PLEM: Prototype Learning with Evidence Match for Improving Few-Shot Document-Level Relation Extraction

Anonymous ACL submission

Abstract

Few-shot document-level relation extraction (FSDLRE) aims to develop a model with the ability to generalize to new categories in the context of document-level relation extraction, using a small number of support samples. Among others, metric based meta-learning methods are widely used in FSDLRE, which involve constructing class prototypes using the contextual representation of the entire document and the representation of entity pairs for relation classification. However, in relation classification, only a subset of sentences in a document, known as evidence, is required to determine the relationship category of entity pairs. In this paper, we propose a prototype learning method with evidence match (PLEM). By introducing an evidence matching auxiliary task in the process of relation prototype construction, the model is guided to focus more on the semantics of evidence sentences when building prototypes, thereby enhancing the relation prototypes. We further design task-specific evidence prototypes, enabling the model to adapt to the evidence semantic space of different relation categories. Extensive experimental results demonstrate that PLEM outperforms the state-of-the-art methods, achieving an average improvement of 1.23% in Macro F1 across various settings of two FSDLRE benchmarks.

1 Introduction

Document-level relation extraction (DocRE) is aimed at classifying the types of relationships between each pair of entities within a document. This task is more closely aligned with real-world scenarios of downstream tasks such as knowledge graph construction and question answering (Zhou et al., 2021; Zhang et al., 2021; Wei and Li, 2022; Sun et al., 2023), compared to sentence-level relation extraction (Zhang et al., 2018; Distiawan et al., 2019; Hu et al., 2021; Liu et al., 2022). However, annotating for DocRE is costly and time-consuming, with data often exhibiting a long-tail distribution, and annotations in many domains are scarce. As a result, recent studies have been moving towards the setting of Few-Shot Document-Level Relation Extraction (FSDLRE), which is a promising solution for relation extraction at scale.

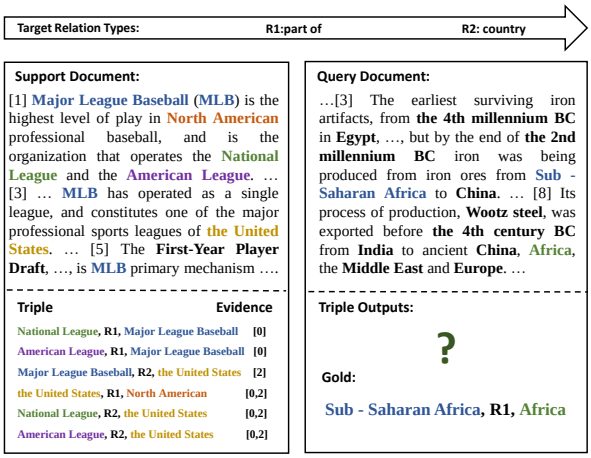

Figure 1: Description of the FSDLRE Task in the 1-DOC Setting with Pre-Annotated Entities in Bold and Colored Triples

Previous studies (Popovic and Färber, 2022; Meng et al., 2023) on FSDLRE primarily adopt a metric-based meta-learning framework (Vinyals et al., 2016; Snell et al., 2017), which aims to learn a metric space where prototypes for each relation category are constructed based on the information from support documents. Classification is performed by calculating the distance from the entity pair instance representations in the query documents to the prototype representations of each category. Through training on a series of sampled FSDLRE tasks, the model acquires general knowledge of FSDLRE, enabling it to quickly generalize to new tasks with novel relation types.

We illustrate an example of the FSDLRE task under the 1-DOC setting in Figure 1, where only a single support document is given, annotated with types of relations and evidence statements. In

such an episode, there are two target relation types: "*part of*" and "*country*". The task requires identifying the types of relations between predefined entities within the query document. To predict the relation between "*National League*" and "*the United States*", we learn from the first sentence that the "*National League*" is a part of "*Major League Baseball (MLB)*", and from the third sentence, that "*MLB*" is a professional sports league located in "*the United States*". Although the mention "*MLB*" also appears in the fifth sentence, this sentence does not semantically contribute to the prediction of this relation. For a pair of entities in the support document, their relation type can be determined based on a few sentences, and only the semantics of these sentences hold referential value for the same relation type in the query document. Including the semantics of irrelevant sentences as part of the representation in constructing relation prototypes and affecting the measurement of distance between instances from the query document and prototypes introduces noise into the model, leading to performance degradation.

Evidence sentences play a crucial role in document-level relation extraction under supervised scenario, where previous methods often jointly train evidence retrieval task with relation extraction, allowing both tasks to mutually enhance each other's performance (Huang et al., 2021; Xie et al., 2022; Ma et al., 2023b). However, such approaches are not suitable for cross-category domain few-shot scenarios due to their tendency to overfit specific categories. Given the effectiveness of evidence sentences in document-level relation extraction, it is vital to explore how to emphasize the semantics of evidence sentences in prototype learning and reduce the impact of irrelevant sentences. Moreover, the semantic spaces of different relation categories vary significantly, and applying the same evidence matching approach to all scenarios without considering the influence of relation categories on evidence matching fails to account for these differences.

In this paper, we propose an auxiliary task for prototype construction, evidence match, and propose a framework, PLEM, that integrates evidence matching into metric learning. During training, we first establish two additional base evidence matching prototypes, MATCH and UNMATCH, to represent whether sentences in a document are evidence for a triple. Guided by the inherent relational semantics of specific relations, we enhance the representation of evidence prototypes using the contextual semantics of specific relation pairs in the supporting documents. By jointly performing relation classification and evidence matching using multi-task learning, we improve the semantic representation in the relation prototype construction process, highlighting the role of evidence sentences in FSDLRE.

Contribution. (1) We propose a prototype learning framework (PLEM) for FSDLRE, which effectively improves the semantic distribution of relation prototypes by incorporating evidence matching into the construction process of the relation prototypes. (2) In PLEM, we introduce two task-specific learnable prototypes, MATCH and UNMATCH, for evidence matching, enabling the model to better adapt to evidence match in episodes of different relation categories. (3) We conduct experiments on two public document-level relation extraction datasets. The experimental results demonstrate the effectiveness of our PLEM model, which achieve state-of-the-art performance across multiple settings of two FSDLRE benchmarks.

2 Problem Formulation

Few-shot document-level relation extraction is conducted under the N-Doc setting (Popovic and Färber, 2022). In each independent FSDLRE task (also called an episode), there are N support documents and one query documents. For each support document D_S, it contains a set of triples T_S which includes all valid triples (e_h, r, e_t) in the document and their evidences $V_{h,t}$ which are the subset of sentences in the document. Here, e_h and e_t represent the head entity and tail entity of a given relation instance, and r is a member of the set $R_{episode}$, which signifies a specific type of relation. The set $R_{episode}$ includes all types of relations that need to be discerned as present or not in instances within the episode. The entity mentions in the query document D_Q are pre-annotated. The goal of FSDLRE is to predict the set of triples T_Q in the query document D_Q using the given information as input. This set includes all valid triples in D_Q, with the relationships being within the scope of the $R_{episode}$.

Our approach adheres to the conventional meta-learning framework, wherein the relationships encompassed within the training and testing phases, denoted as R_{train} and R_{test}, are distinct and non-overlapping. For each task, the relationships involved, referred to as $R_{episode}$, are specifically sub-

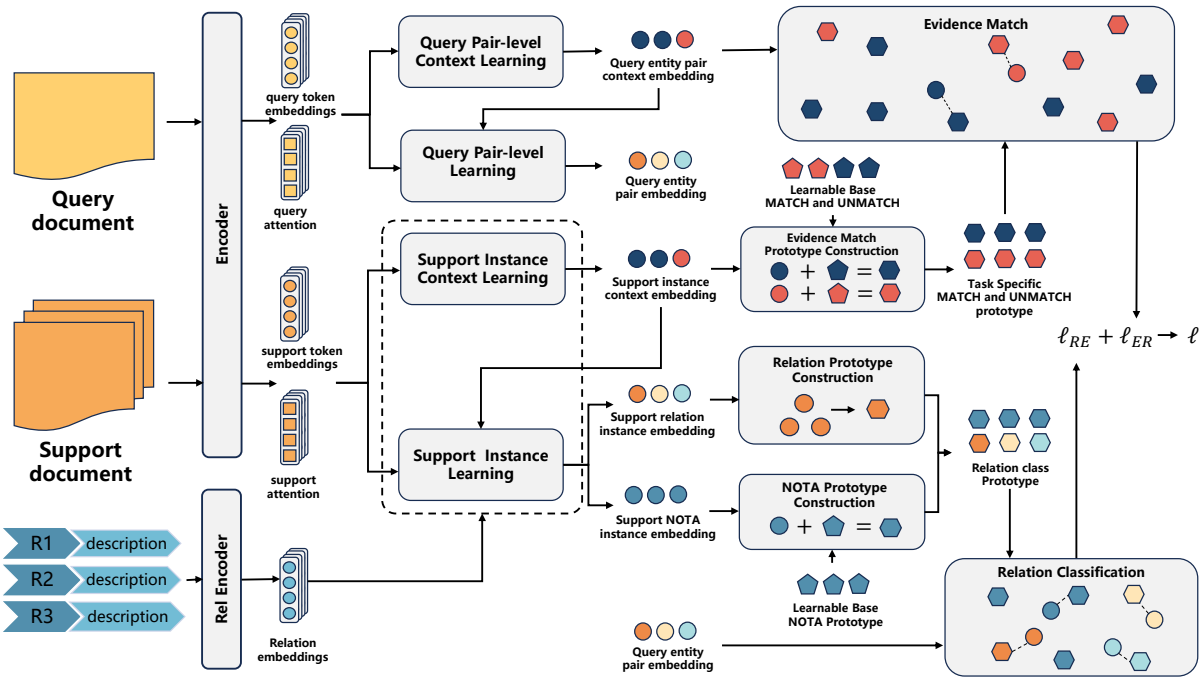

Figure 2: The overall architecture of PLEM.

sets of R_{train} and R_{test} during their respective training and testing stages. The annotations of the support documents are complete, meaning that any candidate entity pair for which no relation type has been assigned can be considered as NOTA (*None-Of-The-Above*).

3 Methodology

An illustration of the framework of PLEM is shown in Figure 2. We first introduce the encoding procedure for documents and entities in Section 3.1. In Section 3.2, We describe the process of constructing evidence prototypes and relationship prototypes. The training and inference processes are finally given in Section 3.3.

3.1 Document Encoding

We use a pre-trained language model (Devlin et al., 2019) to encode the input document. Given a document $D = [h_l]_{l=1}^{L}$ containing L tokens, we insert a special symbol "*" before and after each mention m_i as an entity marker (Zhang et al., 2017), with each mention m_i represented by the embedding of "*" at the start position. For a pre-trained language model with a dimension of d, we input the document D to obtain token embeddings H and attention scores A between tokens:

$$H, A = Encoder([h_1, ..., h_L]) \quad (1)$$

where $H \in \mathbb{R}^{L \times d}$ is the last hidden states and $A \in \mathbb{R}^{L \times L}$ is the average of the attention heads in the last transformer layer. We apply $logsumexp$ pooling (Jia et al., 2019) to obtain representations for each entity e_i from the representations $[h_{m_i}]_{i=1}^{N_e}$ of their corresponding mentions. Formally, for each entity e_i that occurs N_e times in the text, its representation h_e is calculated as follows:

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

In determining different triples, the model may need to focus on different parts of the context. We follow (Meng et al., 2023) to acquire a specific context representation $c^{(h,r,t)}$ for each triple (e_h, e_t, r) by incorporating relation label information. We first obtain the importance scores $A^{(h,t)}$ for each token at the entity pair level (Zhou et al., 2021):

$$A^{(h,t)} = \frac{A_h \odot A_t}{A_h^T A_t} \quad (3)$$

where $\odot$ is the Hadamard Product, $A_h \in \mathbb{R}^L$ and $A_t \in \mathbb{R}^L$ are attention scores for all tokens in the document based on e_h and e_t, and they are obtained by averaging the attention scores of the corresponding mentions of e_h and e_t.Then we use another pre-trained language model to encode the names and descriptions of the relation labels and use the

output of the "[CLS]" token as the relation embedding $h_r \in \mathbb{R}^d$:

$$h_r = Encoder(r) \tag{4}$$

and we calculate the relation-level attention A_r as to represent the degree of attention of the relation to each token in the text:

$$A^r = softmax(\frac{HWh_r}{\sqrt{d}}) \tag{5}$$

where $W \in \mathbb{R}^{d \times d}$ is a learnable parameter. Based on A^r and $A^{(h,t)}$, we get the attention distribution $A^{(h,r,t)}$ for each token of a single instance (e_h, r, e_t).The i-th value of $A^{(h,r,t)}$ is obtained as follows:

$$A_i^{(h,r,t)} = A_i^{(h,t)} + \amalg(i \in topk\%(A^{(h,t)} \odot A^r)) \cdot A_i^r \tag{6}$$

where $topk\%(x)$ returns the indices of the top k% largest values in x, and $\amalg$ is the indicator function. This instance-level attention focuses more on tokens related to the entity pair. Then we calculate specific context embeddings $c^{(h,r,t)} \in \mathbb{R}^d$ for each instance by:

$$c^{(h,r,t)} = H^T A^{(h,r,t)} \tag{7}$$

3.2 Prototype Construction

We construct multiple prototypes based on the encoded representations from the supporting documents, serving as the basis for calculating the relational distance of query document instances.

Evidence Prototype. We define two sets of learnable vectors $M_{base} = \{\boldsymbol{m}_i^{base} \in \mathbb{R}^d\}_{i=1}^{N_{evi}}$ and $U_{base} = \{\boldsymbol{u}_i^{base} \in \mathbb{R}^d\}_{i=1}^{N_{evi}}$ as the base prototypes for evidence matching across all tasks, where N_{evi} is a hyperparameter. Considering that in different episodes, the conceptual representations of evidence matching are not entirely the same due to the variation in relation types, we propose task-specific evidence prototypes built upon the base evidence matching prototypes.

Since the annotations in the supporting documents are complete, we select instances with relations from these documents to enhance the representation of the evidence prototype. We refer to the method for obtaining the specific context representation $c^{(h,r,t)}$ for each instance, and based on the label of the evidence, we obtain sentence representation s for each instance:

$$s^{(h,r,t)} = H^T_{m_s:m_e} A^{(h,r,t)}_{m_s:m_e} \tag{8}$$

where m_s and m_e are the positions of the starting and ending tokens of the sentence. The instance-level sentence representations are divided into two sets, S_{match} and $S_{unmatch}$, S_{match} includes representations of sentences that serve as evidence for a specific instance, whereas $S_{unmatch}$ contains representations of sentences that are not evidence for that specific instance. For each base evidence prototype in M_{base} and U_{base} specific evidence support instances are adaptively selected from S:

$$\begin{aligned}
&(e_h, r, e_t, match) \\
&= \underset{(e_h,r,e_t,match) \in S_{match}}{argmax} (s^{(h,r,t)} \cdot \boldsymbol{m}_i^{base} \\
&- \underset{u_i^{base} \in U_{base}}{max} s^{(h,r,t)} \cdot \boldsymbol{u}_i^{base}) \\
&(e_h, r, e_t, unmatch) \\
&= \underset{(e_h,r,e_t,unmatch) \in S_{unmatch}}{argmax} (s^{(h,r,t)} \cdot \boldsymbol{u}_i^{base} \\
&- \underset{m_i^{base} \in M_{base}}{max} s^{(h,r,t,u)} \cdot \boldsymbol{m}_i^{base})
\end{aligned} \tag{9}$$

where the selection criteria focuses on the affinity with prototypes sharing similar meanings and a divergence from those with contrasting meanings. Then, we integrate the selected evidence sentence prototypes from the support set into the base evidence prototype to obtain the final evidence MATCH and UNMATCH prototype m_i and u_i:

$$\begin{aligned}
\boldsymbol{m}_i &= \alpha s^{(h,r,t,m)} + (1-\alpha)\boldsymbol{m}_i^{base} \\
\boldsymbol{u}_i &= \alpha s^{(h,r,t,u)} + (1-\alpha)\boldsymbol{u}_i^{base}
\end{aligned} \tag{10}$$

where α is a hyperparameter that balances the semantics of task-specific evidence matching with those of the base evidence prototype. By adopting this approach, we obtain two task-specific evidence prototype sets, MATCH and UNMATCH. These sets incorporate both general knowledge from meta-learning and specific knowledge from evidence labels in the support documents.

Relation Prototype. In the construction of relation prototypes, we use label information to build instance-level relation prototypes, enabling each prototype to more effectively focus on relation-relevant information in supporting documents. We first fuse the specific context representation $c^{(h,r,t)}$ with the original semantic entity representation of the PLM to obtain the instance-level head and tail entity representation:

$$z_h^{(h,r,t)} = tanh(W_h[h_{e_h}; c^{(h,r,t)}] + b_h) \tag{11}$$

$$z_t^{(h,r,t)} = tanh(W_t[h_{e_t}; c^{(h,r,t)}] + b_t) \tag{12}$$

where h_{e_h} and h_{e_t} are entity representations computed by Eq.2, $W_h, W_t \in \mathbb{R}^{d\times 2d}$, $b_h, b_t \in \mathbb{R}^d$ are learnable parameters. Then we concatenate the representations of the head and tail entities to obtain the representation $t^{(h,r,t)} = [z_h^{(h,r,t)}, z_t^{(h,r,t)}] \in \mathbb{R}^{2d}$ of a single triple instance of a specific category. Finally, we average the representations of all triple instances corresponding to the same category r in the support document, aggregating them into a prototype representation for r:

$$\boldsymbol{p}^r = \frac{1}{|S_r|} \sum_{(e_h,r,e_t)\in S_r} t^{(h,r,t)} \tag{13}$$

where S_r is the set of all instances of relation r in support documents.

NOTA Prototype. In the query documents, most entities do not have any target relationship, so NOTA (None Of The Above) is also considered a category. To address this common scenario across all tasks, we employ a set of generic NOTA prototypes. However, similar to the case of evidence matching, the semantics of NOTA can vary for different episodes. Therefore, we utilize a task-specific NOTA Prototype construction strategy (Meng et al., 2023). This approach builds upon the generic NOTA prototype to better capture the unique NOTA semantics in each individual task.

Specifically, we first construct a set of learnable vectors $N_{base} = \{\boldsymbol{p}_i^{base} \in \mathbb{R}^{2d}\}_{i=1}^{N_{nota}}$, where N_{nota} represents the hyperparameter. Then we reinforce the semantic representation of the NOTA prototype for a specific task using representations of entity pairs without target relationships from supporting documents. For a NOTA instance $(e_h, nota, e_t)$, we use Eq.11 and Eq.12 for entity representations and combine them into an instance representation $t^{(h,nota,t)} = [z_h^{(h,nota,t)}; z_t^{(h,nota,t)}] \in \mathbb{R}^{2d}$. The set of representations of all entity pairs without target relationships is defined as T_{nota}. We adaptively select a NOTA instance from a specific task for each base NOTA prototype:

$$(e_h, nota, e_t) = \underset{(e_h,nota,e_t)\in T_{nota}}{argmax} (t^{(h,nota,t)} \cdot \boldsymbol{p}_i^{base} - \max_{r\in R_{episode}} t^{(h,nota,t)} \cdot \boldsymbol{p}_r) \tag{14}$$

which select the NOTA instances that are closer to the base NOTA prototype and farther from the target relationship prototype. Then we fuse the selected NOTA instance into the base NOTA prototype to obtain the final NOTA prototype $\boldsymbol{p}_i^{nota} \in \mathbb{R}^{2d}$ with the hyperparameter β:

$$\boldsymbol{p}_i^{nota} = \beta t^{(h,nota,t)} + (1-\beta)\boldsymbol{p}_i^{base} \tag{15}$$

3.3 Training Object

Building upon the various prototypes from supporting documents and meta-learning framework, we train the model using annotations from the query document and the model's predictions. Given an entity pair (e_h, e_t) in a query document, we use Eq.3 to obtain attention scores for each token of the entity pair. We then use the Eq.7 to obtain an instance-level contextual representation $c^{(h,t)}$ and the instance-level sentence representation $S_q = \{s_m^{(h,t)}\}_{m=1}^{N_s}$, where N_s is the number of sentences in the query document. Using Eqs.11 and 12, we compute the representation of the entity pair $q^{(h,t)} = [z_h^{(h,t)}, z_t^{(h,t)}]$. For each target relationship type r in the episode, we compute the probability of r as follows:

$$P_r^{(h,t)} = sigmoid(q^{(h,t)} \cdot \boldsymbol{p}^r - \max_{i\in Y}(q^{(h,t)} \cdot \boldsymbol{p}_i^{nota})) \tag{16}$$

where $Y = \{1, ..., N_{nota}\}$. We assess the similarity between evidence prototypes and sentence representations to calculate the probability of each sentence being evidence for the entity pair:

$$P_m^{(h,t)} = sigmoid(\max_{s_m^{(h,t)}\in S_q} (s_m^{(h,t)} \cdot \boldsymbol{m}_i) - \max_{s_m^{(h,t)}\in S_q} (s_m^{(h,t)} \cdot \boldsymbol{u}_i)) \tag{17}$$

We identify E as the set comprising all entity pairs within the query document, the relation classification loss is computed as follows:

$$\ell_{RE} = \frac{1}{|E|} \sum_{(e_h,e_t)\in E} - \sum_{r\in R_{episode}} (y_r^{(h,t)} log(P_r^{(h,t)}) + (1-y_r^{(h,t)}) log(1-P_r^{(h,t)})) \tag{18}$$

Where $y_r^{(h,t)}$ is set to 1 if the relationship r is present between the entity pair (e_h, e_t), and $y_r^{(h,t)}$ is set to 0 if it is not. We treat evidence matching as an auxiliary task within the meta-learning

framework, with its loss calculated as follows:

$$\ell_{EM} = \frac{1}{|E|} \sum_{(e_h,e_t)\in E} -\sum_{m=1}^{N_s} (y_m^{(h,t)} log(P_m^{(h,t)}) + (1 - y_m^{(h,t)}) log(1 - P_m^{(h,t)})) \quad (19)$$

where the value of $y_m^{(h,t)}$ is set to 1 when the sentence s_m serves as evidence for a relationship r between the entity pair, and $y_m^{(h,t)}$ is set to 0 when the sentence is not evidence for the entity pair. The overall training loss is computed as follows:

$$\ell = \ell_{RE} + \lambda \ell_{EM} \quad (20)$$

where λ is a hyperparameter. During inference, we extract the relation instance (e_h, r, e_t) in the query document if $q^{(h,t)} \cdot \boldsymbol{p}^r > \max_{i \in Y}(q^{(h,t)} \cdot \boldsymbol{p}_i^{nota})$.

4 Experiments

4.1 Datasets and Evaluation

We conduct experiments on two publicly available FSDLRE benchmarks, FREDo and ReFREDo, both providing evidence sentence annotations as additional auxiliary information for each sample.

FREDo benchmark comprises two tasks: in-domain and cross-domain, each with two sub-tasks, 1-DOC and 3-DOC, designed to measure the scalability of models under different settings. For in-domain tasks, both training and testing documents are sourced from DocRED, with a partitioning scheme ensuring disjoint relation types between them. The relation type set of DocRED is divided into three non-overlapping subsets: training (62), development (16), and in-domain test (18). FREDo uses the training set of DocRED as training and development document corpus, and its development set as the document corpus for in-domain testing. For in-domain tasks, a method trained on documents is evaluated on 15k episodes sampled from DocRED. In cross-domain tasks, training documents originate from DocRED, while testing documents are from SciERC, demonstrating significant differences in document themes, relationship types, and textual styles. FREDo utilizes the entire SciERC dataset as the cross-domain test document corpus. A method is first trained on documents sampled from DocRED, then evaluated on 3k episodes generated from documents in SciERC.

Benchmark	Task	N	$K(micro)$	$K(macro)$
FREDo	In-Domain 1-DOC	2.18	2.36	2.24
	In-Domain 2-DOC	3.47	4.30	4.31
ReFREDo	In-Domain 1-DOC	3.50	3.50	3.11
	In-Domain 2-DOC	5.67	6.50	5.73
FREDo	Cross-Domain 1-DOC	4.26	2.73	2.40
	Cross-Domain 2-DOC	6.08	5.55	5.27

Table 1: Average values for N and K across test episodes in FREDo and ReFREDo. K (micro) denotes the average across all episodes, K (macro) denotes the weighted average of mean K for each relation type.

ReFREDo is a revised version of FREDo, where the training, development, and in-domain test document corpus are replaced with Re-DocRED. Re-DocRED increases the number of relational facts in DocRED to 119,991. This expansion addresses the issue of missing labels and offers more comprehensive annotations. The division of relation types for each dataset in ReFREDo mirrors that of FREDo with 15k episodes sampled for in-domain evaluation. The cross-domain test episodes are constructed based on the entire SciERC dataset, as same as FREDo.

To better characterize the relationship of FSDLRE task to the traditional N-way K-shot format of few-shot tasks, we present the distribution of N and K in the test sets in Table 1.we provide detailed descriptions of the various baselines in Appendix A. We further elaborate on the application details of our model in Appendix B. An overview of the relation types and total instance number per relation of two benchmarks is listed in Appendix C.

4.2 Main Results

Our experimental results on FREDo and ReFREDo are shown in Table 2. Our PLEM model demonstrates notable enhancements in terms of macro F1 score, outperforming the RAPL model by an average of 1.17% F1 on FREDo and 1.29% F1 on ReFREDo, showcasing the advantage of our model. In various sub-settings of both benchmarks, our model consistently surpasses both the baselines and the existing state-of-the-art RAPL model, indicating its versatility. PLEM learns during training to more rationally construct relationship prototypes based on evidence information. The performance of relation extraction can be significantly improved by utilizing a small amount of relation and evidence annotations in a few-shot scenario for specific episodes. PLEM shows better performance on 3-Doc than on 1-Doc, demonstrating good scalabil-

Model	FREDo				ReFREDo			
	In-Domain		Cross-Domain		In-Domain		Cross-Domain	
	1-Doc F_1	3-Doc F_1	1-Doc F_1	3-Doc F_1	1-Doc F_1	3-Doc F_1	1-Doc F_1	3-Doc F_1
DL-Base	0.60	0.89	1.76	1.98	1.38	1.84	1.76	1.98
DL-MNAV	7.05 ± 0.18	8.42 ± 0.64	0.84 ± 0.16	0.48 ± 0.21	12.97 ± 0.88	12.43 ± 0.36	1.12 ± 0.38	2.28 ± 0.19
DL-MNAV$_{SIE}$	7.06 ± 0.15	6.77 ± 0.21	1.77 ± 0.60	2.51 ± 0.66	13.37 ± 0.98	12.00 ± 0.80	1.39 ± 0.74	2.92 ± 0.41
DL-MNAV$_{SIE+SBN}$	1.71 ± 0.24	2.79 ± 0.24	2.85 ± 0.12	3.72 ± 0.14	4.59 ± 0.30	5.43 ± 0.24	2.84 ± 0.24	3.86 ± 0.27
KDDocRE	2.59 ± 0.71	4.66 ± 0.83	1.03 ± 0.31	2.00 ± 0.46	4.76 ± 0.55	9.02 ± 0.64	2.30 ± 0.59	3.61 ± 0.43
Eider	2.75 ± 0.77	5.12 ± 0.63	0.98 ± 0.23	2.13 ± 0.52	5.23 ± 0.58	8.66 ± 0.66	2.35 ± 0.55	3.71 ± 0.32
RAPL	8.75 ± 0.80	10.67 ± 0.77	3.33 ± 0.50	5.35 ± 0.72	15.20 ± 0.82	16.35 ± 0.60	3.51 ± 0.79	5.48 ± 0.63
PLEM(Ours)	$\mathbf{10.04 \pm 0.70}$	$\mathbf{11.83 \pm 0.61}$	$\mathbf{4.52 \pm 0.55}$	$\mathbf{6.36 \pm 0.64}$	$\mathbf{16.40 \pm 0.62}$	$\mathbf{17.90 \pm 0.55}$	$\mathbf{4.88 \pm 0.85}$	$\mathbf{6.53 \pm 0.72}$

Table 2: Results on FREDo and ReFREDo benchmarks. The results reported are macro averages for all types of relationships. The scores of existing methods are borrowed from corresponding papers. The best performing method is indicated in bold.

ity. In in-domain settings, PLEM performs significantly better on ReFREDo compared to FREDo. However, in cross-domain settings, there is only a slight performance difference between the two benchmarks, suggesting that improve relationship and evidence annotation does not fully address cross-domain adaptation challenges. Besides, The performance of two supervised DocRE methods, KDDocRE and Eider, is not very impressive. This indicates that supervised methods, whether with additional evidence annotation or not, may not adapt well to few-shot scenarios.

4.3 Ablation Study

Model/ F1	In-Domain		Cross-Domain	
	1-Doc	3-Doc	1-Doc	3-Doc
PLEM	**16.40**	**17.90**	**4.88**	**6.53**
- TSEP	16.10	17.11	4.82	6.46
- EM	14.2	15.45	2.73	4.72
- EM + ER	14.82	16.03	3.15	5.01

Table 3: Ablation study on the In-Domain and Cross-Domain Subtasks of ReFREDo under 1-DOC and 3-DOC Settings.

To evaluate the impact of our proposed evidence matching module, we conduct a series of ablation experiments on the ReFREDo benchmark. The average results are shown in Table 3. The detailed analysis is outlined below:

For "- TSEM", we remove the task-specific evidence prototype construction method and use two base vector sets as evidence prototypes instead. This leads to a decrease in performance in In-Domain tasks, underscoring the effectiveness of utilizing episode-specific information. In Cross-Domain tasks, the elimination of this module results in almost no change in performance. We attribute this to the lack of participation of task-specific evidence information in the construction of relation prototypes during the inference process, and the model learning less task-specific information for cross-domain tasks during training.

For "- EM", we further eliminate the evidence matching task, training the model only with relation classification loss. The macro F1 score demonstrates a decrease of 2.33%, 1.98% in the In-Domain and Cross-Domain tasks, respectively, underscoring the efficacy of incorporating instance-relevant sentence information in the construction of relation prototypes.

For "- EM + ER", after removing the evidence matching task, we integrate the commonly used auxiliary task of Evidence Retrieval from supervised methods into our prototype learning framework. This task utilizes instance-to-sentence attention scores to determine the importance of sentences. We observe a significant decrease in model performance compared to PLEM, yet an improvement compared to scenarios where no evidence information is utilized. This finding indicates that the evidence matching task within the prototype framework has better generalizability, helping the model to more effectively focus on evidence information in few-shot scenarios, thereby improving the performance of relation extraction.

4.4 Impact of Hyperparameters

We explore the impact of different hyperparameters on model performance through experiments conducted under the 3-Doc task on ReFREDo. As illustrated in the figure, the evidence assistance coefficient λ is key in balancing evidence matching and relation classification losses. As λ increases, the macro F1 score first rises then falls, with the

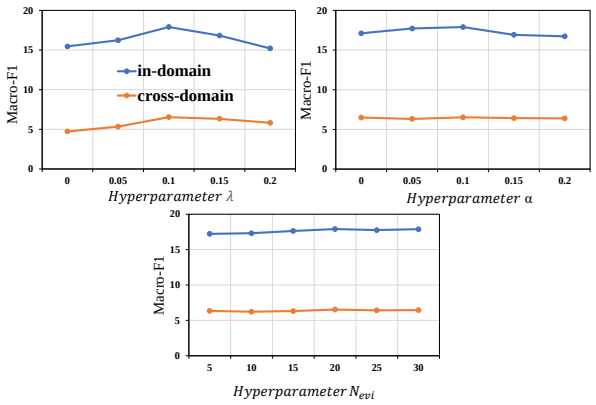

Figure 3: Effect of hyperparameters λ, α and N_{evi} on PLEM under the 3-Doc task setting in ReFREDo

optimal balance between ℓ_{RE} and ℓ_{ER} near 0.1. Clearly, the choice of λ significantly affects the performance of models on the FSDLRE task. Blindly increasing the proportion of support instances in the construction of evidence prototypes, denoted by α, may undermine the semantic representation of the evidence prototypes. Regarding the base number of evidence matching prototypes N_{evi}, it is observed that under the integration of the evidence matching task and task-specific evidence prototypes, its impact on the model is minimal.

5 Related Work

Document-Level Relation Extraction. Most existing research on document-level relation extraction is conducted in supervised scenarios. Current work can be divided into two main categories based on whether they explicitly model the interaction of information between entities: graph-based models and transformer-based models (Vaswani et al., 2017). Graph-based approaches (Zeng et al., 2020; Xu et al., 2021b; Duan et al., 2022; Lu et al., 2023) focus on building a document graph and explicitly learning the information among entities based on the constructed graph. Most current studies define three types of nodes: mentions, entities, and sentences, and connect these nodes using heuristic rules. Graph neural networks are employed for inference on the document with this graph-structured abstraction. Transformer-based methods (Xu et al., 2021a; Tan et al., 2022; Xiao et al., 2022; Xie et al., 2022; Ma et al., 2023b) take only the word sequence as input, implicitly modeling the long-distance context dependencies. Most approaches concentrate on extracting more expressive entity representations from the outputs of Transformers. Existing methods have achieved impressive results; however, these approaches rely on large-scale annotated relationships in documents, making them challenging to adapt to low-resource settings.

Few-Shot Document-Level Relation Extraction. To address the data scarcity issue in real-world DocRE scenarios, Popovic and Färber(2022) formulate the DocRE problem as a few-shot learning task and propose multiple metric-based baseline models. Moreover, Meng et al.(2023) propose a relation-aware prototype learning menthod , constructing prototypes at the instance level to better capture the semantic relations that prototypes represent. We note that, in the context of document-level relations, the process of constructing relation prototypes does not require attention to all contextual information, but rather focuses on context relevant to the instance. However, existing methods using attention-based context representation may introduce noise in the construction of relation prototypes. In this work, we propose a relation prototype learning method that integrates evidence matching, aiming to more effectively capture contextually relevant semantics of relation prototypes.

6 Conclusion

In this paper, we propose an evidence-enhanced prototype learning framework, PLEM, to improve FSDLRE by jointly extracting relations and evidence within a metric-based approach. During training, the evidence matching task enhances the representation of context by emphasizing the role of evidence sentences. This enhancement improves ability of model to depict relation prototypes. Additionally, it refines the method for measuring the proximity between instances and relation prototypes. Experimental results and further analysis demonstrate that PLEM significantly outperforms existing methods across various settings on two public benchmarks FREDo and ReFREDo, highlighting the superiority of our method.

Limitations

Our method has certain limitations that must be acknowledged. Firstly, our approach requires pre-annotated evidence information and entities in the training set, which might impact robustness of the model. Secondly, the addition of evidence instance construction and evidence matching loss calculation increases memory and time expenses. Lastly, PLEM's performance on cross-domain tasks is

lower, prompting us to continue exploring techniques to enhance performance on cross-domain tasks.

Large language models (LLM) have shown promising results in various few-shot tasks (Brown et al., 2020). Some efforts focus on leveraging LLM to tackle few-shot information extraction challenges (Ma et al., 2023a; Wadhwa et al., 2023; Ma et al., 2023c). Compared to the current progress of LLM in FSDLRE, our method demonstrates superior performance but necessitates the use of manually annotated training sets with relation categories that do not overlap with those of the test set. We contend that the potential of LLM in FSDLRE has not been fully explored. This motivates our further investigation into FSDLRE methods based on in-context learning (Rubin et al., 2022) and chain-of-thought (Wei et al., 2022).

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

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

Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240.

Zhichao Duan, Xiuxing Li, Zhenyu Li, Zhuo Wang, and Jianyong Wang. 2022. Not just plain text! fuel document-level relation extraction with explicit syntax refinement and subsentence modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1941–1951.

Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and S Yu Philip. 2021. Gradient imitation reinforcement learning for low resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2737–2746.

Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. 2021. Entity and evidence guided document-level relation extraction. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 307–315.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704.

Shuliang Liu, Xuming Hu, Chenwei Zhang, Lijie Wen, S Yu Philip, et al. 2022. Hiure: Hierarchical exemplar contrastive learning for unsupervised relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5970–5980.

Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. 2023. Anaphor assisted document-level relation extraction. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Xilai Ma, Jing Li, and Min Zhang. 2023a. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023b. 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*, pages 1963–1975.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023c. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*.

Shiao Meng, Xuming Hu, Aiwei Liu, Fukun Ma, Yawen Yang, Lijie Wen, et al. 2023. Rapl: A relation-aware prototype learning approach for few-shot document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5208–5226.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Nicholas Popovic and Michael Färber. 2022. Few-shot document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5733–5746.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Transactions of the Association for Computational Linguistics*, 9:691–706.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang, and Soujanya Poria. 2023. Uncertainty guided label denoising for document-level distant relation extraction. *arXiv preprint arXiv:2305.11029*.

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

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

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. *arXiv preprint arXiv:2305.05003*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Ying Wei and Qi Li. 2022. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2000–2008.

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 *2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 2395–2409. Association for Computational Linguistics (ACL).

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 2022*, pages 257–268.

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

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. Discriminative reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1653–1663.

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.

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. *arXiv preprint arXiv:2106.03618*.

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*, 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 *Conference on Empirical Methods in Natural Language Processing*.

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*, volume 35, pages 14612–14620.

A Baseline

We compare PLEM with both metric-based and supervised methods. **DL-Base** encodes documents using an unfinetuned bert-base (Devlin et al., 2019), averages the representations of entity mentions, and concatenates these representations to obtain embeddings for candidate pairs of entities. The similarity between relation embeddings in a query document and those in a support document is assessed by calculating the dot product between them. The relation type of the support embedding with the highest dot product is then output as the predicted relation

type. **DL-MNAV** (Popovic and Färber, 2022) is an expansion of the state-of-the-art method (Sabo et al., 2021) for sentence-level few-shot relation extraction to the document level, and integrates context pooling and adaptive threshold loss from ATLOP (Zhou et al., 2021). **DL-MNAV**$_{SIE}$ enhances cross-domain inference capabilities by using all individual support instances instead of their aggregated relation prototypes during inference. **DL-MNAV**$_{SIE+SBN}$ employs NOTA instances from the support documents as additional NOTA prototypes during training and exclusively uses NOTA vectors extracted from the support documents, disregarding the learned NOTA vectors, during cross-domain inference. **RAPL** (Meng et al., 2023) reframes the construction of relation prototypes at the instance level and further proposes a relation-weighted contrastive learning method to refine the representations of relation prototypes. Additionally, a task-specific NOTA prototype generation strategy is designed to more effectively capture the NOTA semantics in each task.

We also conduct comparisons using the supervised model **KDDocRE** (Tan et al., 2022), which operates without additional information, and the supervised model **Eider** (Xie et al., 2022), which constructs an auxiliary task of Evidence Retrieval using evidence annotations. The models are first trained on the entire divided training corpus and then fine-tuned on the support set to evaluate the performance of supervised models in few-shot scenarios.

B Implementation Details

Our model is implemented using the PyTorch (Paszke et al., 2019) library and HuggingFace Transformers (Wolf et al., 2019). All experiments are conducted with one RTX 4090 GPU. For a fair comparison, we utilize bert-base as the encoder in our method. AdamW is employed as the optimizer with a learning rate of 1e-5. We implement a linear warmup for the first 4% steps. The batch size is set to 2. We apply Gradient clipping with a maximum norm of 1.0. We train our model over 50,000 episodes and employ early stopping based on the macro F1 values on the development set. The hyperparameters k, N_{nota}, α, β, and N_{evi}, and λ are set to 15, 15, 0.1, 0.1, 20, 0.1 for in-domain tasks, and 10, 20, 0.05, 0.05, 20, 0.1 for cross-domain tasks. We report the mean and standard deviation of the macro F1 scores from training trials conducted with five different random seeds.

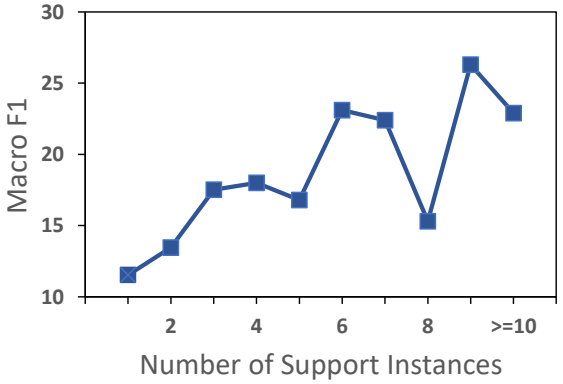

Figure 4: Performance of PLEM under different number of support relation instances on in-domain 3-Doc tasks of ReFREDo.

C Relation Types in benchmarks

In Tables 4 to 8, we list the types of relations for training, development, in-domain testing, and cross-domain testing document corpora in FREDo and ReFREDo. We present the name and description of each relation type.

D Number of Support Relation Instances

We analyze the impact of the number of support relation instances on PLEM's performance. We conduct experiments on the in-domain 3-Doc tasks in ReFREDo. We tally instances for each relation type within test episodes and sort them into 10 categories, with the first 9 for 1-9 instances, and the last for 10 or more. As illustrated in Figure 6. We observe that PLEM's performance generally shows an upward trend as the number of support relation instances increases, although there are fluctuations at certain points. This indicates that our method demonstrates some scalability, but its performance may not be perfectly positively correlated with the number of support relation instances.

Wikidata ID	Name	Description
P6	head of government	head of the executive power of this town, city, municipality, state, country, or other governmental body
P19	place of birth	most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character
P20	place of death	most specific known (e.g. city instead of country, or hospital instead of city) death location of a person, animal or fictional character
P22	father	male parent of the subject
P26	spouse	the subject has the object as their spouse (husband, wife, partner, etc.)
P30	continent	continent of which the subject is a part
P31	instance of	that class of which this subject is a particular example and member. (Subject typically an individual member with Proper Name label.)
P36	capital	primary city of a country, state or other type of administrative territorial entity
P37	official language	language designated as official by this item
P40	child	subject has the object in their family as their offspring son or daughter (independently of their age)
P54	member of sports team	sports teams or clubs that the subject currently represents or formerly represented
P58	screenwriter	author(s) of the screenplay or script for this work
P69	educated at	educational institution attended by the subject
P108	employer	person or organization for which the subject works or worked
P123	publisher	organization or person responsible for publishing books, periodicals, games or software
P127	owned by	owner of the subject
P131	located in the administrative territorial entity	the item is located on the territory of the following administrative entity
P155	follows	immediately prior item in some series of which the subject is part
P156	followed by	immediately following item in some series of which the subject is part
P159	headquarters location	specific location where an organization's headquarters is or has been situated
P161	cast member	actor performing live for a camera or audience
P162	producer	producer(s) of this film or music work (film: not executive producers, associate producers, etc.)
P166	award received	award or recognition received by a person, organisation or creative work
P170	creator	maker of a creative work or other object (where no more specific property exists)
P171	parent taxon	closest parent taxon of the taxon in question
P172	ethnic group	subject's ethnicity (consensus is that a VERY high standard of proof is needed for this field to be used. In general this means 1) the subject claims it him/herself, or 2) it is widely agreed on by scholars, or 3) is fictional and portrayed as such).
P175	performer	performer involved in the performance or the recording of a work
P178	developer	organisation or person that developed this item

Table 4: Relation types and description of training document corpus in FREDo and ReFREDo (continued on next page).

Wikidata ID	Name	Description
P190	sister city	twin towns, sister cities, twinned municipalities and other localities that have a partnership or cooperative agreement, either legally or informally acknowledged by their governments
P205	basin country	country that have drainage to/from or border the body of water
P206	located in or next to body of water	sea, lake or river
P241	military branch	branch to which this military unit, award, office, or person belongs
P264	record label	brand and trademark associated with the marketing of subject music recordings and music videos
P276	location	location of the item, physical object or event is within
P400	platform	platform for which a work has been developed or released / specific platform version of a software developed
P403	mouth of the watercourse	the body of water to which the watercourse drains
P449	original network	network(s) the radio or television show was originally aired on, including
P527	has part	part of this subject. Inverse property of "part of"
P551	residence	the place where the person is, or has been, resident
P569	date of birth	date on which the subject was born
P570	date of death	date on which the subject died
P576	dissolved, abolished or demolished	date or point in time on which an organisation was dissolved/disappeared or a building demolished
P577	publication date	date or point in time a work is first published or released
P580	start time	indicates the time an item begins to exist or a statement starts being valid
P585	point in time	time and date something took place, existed or a statement was true
P607	conflict	battles, wars or other military engagements in which the person or item participated
P676	lyrics by	author of song lyrics
P706	located on terrain feature	located on the specified landform
P710	participant	person, group of people or organization (object) that actively takes/took part in the event (subject)
P737	influenced by	this person, idea, etc. is informed by that other person, idea, etc.
P740	location of formation	location where a group or organization was formed
P749	parent organization	parent organization of an organisation, opposite of subsidiaries
P800	notable work	notable scientific, artistic or literary work, or other work of significance among subject's works
P807	separated from	subject was founded or started by separating from identified object
P840	narrative location	the narrative of the work is set in this location
P937	work location	location where persons were active
P1198	unemployment rate	portion of a workforce population that is not employed
P1336	territory claimed by	administrative divisions that claim control of a given area
P1344	participant of	event a person or an organization was a participant in, inverse of "participant"
P1365	replaces	person or item replaced
P1376	capital of	country, state, department, canton or other administrative division of which the municipality is the governmental seat
P1412	languages spoken, written or signed	language(s) that a person speaks or writes, including the native language(s)

Table 5: Relation types and description of training document corpus in FREDo and ReFREDo (continued).

Wikidata ID	Name	Description
P27	country of citizenship	the object is a country that recognizes the subject as its citizen
P150	contains administrative territorial entity	(list of) direct subdivisions of an administrative territorial entity
P571	inception	date or point in time when the organization/subject was founded/created
P50	author	main creator(s) of a written work (use on works, not humans)
P1441	present in work	work in which this fictional entity or historical person is present
P57	director	director(s) of this motion picture, TV-series, stageplay, video game or similar
P179	series	subject is part of a series, whose sum constitutes the object
P136	genre	a creative work's genre or an artist's field of work
P112	founded by	founder or co-founder of this organization, religion or place
P137	operator	person or organization that operates the equipment, facility, or service
P355	subsidiary	subsidiary of a company or organization, opposite of parent company
P176	manufacturer	manufacturer or producer of this product
P86	composer	person(s) who wrote the music
P488	chairperson	presiding member of an organization, group or body
P1056	product or material produced	material or product produced by a government agency, business, industry, facility, or process
P1366	replaced by	person or item which replaces another

Table 6: Relation types and description of development document corpus in FREDo and ReFREDo.

Wikidata ID	Name	Description
P17	country	sovereign state of this item; don't use on humans
P495	country of origin	country of origin of the creative work or subject item
P361	part of	object of which the subject is a part. Inverse property of "has part"
P3373	sibling	the subject has the object as their sibling (brother, sister, etc.)
P463	member of	organization or club to which the subject belongs
P102	member of political party	the political party of which this politician is or has been a member
P1001	applies to jurisdiction	the item (an institution, law, public office ...) belongs to or has power over or applies to the value (a territorial jurisdiction: a country, state, municipality, ...)
P140	religion	religion of a person, organization or religious building, or associated with this subject
P674	characters	characters which appear in this item (like plays, operas, operettas, books, comics, films, TV series, video games)
P194	legislative body	legislative body governing this entity; political institution with elected representatives, such as a parliament/legislature or council
P118	league	league in which team or player plays or has played in
P35	head of state	official with the highest formal authority in a country/state
P272	production company	company that produced this film, audio or performing arts work
P279	subclass of	all instances of these items are instances of those items; this item is a class (subset) of that item
P364	original language of work	language in which a film or a performance work was originally created
P582	end time	indicates the time an item ceases to exist or a statement stops being valid
P25	mother	female parent of the subject
P39	position held	subject currently or formerly holds the object position or public office

Table 7: Relation types and description of in-domain test document corpus in FREDo and ReFREDo.

Wikidata ID	Name	Description
HYPONYM-OF	hyponym of	subject is a hyponym of the object; subject is a type of the object.
PART-OF	part of	subject is a part of the object.
USED-FOR	used for	subject is used for the object; subject models the object; object is trained on the subject; subject exploits the object; object is based on the subject.
COMPARE	compare	compare two models/methods, or listing two opposing entities.
EVALUATE-FOR	evaluate for	evaluate for
FEATURE-OF	feature of	subject belongs to the object; subject is a feature of the object; subject is under the object domain.
CONJUNCTION	conjunction	function as similar role or use/incorporate with.

Table 8: Relation types and description of cross-domain test document corpus in SciERC.