

Enhancing Low-Resource Relation Representations through Multi-View Decoupling

Chenghao Fan^{1,2}, Wei Wei^{*1,2}, Xiaoye Qu^{1,2}, Zhenyi Lu^{1,2},
Wenfeng Xie³, Yu Cheng⁴, Dangyang Chen³

¹Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology

²Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)

³Ping An Property & Casualty Insurance Company of China, Ltd.

⁴The Chinese University of Hong Kong

facicofan@gmail.com, {weiw, quxiaoye}@hust.edu.cn, luzhenyi529@gmail.com, xiewenfeng801@pingan.com.cn
chengyu@cse.cuhk.edu.hk, chendangyang273@pingan.com.cn

Abstract

Recently, prompt-tuning with pre-trained language models (PLMs) has demonstrated the significantly enhancing ability of relation extraction (RE) tasks. However, in low-resource scenarios, where the available training data is scarce, previous prompt-based methods may still perform poorly for prompt-based representation learning due to a superficial understanding of the relation. To this end, we highlight the importance of learning high-quality relation representation in low-resource scenarios for RE, and propose a novel prompt-based relation representation method, named MVRE (Multi-View Relation Extraction), to better leverage the capacity of PLMs to improve the performance of RE within the low-resource prompt-tuning paradigm. Specifically, MVRE decouples each relation into different perspectives to encompass multi-view relation representations for maximizing the likelihood during relation inference. Furthermore, we also design a Global-Local loss and a Dynamic-Initialization method for better alignment of the multi-view relation-representing virtual words, containing the semantics of relation labels during the optimization learning process and initialization. Extensive experiments on three benchmark datasets show that our method can achieve state-of-the-art in low-resource settings.

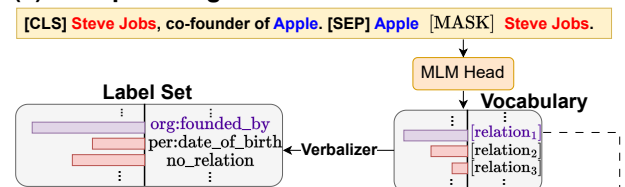
Introduction

Relation Extraction (RE) aims to extract the relation between two entities (Qu et al. 2023; Gu et al. 2022b) from an unstructured text (Cheng et al. 2021). Given the significance of inter-entity relations within textual information, the practice of relation extraction finds extensive utility across various downstream tasks, including dialogue systems (Lu et al. 2023; Liu et al. 2018), information retrieval (Dietz, Kotov, and Meij 2018; Yang 2020; Yu et al. 2023), information extraction (Zhu et al. 2023, 2021), and question answering (Yasunaga et al. 2021; Qu et al. 2021).

Following the emergence of the paradigm involving pre-trained models and fine-tuning for downstream tasks (Kenton and Toutanova 2019; Radford et al. 2018), many recent relation extraction studies have embraced the utilization of

*Corresponding author

(a) Prompt-tuning For RE



(b) Multi-view Decoupling

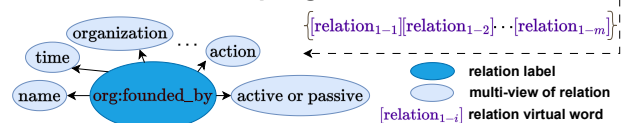


Figure 1: (a) An example of prompt-tuning for RE. Red-colored words indicate the subject, while blue-colored words indicate the object. (b) The concept of multi-view decoupling attempts to encompass various aspects of a relation using multiple relation representations.

large language models (Ye et al. 2020; Soares et al. 2019; Zhou and Chen 2022; Ye et al. 2022). In these works, the language models are integrated with classification heads and fine-tuned specifically for relation extraction tasks, resulting in promising results. However, the effective training of additional classification heads becomes challenging in situations where task-specific data is scarce. This challenge arises from the disparity between pre-training tasks, such as masked language modeling, and the subsequent fine-tuning tasks encompassing classification and regression. This divergence hampers the seamless adaptation of pre-trained language models (PLMs) to downstream tasks.

Recently, prompt tuning has emerged as a promising direction for facilitating few-shot learning, which effectively bridges the gap between the pre-training and the downstream task (Gao, Fisch, and Chen 2021; Jin et al. 2023). Conceptually, prompt-tuning involves template and verbalizer engineering, aiming to discover optimal templates and answer spaces. For example, as shown in Figure 1 (a), given a sentence “Steve Jobs, co-founder of Apple” for

relation extraction, the text will first be enveloped with relation-specific templates, namely transforming the original relation extraction task into a relation-oriented cloze-style task. Subsequently, the PLM will predict words in the vocabulary to fill in the [MASK] position, and these predicted words are finally mapped to corresponding labels through a verbalizer. In this example, the filled word “[relation₁]” (e.g., “founded”) can be linked to the label “org:founded_by” through the verbalizer. However, for complex relation representations, such as “per: country_of_birth” and “org: city_of_headquarters,” obtaining suitable vocabulary labels is much more challenging. To address this issue, previous work (Han et al. 2022) applies logic rules to decompose complex relations into descriptions related to the subject and object entity types. Some works construct virtual words for each relation (a trainable “[relation₁]”) to substitute the corresponding answer space of the complex relation (Chen et al. 2022b,a). This paradigm focuses on optimizing the relation representation space and demands PLMs to learn representations for words that are not present in the vocabulary. However, in extremely low-resource scenarios, such as one-shot RE, building robust relation representations with this paradigm is difficult, thus leading to a performance drop.

To mitigate the above issue, in this paper, we introduce Multi-view Relation Extraction (MVRE), which improves low-resource prompt-based relation representations with a multi-view decoupling framework. As illustrated in Figure 1 (b), considering that relations may contain multiple dimensions of information, for instance, “org:founded_by” may entail details about organizations, people’s names, time, the action of founding, and so on. According to theoretical analysis, being limited to a single vector representation, the model may face the upper boundary of representation capacity and fail to construct robust representations in low-resource scenarios. Therefore, we propose to optimize the latent space by decoupling it into a joint optimization of multi-view relation representations, thereby maximizing the likelihood during relation inference. By sampling a greater number of relation representations, as denoted “[relation_{1-*i*]” in Figure 1 (b)), we promote the learned latent space to include more kinds of information about the corresponding relation. In detail, we achieve this decoupling process by disassembling the virtual words into multiple components and predicting these components through successive [MASK] tokens. Furthermore, we introduce a Global-Local loss and Dynamic Initialization approach to optimize the process of relation representations by constraining semantic information of relations. We evaluate MVRE on three relation extraction datasets. Experimental results demonstrate that our method significantly outperforms previous approaches. To sum up, our main contributions are as follows:}

- To the best of our knowledge, this paper presents the first attempt to improve low-resource prompt-based relation representations with multi-view decoupling learning. In this way, the PLM can be comprehensively utilized for generating robust relation representations from limited

data.

- To optimize the learning process of multi-view relation representations, we introduce the Global-Local Loss and Dynamic Initialization to impose semantic constraints between virtual relation words.
- We conduct extensive experiments on three datasets and our proposed MVRE can achieve state-of-the-art performance in low-resource scenarios.

Background and Related Work

Prompt-Tuning for RE

Inspired by the “in context learning” proposed in GPT-3 (Brown et al. 2020), the approach of stimulating model knowledge through a few prompts has recently attracted increasing attention. In text classification tasks, significant performance gains can be achieved by designing a tailored prompt for a specific task, particularly in few-shot scenarios (Schick and Schütze 2021; Gao, Fisch, and Chen 2021). In order to alleviate the labor-intensive process of manual prompt creation, there has been extensive exploration into automatic searches for discrete prompts (Schick, Schmid, and Schütze 2020; Wang, Xu, and McAuley 2022) and continuous prompts (Huang et al. 2022; Gu et al. 2022a).

For RE with prompt-tuning, a template function can be defined in the following format: $T(x) = x : w_s : [\text{MASK}] : w_o$, where “:” signifies the operation of concatenation. By employing this template function, the instance x is modified to incorporate the entity pair (w_s, w_o) , resulting in the formation of $x_{prompt} = T(x)$. In this process, x_{prompt} is the corresponding input of model M with a [MASK] token in it. Here, Y refers to the label words set, and \mathcal{V} donates the relation set within the prompt-tuning framework. A verbalizer v is a mapping function $v : Y \rightarrow \mathcal{V}$, establishing a connection between the relation set and the label word set, where $v(y)$ means label words corresponding to label y . The probability distribution over the relation set is calculated as:

$$p(y|x) = p_M([\text{MASK}] = v(y)|T(x)) \quad (1)$$

In this way, the RE problem can be transferred into a masked language modeling problem by filling the [MASK] token in the input.

However, for relation extraction, the complexity and diversity of relations pose challenges in employing these methods to discover suitable templates and answer spaces. Han et al. (2022) propose prompt-tuning methods for RE by applying logic rules to construct hierarchical prompts. Lu et al. (2022) make prompts for each relation and converts RE into a generative summarization problem. These works translate the prediction of a relation into predicting a specific sentence, which to some extent addresses the complexity of relations. However, summarizing the intricate information of a relation using these words remains challenging.

Virtual Relation Word

Chen et al. (2022b) introduce virtual relation words and leverage prompt-tuning for RE by injecting semantics of relations and entity types. Chen et al. (2022a) propose retrieval-enhanced prompt-tuning by incorporating retrieval

of representations obtained through prompt-tuning. These studies devise virtual words for each relation in prompt-tuning, circumventing the need to search for complex answer spaces (Liu et al. 2023).

The corresponding verbalizer v^* for this approach function as $v^* : Y \rightarrow \mathcal{V}^*$, where $\mathcal{V}^* = \{\mathcal{V}, \mathcal{V}^Y\}, |Y| = |\mathcal{V}^Y|, v^*(y) \in \mathcal{V}^Y, y \in Y$. The \mathcal{V}^Y corresponds to virtual relational words, representing the set of words created for each relation. The acquisition of this virtual word for a relation is equivalent to obtaining a latent space representation for that relation. As the relation virtual words do not exist in the pre-trained model’s vocabulary, ensuring robust representations often requires a sufficient amount of data or semantic constraints to the prompt-based instance representation (Chen et al. 2022a).

Given an instance x , the prompt-based instance representation h^x can be computed by leveraging the output embedding of the “[MASK]” token of the last layer of the underlying PLM:

$$h^x = M(T(x))_{[\text{MASK}]} \quad (2)$$

The prompt-based instance representation h^x can capture the relation corresponding to the instance x , and ultimately, through the “MLM head”, derive the classification probabilities for the respective virtual relation word (Chen et al. 2022b,a). Most of these approaches confine a complex relation to a single prompt-based vector, which limits the learning of relation latent space in low-resource scenarios.

Method

Preliminaries

Formally, a RE dataset can be denoted as $D = \{X, Y\}$, where X is the set of examples and Y is the set of relation labels. For each example $x = \{w_1, w_2, w_s, \dots, w_o, \dots, w_n\}$, the goal of RE is to predict the relation $y \in Y$ between subject entity w_s and object entity w_o (since an entity may have multiple tokens, we simply utilize w_s and w_o to represent all entities briefly).

Previous Prompt-Tuning in Standard Scenario In the prompt-based instance learning for relations, it is assumed that for each class y_i , we learn a latent space representation H_{y_i} such that $F^{-1}(y_i) = H_{y_i}$, where F denotes the function mapping between labels and representations. In the case of a standard scenario, where all available data can be used, the model minimizes the following loss function:

$$\mathbb{E}_{x \sim \mathcal{X}} [-\log p(y|x)] = -\frac{1}{N} \sum_{i=1}^N \log p(y_i, H_{y_i}|x_i) \quad (3)$$

where N represents the total data volume across all classes. In this process, focusing solely on a specific relation y_e , the learned latent space representation $\hat{H}_{y_e}^{\text{standard}}$ for class y_e satisfies $F(h^{x_i^e}) = y_e$, where $1 \leq i \leq \#y_e$ and $(x_i^e, y_e) \in (\mathcal{X}, \mathcal{Y})$. Here, $\#y_e$ represents the number of instances in the data with the label y_e . The process of obtaining $\hat{H}_{y_e}^{\text{standard}}$ is akin to optimizing the following expression:

$$\min_{\theta} \sum_{(x_i^e, y_e) \in (\mathcal{X}, Y)} \text{sim}(H_{y_e}, F^{-1}(y_e, \theta)) \quad (4)$$

where “*sim*” represents the degree of similarity between the latent space representations. However, in low-resource scenarios, the value of $\#y_e$ can constrain the optimization effectiveness of Eq 4.

Multi-View Decoupling Learning

Therefore, we assume that in the process of learning the complex relation latent space H_{y_i} , it is feasible to decompose this space into multiple perspectives and learn from various viewpoints. Consequently, we consider the learning process for single data pair (x_i, y_i) as follows:

$$\begin{aligned} p(y_i, H_{y_i}|x_i) &= \sum_h p(y_i, h|x_i) \\ &= \sum_h p(y_i|x_i, h)p(h|x_i) \\ &= \mathbb{E}_{h \sim p(h|x_i)} p(y_i|x_i, h) \end{aligned} \quad (5)$$

where h represents a perspective in which the relation y_i is decomposed, we transform the learning of relations into the process of learning each relation’s various perspectives. Ultimately, we merge the information from multiple perspectives to optimize the relation inference process.

Similar to Eq 4, when there is only one pair of data for a given relation, the learning of its latent space is as follows:

$$\min_{\theta} \sum_{(x^e, y_e) \in (\mathcal{X}, \mathcal{Y}), y_e^j \in y_e} \text{sim}(H_{y_e}, F^{-1}(y_e^j, \theta)) \quad (6)$$

In this process, the learned latent space representation $\hat{H}_{y_e}^{1\text{-shot}}$ for class y_e satisfies $F(h_j^{x^e}) = y_e$, where $1 \leq j \leq m$ and $(x^e, y_e) \in (\mathcal{X}, \mathcal{Y})$. Here, m represents the number of decomposed perspectives for the relation y_e .

Sampling of Relation Latent Space Under normal circumstances, the latent space learned in a low-resource setting tends to be inferior compared to the standard scenario, resembling $\text{sim}(\hat{H}_{y_e}^{1\text{-shot}}, H_{y_e}) \geq \text{sim}(\hat{H}_{y_e}^{\text{standard}}, H_{y_e})$. Hence, as can be seen in Figure 1 (a), our objective is for the low-resource acquired latent space to closely resemble that learned in the standard scenario, as $E(\hat{H}_{y_e}^{1\text{-shot}}) \sim E(\hat{H}_{y_e}^{\text{standard}})$. Combining Eq 4 and Eq 6, the representation set $\{h_j^{x^e} | 1 \leq j \leq m\}$ we acquire needs to resemble the representation set $\{h^{x_i} | 1 \leq i \leq \#y_e\}$ obtained under standard conditions. This highlights the necessity of sampling a substantial number of $h_j^{x^e}$ ($m \geq 1$) instances with similar distribution to ensure alignment of the obtained relation latent space with that in standard scenarios. The value of m will be discussed in the experimental section.

According to the Eq 2, h is determined by the parameters of model M , the structure of template T , and the expression “[MASK] = $v(y_i)$ ”:

$$\begin{aligned} p(y_i|x_i, h^{x_i}) &= p(y_i|x_i, M(T(x_i))_{[\text{MASK}]}) \\ &= p_M([\text{MASK}] = v(y_i)|T(x_i)) \end{aligned} \quad (7)$$

To ensure a consistent interpretation of h^{x_i} obtained from single data pair, while simultaneously covering various perspectives of a relation, we sample h^{x_i} based on the expression “[MASK] = $v(y_i)$ ”. Specifically, we expand the token

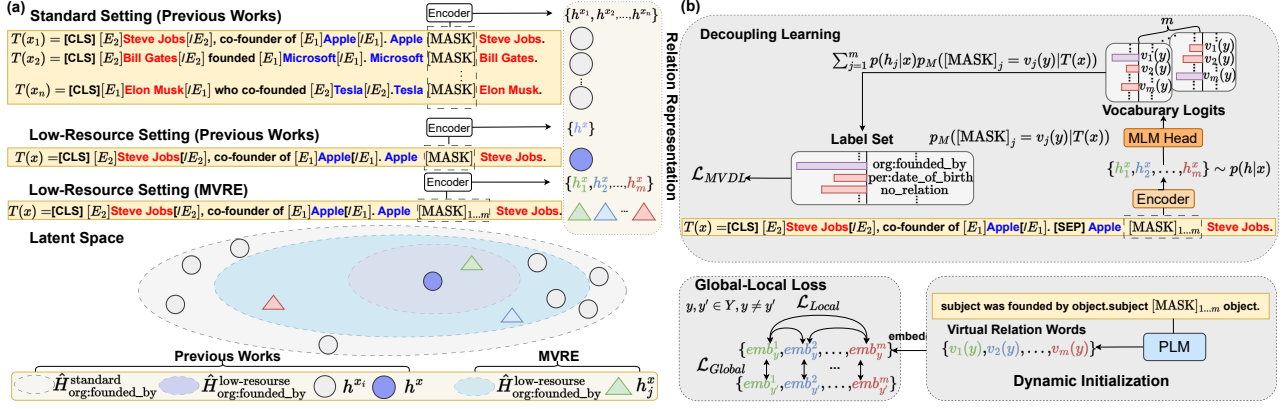


Figure 2: (a) An illustrative comparison of the relation latent space learning process between MVRE and previous prompt-based works. We employ multi-view relation representations to cover a broader latent space in low-resource scenarios. (b) The MVRE framework incorporates Multi-view Decoupling Learning, Global-Local Loss and Dynamic Initialization processes.

“[MASK]” into multiple contiguous tokens within the template, each “[MASK]” corresponds to as follows:

$$T(x) = x : [sub] : [MASK]_{\{1..m\}} : [obj] \quad (8)$$

the sampling method for $h_j^{x_i}$ is as follows, $h_j^{x_i} = M(T(x))_{[MASK]_j}$. It’s important to note that a relation in text can be represented by a continuous segment of text. Therefore, this approach has the potential to capture multi-view representations of a relation.

Based on our sampling method for latent space representation, we derive the probability distribution of y_i as follows:

$$p(y_i|x_i, h_j^{x_i}) = p_M([MASK]_j = v_j(y_i)|T(x_i)) \quad (9)$$

Due to the challenge of finding suitable words in the vocabulary to match different perspectives of a relation, we introduce m new multi-view virtual relation words, denoted as $v_j(y)$, for each relation y_i . Combining Eq 5, the final loss function $\mathcal{L}_{MVDL}(x_i, y_i)$ that the model needs to minimize is as follows:

$$\sum_{j=1}^m \left[-\log \left(p(h_j^{x_i}|x_i) p_M([MASK]_j = v_j(y_i)|T(x_i)) \right) \right] \quad (10)$$

Here, we employ a matrix W_h to learn the posterior probability of $h_j^{x_i}$, the formula is as follows $p(h_j^{x_i}|x_i) =$

$$\frac{\sigma(W_h^T h_j^{x_i})}{\sum_{k=1}^m \sigma(W_h^T h_k^{x_i})},$$

where σ represents the sigmoid function.

Considering all the data, the loss function is given by:

$$\mathcal{L}_{MVDL} = \sum_{(x_i, y_i) \in (X, Y)} \mathcal{L}_{MVDL}(x_i, y_i) \quad (11)$$

Global-Local Loss The contrastive learning methods to enhance representation learning have been employed in many previous works (Gao, Yao, and Chen 2021; Zhang et al. 2022). To encourage better alignment of multi-view virtual relation words $v_j(y)$ with diverse semantic meanings, we introduce the Global-Local Loss(referred to as “GL”)

to optimize the learning process of multi-view relation virtual words. The Local Loss encourages virtual words representing the same relation to focus on similar information, while the Global Loss ensures that virtual words representing different relations emphasize distinct aspects. Their expressions are as follows:

$$\mathcal{L}_{Local} = -\frac{1}{|Y|m^2} \sum_{r \in Y} \left[\sum_{i, j \in [1, m]} \text{sim}(emb_r^i, emb_r^j) \right]$$

$$\mathcal{L}_{Global} = \frac{1}{|Y|^2 m} \sum_{i=1}^m \left[\sum_{ru, rv \in \mathcal{R}} \text{sim}(emb_{ru}^i, emb_{rv}^i) \right] \quad (12)$$

where $\text{sim}(x, y) = \cos(\frac{x}{\|x\|}, \frac{y}{\|y\|})$, emb_r^i denotes the embedding of the virtual word for relation $v_i(r)$.

Finally, the loss function of MVRE is as follows:

$$\mathcal{L}_{MVRE} = \mathcal{L}_{MVDL} + \alpha * \mathcal{L}_{Local} + \beta * \mathcal{L}_{Global} \quad (13)$$

where α and β are hyperparameters. The framework of MVRE is illustrated in Figure 2 (b).

Dynamic Initialization The virtual word for a relation also involves learning a new word that does not exist in the original vocabulary. Therefore, efficient initialization is crucial for achieving desirable results in this process. However, in MVRE, it is essential to have meaningful initialization methods that consider the actual positions of each virtual word in the text.

We introduce Dynamic Initialization (referred to as “DI”), which leverages the PLM’s cloze-style capability to identify appropriate initialization tokens for relation-representing virtual words. Specifically, we first create a manual template for each relation and insert a prompt after it. Then, we employ the model to find the token with the highest probability, which serves as the initialization token for the respective virtual word. To enhance the construction of relation information, we incorporate the entity information corresponding to the label itself. This knowledge is not involved in

Dataset	Train	Dev	Test	Relation
SemEval	6,507	1,493	2,717	19
TACRED	68,124	22,631	15,509	42
TACREV	68,124	22,631	15,509	42

Table 1: The statistics of different RE datasets.

the model’s training process and is similar to prompts, as it leverages the inherent abilities of the model, thus preserving the characteristics of low-resource scenarios.

To mitigate the potential generation of irrelevant tokens during dynamic initialization, particularly with larger m values, we merge the static and dynamic initialization techniques. Inspired by Chen et al. (2022b), we introduce **Static Initialization** (referred to as “SI”), where words for initialization are derived from the labels corresponding to each relation. We integrate the two methods by averaging the tokens’ embedding obtained from different initializations.

Experiments

Datasets

For comprehensive experiments, we conduct experiments on three RE datasets: SemEval 2010 Task 8 (SemEval) (Hendrickx et al. 2010), TACRED (Zhang et al. 2017), and TACRED-Revisit (TACREV) (Alt, Gabryszak, and Hennig 2020). Here we briefly describe them below. The detailed statistics are provided in Table 1.

SemEval is a traditional dataset in relation extraction that does not provide entity types. It covers 9 relations with two directions and one special relation “Other”.

TACRED is one large-scale sentence-level relation extraction dataset drawn from the yearly TACKBP4 challenge, which contains 41 common relation types and a special “no relation” type.

TACREV builds on the original TACRED dataset. They find out and correct the errors in the original development set and test set of TACRED, while the training set is left intact. TACREV and TACRED share the same set of relation types.

Compared Methods

To evaluate our proposed MVRE, we compare with the following methods: (1) **FINE-TUNING** employs a conventional fine-tuning approach for PLMs to relation extraction. (2) **GDPN_{ET}** utilizes the multi-view graph for relation extraction (Xue et al. 2021) (3) **PTR** (Han et al. 2022) propose prompt-tuning methods for RE by applying logic rules to partition relations into sub-prompts; (4) **Know-Prompt** (Chen et al. 2022b) utilize virtual relation word to prompt-tuning; (5) **RetrieveRE** (Chen et al. 2022a) employ retrieval to enhance prompt-tuning.

Implementation Details

We utilize Roberta-large for all experiments to make a fair comparison. For test metrics, we use micro F_1 scores of RE

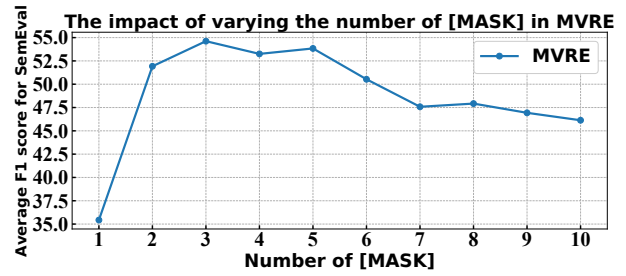


Figure 3: Effect of the number of [MASK] on MVRE.

as the primary metric to evaluate models, considering that F_1 scores can assess the overall performance of precision and recall.

Low-Resource Setting. we adopt the same setting as RetrievalRE (Chen et al. 2022a) and perform experiments using 1-, 5-, and 16-shot scenarios to evaluate the performance of our approach in extremely low-resource situations. To avoid randomness, we employ a fixed set of seeds to randomly sample data five times and record the average performance and variance. During the sampling process, we select k instances for each relation label from the original training sets to compose the few-shot training sets.

Standard Setting. In the standard setting, we leverage full trainsets to conduct experiments and compare with previous prompt-tuning methods, including PTR, Know-Prompt, and RetrievalRE.

Low-Resource Results

We present our results on low-resource settings in Table 2. Notably, across all datasets, our MVRE consistently outperforms all previous prompt-tuning models. Particularly remarkable is the substantial improvement in the 1-shot scenario, with gains of 63.9%, 8.7%, and 9.6% over RetrievalRE in SemEval, TACRED, and TACREV respectively. When k is set to 5 or 16, the magnitude of improvement decreases. In the TACRED and TACREV datasets, when k is set to 16, there’s a slight decrease compared to the retrieval-enhanced RetrievalRE. However, overall, the performance remains better than KnowPrompt, a fellow one-stage prompt-tuning method similar to ours. Similar to previous works (Chen et al. 2022b,a), the comparison of performance between fine-tuning-based methods (FINE-TUNING, GDPN_{ET}) and MVRE demonstrates the superiority of prompt-based methods in low-resource settings.

It’s noteworthy that our method doesn’t exhibit the same significant improvements in TACRED and TACREV as observed in SemEval. Our speculation is attributed to two reasons: (1) In TACRED and TACREV, the high proportion of “other” relations (78% in TACRED/V, 17% in SemEval) can make it challenging to categorize relations as “other” in the low-resource scenario. (2) There are more similar relations than SemEval, such as “org:city_of_headquarters” and “org:stateorprovince_of_headquarters”, making it more difficult to distinguish them in low-resource scenarios.

Model	SemEval			TACRED			TACREV		
	K=1	K=5	K=16	K=1	K=5	K=16	K=1	K=5	K=16
<i>Compared Methods</i>									
FINE-TUNING	18.5(±1.4)	41.5(±2.3)	66.1(±0.4)	7.6(±3.0)	16.6(±2.1)	26.8(±1.8)	7.2(±1.4)	16.3(±2.1)	25.8(±1.2)
GDPN _{ET}	10.3(±2.5)	42.7(±2.0)	67.5(±0.8)	4.2(±3.8)	15.5(±2.3)	28.0(±1.8)	5.1(±2.4)	17.8(±2.4)	26.4(±1.2)
PTR	14.7(±1.1)	53.9(±1.9)	80.6(±1.2)	8.6(±2.5)	24.9(±3.1)	30.7(±2.0)	9.4(±0.7)	26.9(±1.5)	31.4(±0.3)
KnowPrompt	28.6(±6.2)	66.1(±8.6)	80.9(±1.6)	17.6(±1.8)	28.8(±2.0)	34.7(±1.8)	17.8(±2.2)	30.4(±0.5)	33.2(±1.4)
RetrievalRE	33.3(±1.6)	69.7(±1.7)	81.8(±1.0)	19.5(±1.5)	30.7(±1.7)	36.1 (±1.2)	18.7(±1.8)	30.6(±0.2)	35.3 (±0.3)
<i>Ours</i>									
MVRE(w/o GL&DL)	35.3(±4.6)	74.6(±1.7)	81.3(±1.4)	21.0(±2.1)	31.4(±1.0)	32.9(±2.5)	20.2(±0.7)	31.0 (±1.1)	34.1(±2.1)
MVRE	54.6 (±2.8)	77.6 (±3.6)	82.5 (±0.8)	21.2 (±2.2)	32.4 (±1.2)	34.8(±0.8)	20.5 (±1.9)	31.0 (±1.4)	34.3(±1.1)

Table 2: Performance of RE models in the low-resource setting. We report the mean and standard deviation performance of micro $F1$ scores (%) over 5 different splits. The best numbers are highlighted in each column.

x=The National Congress of American Indians was founded in 1944 in response to the implementation of assimilation policies on tribes by the federal government.		
[sub]=National Congress of American Indians [obj]=1944		
m	top-1 token; T(x)=x [sub] [MASK]*m [obj]	top-1 token; T(x)=x [obj] [MASK]*m [sub]
1	in(0.42)	.(0.31)
2	founded(0.48) in(0.70)	.(0.18) The(0.19)
3	was(0.87) founded(0.92) in(0.93)	.(0.08) of(0.59) the(0.48)
4	was(0.46) was(0.16) founded(0.19) in(0.55)	</s>(0.07) of(0.03) of(0.53) the(0.55)
5	was(0.44) founded(0.31) in(0.29) founded(0.03) ,(0.74)	</s> (0.09) the(0.05) founding(0.09) of(0.63) the(0.70)
x=The series reflected on the changes that had taken place in Ireland since the 1960s.		
[sub]=series [obj]=changes		
m	top-1 token; T(x)=x [sub] [MASK]*m [obj]	top-1 token; T(x)=x [obj] [MASK]*m [sub]
1	on(0.20)	the(0.41)
2	reflected(0.24) those(0.34)	in(0.53) the(0.83)
3	reflected(0.69) on(0.87) those(0.40)	to(0.06) in(0.18) the(0.64)
4	reflected(0.15) on(0.10) on(0.27) those(0.41)	are(0.12) reflected(0.05) throughout(0.35) the(0.69)
5	reflected(0.08) the(0.12) some(0.06) of(0.22) those(0.43)	that(0.10) been(0.08) reflected(0.07) in(0.30) the(0.69)

Table 3: Case study of Dynamic Initialization. Each line represents the top-1 token generated for each [MASK] and its corresponding probability when the number of [MASK] is m . We highlight the parts that represent the relation more accurately

Method	GL	SI	DI	K=1	K=5	K=16	Full
MVRE	✓	✓	✓	54.6	77.6	82.5	90.2
		✓	✓	54.6	77.1	82.1	89.3
	✓		✓	44.9	74.1	82.4	89.8
			✓	43.3	73.1	82.5	89.5
	✓	✓		37.5	72.9	81.5	89.5
		✓	35.3	74.6	81.3	89.9	
Prompt-tuning Pre-trained Model(For Reference)							
PTR				14.7	53.9	80.6	89.9
KnowPrompt		✓		28.6	66.1	80.9	90.2
RetrievalRE		✓		33.3	69.7	81.8	90.4

Table 4: Ablation Study on SemEval: Investigating the Impact of Global-Local Loss (GL), Static Initialization (SI), and Dynamic Initialization (DI). The "Full" column indicates the results under the standard setting.

Ablation Study

To prove the effects of the components of MVRE, including Global-Local Loss(GL), Dynamic Initialization(DI), and Static Initialization(SI), we conduct the ablation study on SemEval and present the results in Table 4. Additionally, we present the results under the standard setting in Table 4.

Standard Results Under the full data scenario, MVRE and KnowPrompt yield equivalent results, indicating that our approach remains applicable and does not compromise model performance when enough data is available.

Global-Local Loss As observed in Table 4, the incorporation of the Global-Local Loss(GL) consistently yields improvements across various scenarios, resulting in an enhancement of the relation $F1$ score by 0.5, 0.4, and 0.5 in the 5-shot, 16-shot, and standard settings, respectively. This phenomenon demonstrates that constraining the semantics of virtual relation words' embedding through a comparative method can optimize the representation of multi-perspective relations.

The Initialization of Virtual Relation Words We also conduct an ablation study to validate the effectiveness of the initialization of relation virtual words. Previous studies have revealed that achieving satisfactory relation representations with random initialization is challenging (Chen et al. 2022b). Hence, to ensure model performance, it is essential to use either Static Initialization(SI) or Dynamic Initialization(DI) during the experiment. When both are employed simultaneously, their corresponding tokens' embedding is av-

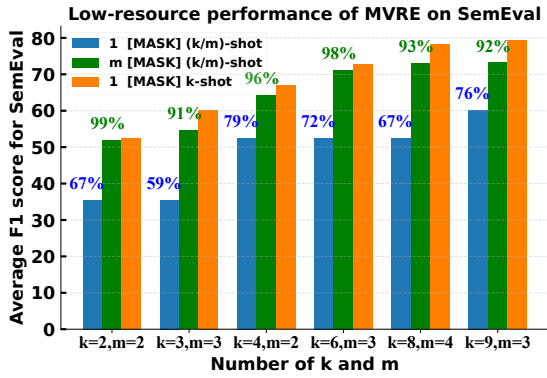


Figure 4: MVRE under low-resource conditions vs. MVRE with only one [MASK] under more resource-rich conditions.

eraged to integrate these two methods. Table 4 demonstrates that adopting Dynamic Initialization leads to a significant enhancement in model performance compared to Static Initialization. Furthermore, combining both initialization methods also yields substantial improvements.

Effect of m Number of [MASK] Due to the introduction of noise when inserting “[MASK]” and further, the efficiency of decoupling learning presents significant challenges. Therefore, simply increasing the number of “[MASK]” tokens cannot enhance performance in low-resource scenarios. As shown in Figure 3, we conduct experiments to investigate the impact of varying quantities of “[MASK]” tokens on relation extraction effectiveness, aiming to identify the optimal value for m . The performance of the model shows a trend of initially increasing and then decreasing as the value of m increases. Specifically, the value of m reaches its peak within the range of [3, 5]. As m increases from 1 to 3, there is a sudden improvement in performance, indicating that the decoupling of relation latent space into multiple perspectives contributes significantly to the construction of relation representations. However, when $m \geq 5$, the model’s performance exhibits a gradual decline. This trend suggests that with a higher number of consecutive “[MASK]” tokens, the prompt-based instance representation obtained by the model tends to contain more noise, thereby adversely affecting the overall model performance.

Case Study of Dynamic Initialization

We illustrate the feasibility of multiple “[MASK]” tokens and the effectiveness of our Dynamic Initialization through a case study, presented in Table 3.

Specifically, for a sentence x , we wrap it into $T(x)$ and input $T(x)$ into the model (RoBERTa-large). At each “[MASK]” position, we obtain the token with the highest probability from the model. This token represents the word that the model identifies as best representing the relation based on the given sentence. During the Dynamic Initialization process, we utilize the embedding of the token with the highest probability to initialize the corresponding position of the virtual relation word.

Given the existence of many relations with reversed subject and object roles in the dataset, it is challenging to model them accurately without confusion. Therefore, in Table 3, we illustrate our method’s unique treatment of relations that are mutually passive and active by interchanging the subject and object orders (we controlled the active and passive voice of relations by swapping the order of [sub] and [obj]). It can be observed that, by increasing the number of [MASK] tokens, RoBERTa-large in the zero-shot scenario effectively captures both active (“was founded in” and “reflected on”) and passive (“the founding of” and “been reflected in”) voice forms for these two relations. However, when there is only one [MASK] token, the generated tokens are largely unrelated to these relations. This indicates that increasing the number of [MASK] tokens enables the PLM to utilize a broader range of words to depict a complex relation, potentially enhancing the PLM’s capacity to represent relations.

Effectiveness of Low-Resource Decoupling Learning

We conduct experiments to demonstrate the effectiveness of decoupling learning in MVRE, which can be formalized as the following equation in our methods: $E(\hat{H}_{y_e}^{1-shot}) \sim E(\hat{H}_{y_e}^{standard})$. To evaluate the effectiveness of our proposed method, we compare the performance in scenarios with relatively low and enough resources. To be specific, we compare MVRE with one [MASK] against MVRE with m [MASK]. One-[MASK] MVRE is tested in k -shot settings, while m -[MASK] MVRE is tested in (k/m) -shot settings, ensuring consistent relation representation sampling. Additionally, we test one-[MASK] MVRE in (k/m) -shot scenarios for result comparison. The results are as shown in Figure 4. We employ the proportion of model result similarity to represent the overall similarity of obtained relation representations, as represented by the formula: $sim(H-model1, H-model2) = \frac{F1-score-model1}{F1-score-model2}$. Experimental results show that, with an equal number of h , the similarity of relation representations obtained under low-resource scenarios surpasses 90% when compared to higher-resource scenarios. This indicates a 20% improvement over the one-[MASK] approach. This demonstrates that decoupling relation representations into multi-view perspectives can significantly enhance relation representation capabilities in low-resource scenarios.

Conclusion

In this paper, we present MVRE for relation extraction, which improves low-resource prompt-based relation representations with multi-view decoupling. Meanwhile, we propose the Global-Local Loss and Dynamic Initialization techniques to constrain the semantics of virtual relation words, optimizing the learning process of relation representations. Experimental results demonstrate that our method significantly outperforms existing state-of-the-art prompt-tuning approaches in low-resource settings.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110,

No. 62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

- Alt, C.; Gabryszak, A.; and Hennig, L. 2020. TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1558–1569.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Chen, X.; Li, L.; Zhang, N.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022a. Relation Extraction as Open-book Examination: Retrieval-enhanced Prompt Tuning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2443–2448.
- Chen, X.; Zhang, N.; Xie, X.; Deng, S.; Yao, Y.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2022b. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*, 2778–2788.
- Cheng, Q.; Liu, J.; Qu, X.; Zhao, J.; Liang, J.; Wang, Z.; Huai, B.; Yuan, N. J.; and Xiao, Y. 2021. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2819–2831.
- Dietz, L.; Kotov, A.; and Meij, E. 2018. Utilizing knowledge graphs for text-centric information retrieval. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 1387–1390.
- Gao, T.; Fisch, A.; and Chen, D. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3816–3830.
- Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910.
- Gu, Y.; Han, X.; Liu, Z.; and Huang, M. 2022a. PPT: Pre-trained Prompt Tuning for Few-shot Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8410–8423.
- Gu, Y.; Qu, X.; Wang, Z.; Zheng, Y.; Huai, B.; and Yuan, N. J. 2022b. Delving deep into regularity: a simple but effective method for Chinese named entity recognition. *arXiv preprint arXiv:2204.05544*.
- Han, X.; Zhao, W.; Ding, N.; Liu, Z.; and Sun, M. 2022. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3: 182–192.
- Hendrickx, I.; Kim, S. N.; Kozareva, Z.; Nakov, P.; Séaghdha, D. O.; Padó, S.; Pennacchiotti, M.; Romano, L.; and Szpakowicz, S. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals. *ACL 2010*, 33.
- Huang, Y.; Qin, Y.; Wang, H.; Yin, Y.; Sun, M.; Liu, Z.; and Liu, Q. 2022. FPT: Improving Prompt Tuning Efficiency via Progressive Training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6877–6887.
- Jin, F.; Lu, J.; Zhang, J.; and Zong, C. 2023. Instance-aware prompt learning for language understanding and generation. *ACM Transactions on Asian and Low-Resource Language Information Processing*.
- Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.
- Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.
- Liu, S.; Chen, H.; Ren, Z.; Feng, Y.; Liu, Q.; and Yin, D. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1498.
- Lu, K.; Hsu, I.-H.; Zhou, W.; Ma, M. D.; and Chen, M. 2022. Summarization as Indirect Supervision for Relation Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6575–6594.
- Lu, Z.; Wei, W.; Qu, X.; Mao, X.; Chen, D.; and Chen, J. 2023. MIRACLE: Towards Personalized Dialogue Generation with Latent-Space Multiple Personal Attribute Control. *arXiv preprint arXiv:2310.18342*.
- Qu, C.; Zamani, H.; Yang, L.; Croft, W. B.; and Learned-Miller, E. 2021. Passage retrieval for outside-knowledge visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1753–1757.
- Qu, X.; Zeng, J.; Liu, D.; Wang, Z.; Huai, B.; and Zhou, P. 2023. Distantly-supervised named entity recognition with adaptive teacher learning and fine-grained student ensemble. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13501–13509.
- Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.
- Schick, T.; Schmid, H.; and Schütze, H. 2020. Automatically Identifying Words That Can Serve as Labels for Few-Shot Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5569–5578.

- Schick, T.; and Schütze, H. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269.
- Soares, L. B.; Fitzgerald, N.; Ling, J.; and Kwiatkowski, T. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2895–2905.
- Wang, H.; Xu, C.; and McAuley, J. 2022. Automatic Multi-Label Prompting: Simple and Interpretable Few-Shot Classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5483–5492.
- Xue, F.; Sun, A.; Zhang, H.; and Chng, E. S. 2021. Gdpnet: Refining latent multi-view graph for relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14194–14202.
- Yang, Z. 2020. Biomedical information retrieval incorporating knowledge graph for explainable precision medicine. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2486–2486.
- Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 535–546.
- Ye, D.; Lin, Y.; Du, J.; Liu, Z.; Li, P.; Sun, M.; and Liu, Z. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7170–7186.
- Ye, D.; Lin, Y.; Li, P.; and Sun, M. 2022. Packed Levitated Marker for Entity and Relation Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4904–4917.
- Yu, S.; Fan, C.; Xiong, C.; Jin, D.; Liu, Z.; and Liu, Z. 2023. Fusion-in-T5: Unifying Document Ranking Signals for Improved Information Retrieval. arXiv:2305.14685.
- Zhang, S.; Liang, Y.; Gong, M.; Jiang, D.; and Duan, N. 2022. Multi-View Document Representation Learning for Open-Domain Dense Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5990–6000.
- Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*.
- Zhou, W.; and Chen, M. 2022. An Improved Baseline for Sentence-level Relation Extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 161–168.
- Zhu, T.; Qu, X.; Chen, W.; Wang, Z.; Huai, B.; Yuan, N. J.; and Zhang, M. 2021. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph. arXiv preprint arXiv:2112.06013.
- Zhu, T.; Ren, J.; Yu, Z.; Wu, M.; Zhang, G.; Qu, X.; Chen, W.; Wang, Z.; Huai, B.; and Zhang, M. 2023. Mirror: A Universal Framework for Various Information Extraction Tasks. arXiv preprint arXiv:2311.05419.