	61.39	63.40
SSAN (Xu et al., 2021)	RoBERTa_large	60.25	62.08	-	-	59.47	61.42
DocuNet (Zhang et al., 2021)	RoBERTa_large	62.23	64.12	-	-	62.39	64.55
KD-DocRE (Tan et al., 2022a)	RoBERTa_large	62.16	64.19	-	-	62.57	64.28
DREEAM (Ma et al., 2023)	RoBERTa_large	62.29	64.20	-	-	62.12	64.27
Eider (Xie et al., 2022)	RoBERTa_large	62.34	64.27	70.36	56.53	62.85	64.79
SAIS (Xiao et al., 2022)	RoBERTa_large	62.23	65.17	-	-	63.44	65.11
AA (Lu et al., 2023)	RoBERTa_large	<u>63.15</u>	<u>65.19</u>	<u>71.09</u>	<u>57.83</u>	62.88	<u>64.98</u>
Ours	RoBERTa_large	63.74±0.07	65.64±0.06	71.15±0.12	58.94±0.14	<u>62.96</u>	64.86

Table 1: Performance comparison on the DocRED dataset. Results of other models are from the original papers. We mark the best results in **bold** and the second-best underlined.

Model	Dev		Test			
	Ign-F1	F1	Ign-F1	F1	Intra-F1	Inter-F1
ATLOP (Zhou et al., 2021) [†]	76.88	77.63	76.94	77.73	80.18	75.13
DocuNet (Zhang et al., 2021) [†]	77.53	78.16	77.27	77.92	79.91	76.64
KD-DocRE (Tan et al., 2022a) [†]	77.92	78.65	77.63	78.35	79.57	77.26
DREEAM (Ma et al., 2023)	-	-	79.66	80.73	-	-
PEMSCL (Guo et al., 2023)	79.02	79.89	79.01	79.86	-	-
AA (Lu et al., 2023)	<u>80.04</u>	<u>81.15</u>	<u>80.12</u>	<u>81.20</u>	<u>83.41</u>	<u>79.24</u>
TTM-RE (Gao et al., 2024)	78.22	78.25	78.54	80.08	-	-
Ours	80.60±0.08	81.26±0.11	81.00±0.13	81.61±0.04	83.60±0.05	79.88±0.08

Table 2: Experimental results on the Re-DocRED dataset based on RoBERTa_large. Results with [†] are sourced from (Tan et al., 2022b), while others are from the original papers.

5 Experimental Settings

5.1 Datasets

We evaluate our model on two public datasets: DocRED and Re-DocRED. The dataset statistics are shown in Table 3. DocRED, proposed by (Yao et al., 2019), is a large-scale dataset for document-level relation extraction. It includes 97 relation types, with approximately 40.7% of relational facts requiring extraction from multiple sentences.

Due to issues with under-labeling and mislabeling in DocRED, Tan et al. (2022b) propose Re-DocRED, which provides a re-annotated version of DocRED. Re-DocRED contains more than twice the number of triples as DocRED and provides cleaner dev and test sets, enabling a more realistic performance evaluation for DocRE.

Datasets	DocRED			Re-DocRED		
	Train	Dev	Test	Train	Dev	Test
#Docs	3,053	1,000	1,000	3,053	500	500
Avg. #Entities	19.5	19.6	19.5	19.4	19.4	19.6
Avg. #Entity Pairs	392.6	392.2	398.2	390.8	386.5	397.3
Avg. #Triples	12.5	12.3	-	28.1	34.6	34.9
Avg. #Sentences	7.9	8.1	7.9	7.9	8.2	7.9
Avg. #Relations	5.4	5.3	-	8.6	10.0	9.5

Table 3: Statistics of DocRED and Re-DocRED. Avg.# represents the average number of items per document.

5.2 Implementation Details

Our model is implemented based on PyTorch (Paszke et al., 2019) and Huggingface’s Transformers (Wolf et al., 2019). We use BERT_{base} (Devlin et al., 2019) and RoBERTa_{large} (Liu et al., 2019) as encoders, and optimize the model using the AdamW

optimizer (Kingma and Ba, 2015). A linear warm-up is applied during the first 6% of the training steps. The model embedding dimension and hidden dimension are set to 768. The learning rate for the BERT encoder is set to $3e-5$, while the learning rate for other encoders is $5e-5$. The number of layers and iterations for the GCN is 2. All hyperparameters are tuned based on the development set. All experiments are conducted on a single NVIDIA RTX 3090 GPU. Some hyperparameters are listed in Table 4.

Dataset	DocRED		Re-DocRED
	BERT	RoBERTa	RoBERTa
epoch	30	30	30
lr_encoder	$5e-5$	$3e-5$	$3e-5$
lr_classifier	$1e-4$	$1e-4$	$1e-4$
batch size	4	4	4
warmup_ratio	$6e-2$	$6e-2$	$6e-2$
α	$1e-1$	$1e-1$	$1e-1$
β	$1e-1$	$3e-2$	$5e-2$

Table 4: Best hyper-parameters of our model observed on the dev set.

5.3 Evaluation Metrics

We use F1 and Ign_F1 as the primary evaluation metrics for relation extraction. Ign_F1 excludes relation triples that appear in both the training and dev/test sets. We also report Intra_F1 for relations within a single sentence and Inter_F1 for relations that span multiple sentences. To reduce potential bias, we report the average results and standard deviations over 5 independent runs.

5.4 Baselines

We conduct comparisons with two main categories of baseline models (as detailed in Section 2): (i) Graph-based models, including GAIN, AA, etc. (ii) Transformer-based models, including ATLOP, KD-DocRE, Eider, DREEAM, SRF, SAIS, etc. Moreover, in Table 2, we report the results of TTM-RE without utilizing distantly supervised training data, aligning with the experimental setup adopted in the other baselines and our work.

6 Main Results and Analysis

6.1 Main Results

Table 1 reports the results on the DocRED dataset. Our model GREP outperforms most of baseline methods. On the test set, our BERT_base model

outperforms the ATLOP_base model, achieving F1 and Ign-F1 improvements of 3.01 and 2.88. Against the state-of-the-art AA model, our model shows gains of 0.72 and 0.79 in F1 and Ign-F1 on the dev set. Additionally, Inter-F1 and Intra-F1 scores improve by 1.2 and 0.3, respectively, demonstrating the enhanced intra-sentence and cross-sentence reasoning capabilities of our model.

Table 2 shows our results on the Re-DocRED dataset, where the performance gap between our model and baselines is more pronounced. Our RoBERTa_large-based model outperforms ATLOP, achieving F1 and Ign-F1 scores that exceed it on the test set by 3.88 and 4.06, respectively. Additionally, our model surpasses the previous state-of-the-art model in Ign-F1 by 0.56 on the dev set and 0.88 on the test set. These improvements collectively demonstrate the effectiveness of our model in handling multi-step reasoning and complex relation judgment between entity pairs.

Model	Ign-F1	F1	Intra-F1	Inter-F1
Ours-BERT_base	62.10	64.10	69.71	57.12
w/o Graph	61.34	62.98	68.75	55.77
w/o GRPM	61.58	63.45	69.14	56.43
w/o EEM	61.42	63.35	68.94	56.39

Table 5: Ablation study on DocRED dev set.

6.2 Ablation Study

To investigate the effectiveness of different components in our model, we conducted a series of ablation studies on DocRED. The results are shown in Table 5, and a detailed analysis is outlined as follows:

w/o Graph: Removing the *entity pair graph* leads to a performance drop, with the F1 score decreasing by 1.12 and the Ign-F1 score by 0.76, indicating that the entity pair graph can explicitly capture reasoning paths between entities, enhancing the model’s multi-step reasoning ability.

w/o GRPM: Removing the *global relation prediction module* leads to a performance decline, with F1 dropping by 0.65 and Ign-F1 by 0.52. This demonstrates that the module is crucial for accurately identifying relations within the document, and its absence weakens the model’s overall performance.

w/o EEM: We exclude *evidence extraction module* during training. The F1 score drops by 0.75, which indicates that the model’s performance decreases due to the lack of focus on evidence sentences within the document.

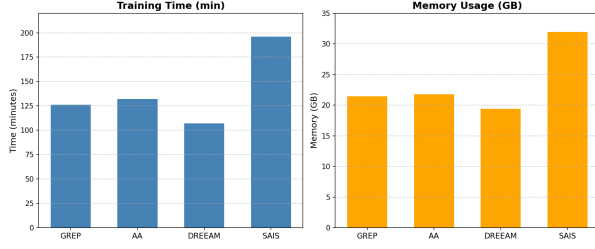


Figure 3: Comparison of training costs between GREP and SOTA models.

EE	GRP	RE	Ign-F1	F1
		✓	61.16	63.11
✓		✓	61.58	63.45
	✓	✓	61.42	63.35
✓	✓	✓	62.10	64.10

Table 6: Exploring the effectiveness of two complementary tasks, i.e., evidence extraction (EE) and global relation prediction (GRP), for relation extraction (RE) on the DocRED dev set.

7 Further Analysis

7.1 Model Cost Comparison

In Figure 3, we compare the training costs of the GREP model with previous SOTA models. GREP trains in 126 minutes with 21.47 GB of memory. Compared to the non-graph-based method DREEAM (Ma et al., 2023), GREP requires slightly more time and memory, but it outperforms DREEAM in terms of performance. In contrast to the multitask-based SAIS (Xiao et al., 2022), our model incurs lower computational costs. Compared to the previous SOTA method AA (Lu et al., 2023), which uses both transformers and graphs, our model is more efficient in training time and memory consumption. These results indicate that GREP strikes a good balance between the training cost and the relation extraction performance.

7.2 Impact of Multi-Task Learning on DocRE

We explore the impact of multi-task learning on relation extraction (RE). In Table 6, by incorporating evidence extraction (EE) task, the model better focuses on key evidence, leading to improved performance. Our proposed global relation prediction (GRP) task further enhances accuracy by guiding attention towards document-level relations. Combining these tasks significantly boosts the model’s overall performance.

7.3 Generalization Analysis of Global Relation Prediction

In addition to its effectiveness in our model, the global relation prediction (GRP) module also improves performance in other state-of-the-art models, as demonstrated in Table 7. We select ATLOP, DREEAM, and AA, re-running these models under their original configurations while incorporating the GRP module. Experimental results show that adding the GRP module consistently enhances the performance of these models. The results once again demonstrate that the GRP sub-task we propose is simple yet effective for the DocRE task.

Model	Dev		Test	
	F1	Ign F1	F1	Ign F1
ATLOP	61.01	59.11	61.30	59.31
+Global Relation Prediction	61.21	59.26	61.38	59.49
DREEAM	62.55	60.51	62.49	60.03
+Global Relation Prediction	62.69	60.64	62.65	60.17
AA	63.38	61.31	63.10	60.84
+Global Relation Prediction	63.55	61.48	63.25	60.98

Table 7: Evaluating the generalizability of global relation prediction across different models on DocRED.

7.4 Exploring the Impact of the Number of Entity Pairs in a Document on DocRE

In DocRE tasks, documents with a higher number of related entity pairs may tend to require more entity pairs for reasoning. To assess the model’s ability to learn potential association between entity pairs, we conduct an evaluation on the DocRED dev set, categorizing documents by the number of related entity pairs they contain. We then measure the performance of our model and the previously competitive ATLOP (Zhou et al., 2021), DREEAM (Ma et al., 2023) models across these categories. The experimental results in Figure 4 demonstrate that our model consistently achieves better performance across all document categories.

7.5 Scalability Across Entity Quantities

We evaluate the scalability of our model with respect to entity count on the DocRED development set, grouping documents into five bins based on the number of entities: 0–10 (44 documents), 10–20 (527 documents), 20–30 (399 documents), 30–40 (29 documents), and 40–50 (1 document). As illustrated in Figure 5, GREP consistently outperforms strong baselines across all entity ranges, with particularly notable gains on documents containing a larger number of entities.

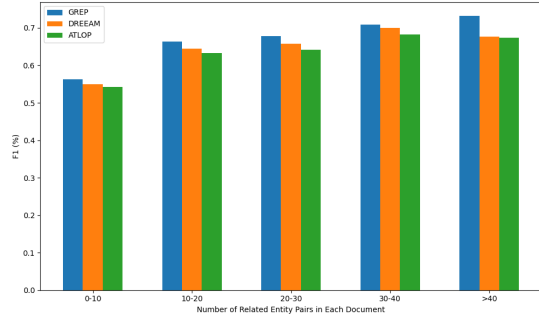


Figure 4: F1 scores of documents with different numbers of entity pairs on the DocRED dev set.

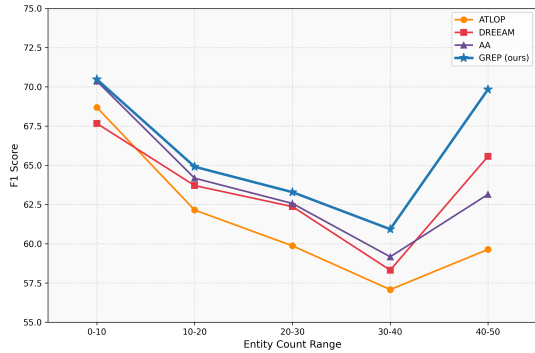


Figure 5: F1 scores across different entity count ranges on the DocRED dev set.

In addition to achieving higher overall F1 scores, GREP demonstrates improved robustness, exhibiting only a 13.55% relative performance drop from the 0–10 to the 30–40 group. This decline is substantially smaller than those observed for ATLOP (16.90%) and AA (15.91%), suggesting that GREP maintains better scalability as relational complexity increases. These results highlight the model’s ability to effectively handle documents with high entity density.

7.6 Impact of Coefficients α and β

We evaluate the impact of α on model performance, where α serves as the balancing coefficient for the Global Relation Prediction module in the overall loss. As shown in Figure 6, we select five values for α : 0, 0.05, 0.1, 0.15, and 0.2. We can observe that the F1 score on the DocRED dev set increases as α grows, reaching its peak when α is set to 0.1. However, as α continues to increase, the F1, Inter-F1, and Intra-F1 scores gradually decrease. Therefore, we choose $\alpha = 0.1$ as the balancing coefficient for model training. For β , following all previous evidence extraction work (Xie et al., 2022; Ma et al., 2023), β is set to 0.1.

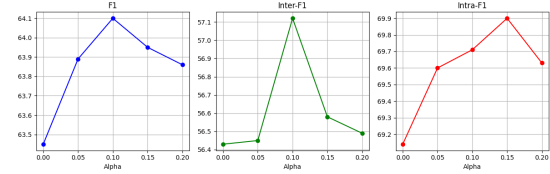


Figure 6: Impact of Coefficient α .

7.7 Case Study

We conduct case studies, and the results show that by employing multi-step reasoning with entity pairs and global relations, our model better understands entity interactions, leading to more accurate document-level relation prediction. For detailed results, please refer to Appendix A.1.

8 Conclusions

In this paper, we propose a novel DocRE framework based on global relations and entity pair reasoning. We first introduce a new task specifically for DocRE that predicts all possible relations that exist in a document, helping to filter out the most unlikely relations. The task is simple but effectively enhances relation extraction performance and can be incorporated into other DocRE models. Further, we propose to construct an entity pair graph for a document to capture fine-grained interactions and perform multi-step reasoning at the entity pair level. Empirical studies demonstrate the effectiveness of our method, outperforming previous SOTA models.

Limitations

Despite our framework GREP demonstrating advantages in the document-level relation extraction task, there are still some limitations. First, when processing long documents containing a large number of entity pairs, it may lead to increased computational overhead, thus affecting the efficiency of the model. Second, while predicting all possible relations can improve document-level relation extraction, it may also introduce noise and potentially affect the effectiveness of the model. In our near future work, we plan to conduct more in-depth and extensive exploration to address these limitations and further improve the framework’s performance.

Acknowledgments

The authors sincerely thank the reviewers for their valuable comments, which improved the paper. The work is supported by the National Natural Science Foundation of China (62276057).

References

- Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4924–4935.
- Julien Delaunay, Thi Hong Hanh Tran, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Nicolas Sidere, and Antoine Doucet. 2023. A comprehensive survey of document-level relation extraction (2016-2023). *arXiv preprint arXiv:2309.16396*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186.
- Chufan Gao, Xuan Wang, and Jimeng Sun. 2024. TTM-RE: Memory-augmented document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 443–458.
- Jia Guo, Stanley Kok, and Lidong Bing. 2023. Towards integration of discriminability and robustness for document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2598–2609.
- Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multi-scale representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 3693–3704.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.
- Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations (ICLR)*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. 2023. Anaphor assisted document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15453–15464.
- Youmi Ma, An Wang, and Naoaki Okazaki. 2023. 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 (EACL)*, pages 1963–1975.
- Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing (NeurIPS)*, pages 8024–8035.
- Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.
- Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2020. Recurrent interaction network for jointly extracting entities and classifying relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3722–3732.
- Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics (ACL)*, pages 1672–1681.
- Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8472–8487.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3711–3721.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

- Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. 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 (NAACL)*, pages 2395–2409.
- Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: (ACL)*, pages 257–268.
- Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 35, pages 14149–14157.
- Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 764–777.
- Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640.
- Fu Zhang, Qi Miao, Jingwei Cheng, Hongsen Yu, Yi Yan, Xin Li, and Yongxue Wu. 2024. SRF: Enhancing document-level relation extraction with a novel secondary reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15426–15439.
- Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3999–4006.
- Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2205–2215.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 35–45.
- Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 35, pages 14612–14620.
- Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1331–1339.

A Appendix

A.1 Case Study

Figure 7 presents one of case studies, where we examine four sentences from a document containing mentions of *Philip*, *United Kingdom*, *Elizabeth II*, and *Commonwealth*. Both ATLOP and AA face challenges in predicting the *country of citizenship* relation between *Philip* and *United Kingdom*, as well as the *head of state* relation between *Elizabeth II* and *United Kingdom*, which may need to perform multi-step reasoning based on the connections between entity pairs. In contrast, our model effectively applies multi-step reasoning based on the relations between entity pairs, which helps provide a more comprehensive understanding of the interactions between *Philip* and *United Kingdom*, as well as *Elizabeth II* and *United Kingdom*, ultimately leading to the successful identification of their relations.

<p>[1]Elizabeth II was Queen of Mauritius from 1968 to 1992.</p> <p>[3]The Queen was also the monarch of the United Kingdom and the other Commonwealth realms.</p> <p>[8]The Queen and her husband Prince Philip , Duke of Edinburgh , visited Mauritius for three days (24–26 March) in 1972 , as part of a tour of Asia and Africa .</p>			
ATLOP	<pre> graph TD Philip -- spouse --> ElizabethII[Elizabeth II] UK[United Kingdom] -- member of --> Commonwealth ElizabethII -- chairperson --> Commonwealth </pre>	AA	<pre> graph TD Philip -- spouse --> ElizabethII[Elizabeth II] UK[United Kingdom] -- country of citizenship --> ElizabethII UK -- member of --> Commonwealth ElizabethII -- chairperson --> Commonwealth </pre>
Ours	<pre> graph TD Philip -- country of citizenship --> UK[United Kingdom] Philip -- spouse --> ElizabethII[Elizabeth II] Philip -- head of state --> Commonwealth UK -- country of citizenship --> ElizabethII UK -- member of --> Commonwealth ElizabethII -- chairperson --> Commonwealth </pre>	label	<pre> graph TD Philip -- country of citizenship --> UK[United Kingdom] Philip -- spouse --> ElizabethII[Elizabeth II] Philip -- head of state --> Commonwealth UK -- country of citizenship --> ElizabethII UK -- member of --> Commonwealth ElizabethII -- chairperson --> Commonwealth </pre>

Figure 7: Case study comparison between our method and the existing models.