Clear Up Confusion: Advancing Cross-Domain Few-Shot Relation Extraction through Relation-Aware Prompt Learning

Ge Bai [*], **Chenji Lu**[*], **Daichi Guo**[*], **Shilong Li, Ying Liu, Zhang Zhang, Guanting Dong, Ruifang Liu, Yong Sun**[†]
Beijing University of Posts and Telecommunications, China

Abstract

Cross-domain few-shot Relation Extraction (RE) aims to transfer knowledge from a source domain to a different target domain to address low-resource problems. Previous work utilized label descriptions and entity information to leverage the knowledge of the source domain. However, these models are prone to confusion when directly applying this knowledge to a target domain with entirely new types of relations, which becomes particularly pronounced when facing similar relations. In this work, we propose a relation-aware prompt learning method with pre-training. Specifically, we empower the model to clear confusion by decomposing various relation types through an innovative label prompt, while a context prompt is employed to capture differences in different scenarios, enabling the model to further discern confusion. Two pre-training tasks are designed to leverage the prompt knowledge and paradigm. Experiments show that our method outperforms previous sota methods, yielding significantly better results on cross-domain few-shot RE tasks.

1 Introduction

Relation Extraction (RE) is one of the key tasks of Natural Language Processing (NLP), which aims to identify the relations between given entities. Traditional supervised models (Zhang et al., 2017; Tran et al., 2019; Peng et al., 2020a; Yamada et al., 2020) have impressive performance in RE tasks. However, collecting sufficient data for certain classes may be laborious in practice. Inspired by the advances in few-shot learning (Mishra et al., 2018; Nichol et al., 2018), finetuning prompt-based pretrained language models have shown superior performance in few-shot RE (He et al., 2023; Liu et al., 2022) and some other tasks (Lee et al., 2021; Cui et al., 2021; Dong et al., 2023b; Sun et al., 2023).

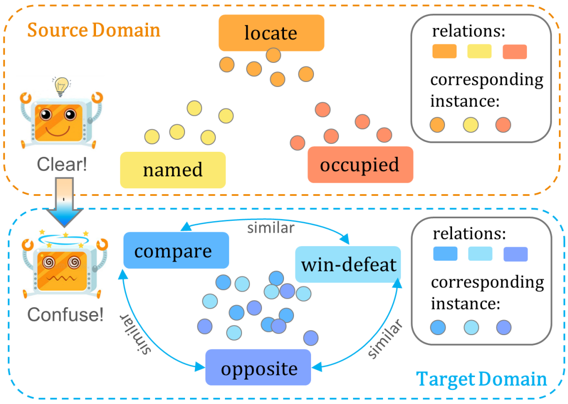

Figure 1: Even though the model performs well in the source domain, the entirely new scenarios in the target domain make it challenging to differentiate between easily confused types of relations.

However, these methods lack robustness in cross-domain scenarios, which is particularly important in low-resource RE tasks.

Domain adaptation methods (Ganin et al., 2016; Shen et al., 2018; Li et al., 2022, 2023) offer new insights by transferring knowledge between domains through shared feature representations extracted from multiple domains. However, these methods only work when the classes in the source and target domain have the same labels (Gao et al., 2019). To better learn the knowledge from the source domain, Soares et al., 2019 build task-agnostic relation representations solely from the entity-linked text. Zhang and Lu, 2022 proposed a label descriptions prompt dropout approach to leverage the label information, which helps the model learn class representations effectively. Nevertheless, as shown in Figure 1, simply relying on memorizing and understanding elements from the source domain does not assist the model in maintaining better performance in the target domain (Wu et al., 2021; Liu et al., 2023; Dong et al., 2023a). Therefore, the model needs the ability to figure out the connections and distinctions between relations to decompose various relation types.

[*]The first three authors contribute equally.
[†]Yong Sun is the corresponding author.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 70–78
June 16-21, 2024 ©2024 Association for Computational Linguistics

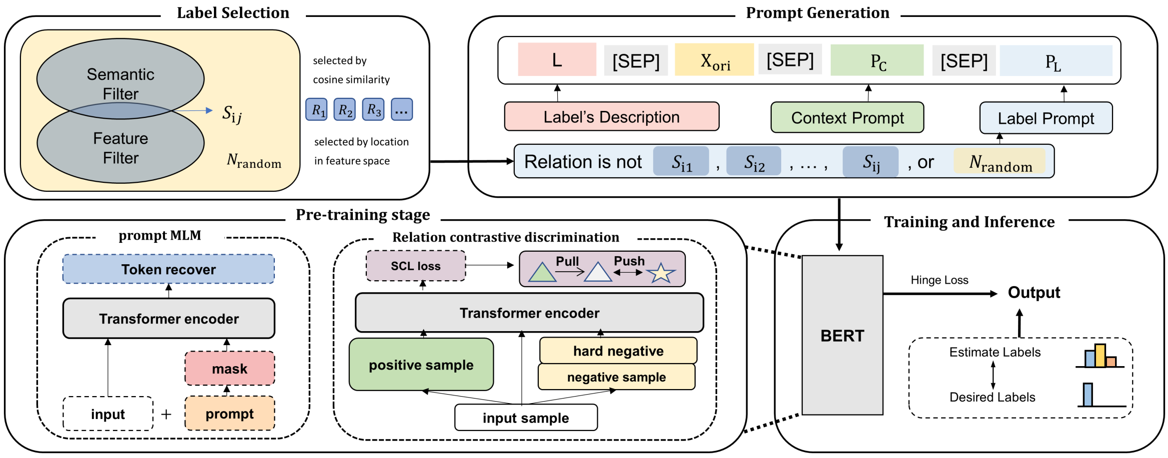

Figure 2: The overall architecture of ClearRE framework. The framework includes selecting relations by two filters, prompt generation, pre-training stage with two different tasks, and training and inference.

To bypass this issue, we propose a novel relation-aware prompt learning approach with pre-training (ClearRE). Specifically, we design a label prompt to assist the model in distinguishing similar relations, which tend to be confusing during predictions. We identify similar relations by introducing two filters: semantic filter and feature filter. Meanwhile, a context prompt is proposed to further capture the differences among relations in different scenarios. We combine these two components into our ultimate prompt, effectively enhancing the model's ability in cross-domain scenarios by alleviating confusion. Motivated by Chen et al. (2022); Dong et al. (2023c), we further design two pre-training tasks to facilitate the utilization of prompts. The first task is prompt-based MLM, which learns the prompt paradigm by predicting masked tokens in prompt sentences. The second task is relation contrastive discrimination in which we construct positive, negative, and hard negative samples and employ contrastive learning to enhance the ability of the model to differentiate between relations, especially for similar ones.

We summarize the contributions as follows:

- We design a context prompt and a label prompt to enable the model to decompose various relations in different domains.
- We propose two pre-training tasks, namely prompt MLM and relation contrastive discrimination, for learning the prompt paradigm and distinguishing similar relations.
- Experiments on three benchmarks show that ClearRE outperforms previous state-of-the-art methods in all scenarios. A visualization study demonstrates the effectiveness of our approach in clearing confusion.

2 Methodology

The overall framework of ClearRE is shown in Figure 2. Section 2.1 briefly illustrates the task definition of RE. Section 2.2 shows the details of similar relations selection. Section 2.3 explains the design of the prompt. Section 2.4 provides a comprehensive interpretation of prompt MLM and relation contrastive discrimination pre-training tasks. Training and inference are shown in Appendix A

2.1 Task Definition

$X_{ori} = \{x_1, x_2, ..., x_m\}$ is the original input sentence, including m tokens. Entity position $E_{key} = \{e_{head}, e_{tail}\}$ refers to head and tail entities spans which warped with special tokens [E1], [/E1], [E2] and [/E2] following ERNIE (Zhang et al., 2019). RE tasks aim to learn a mapping function: $f : (X_{ori}, E_{key}) \rightarrow y$, where y is the label.

2.2 Similar Relation Filters

The identification of similar relations is essential for the model to alleviate confusion. Thus, we propose two strategies to filter out similar relations in different aspects.

Semantic filter: Relation types with more semantically similar label descriptions are generally prone to confusion. We collect the label descriptions of the training instances and feed them into sentence-transformers (Reimers and Gurevych, 2019) to generate embeddings. For each relation, we calculate the cosine similarity with other relations and select relations with top-k highest scores as the final candidates.

Feature filter: For similar relations, the overall feature distributions of their corresponding samples are close. Therefore, we randomly select multiple samples for each relation type and calculate the average of their embeddings as the overall representations in feature space. For each relation, we calculate the Euclidean distances between representations and select top-k relations that are closest as the final candidates.

We set $k = 5$ and take the intersection of these two candidate sets as the final similar relation set. Specifically, for a relation type a, we define the set of similar relations as $R_a = \{S_{a1}, S_{a2}, ..., S_{aj}...\}$, where S_{aj} are similar relations. Relations that are not in this set are non-similar relations.

2.3 Prompt Generation

We designed an innovative label prompt to assist the model in differentiating between relations. The similar relations obtained by filters are introduced into the prompt, aiming to focus on those confusing labels. To avoid the model over-avoiding similar relations and making radical predictions, we also include a randomly selected non-similar relation in the template. The Label Prompt P_L is of the following form: "***relation is not*** S_{i1}***, ...,*** S_{ij}... ***or*** N_{random}", where S_{ij} are from the similar relations set R_i. N_{random} represents the non-similar relation which is selected randomly.

Additionally, we construct a context prompt P_C to further capture the differences in application scenarios of relations. We extract the context by replacing each entity in the sentence with a special token [BLANK]. The final input instance is formed by concatenating the correct label L, original sentence X_{ori}, context prompt P_C, and label prompt P_L in sequence with special tokens [SEP].

$$X_{all} = L\ [SEP]\ X_{ori}\ [SEP]\ P_C\ [SEP]\ P_L \qquad (1)$$

2.4 Pre-training Task

As shown in figure 2, we propose two pre-training tasks to help the model learn the prompt format and teach the model how to distinguish relation types.

Prompt MLM. We follow the design of masked language model(MLM) in BERT (Devlin et al., 2019) and utilize this method on the context prompt and label prompt. This pre-training task allows the model to fit the corresponding parts of the prompt and learn how to extract useful information from them. Specifically, we randomly select the words in the context or labels in the prompt and replace them with [MASK]. All tokens will be masked if the labels consists of multiple tokens. We set the loss function of prompt MLM as:

$$\mathcal{L}_{MLM} = -\sum\nolimits_{n=1}^{M} \log P(x_n) \qquad (2)$$

where M is the number of masked tokens and $P(x_n)$ is the predicted probability of token x_n over the vocabulary size.

Relation Contrastive Discrimination. We introduce relation contrastive discrimination to optimize the distribution of relations in semantic space, which equips the model with a better ability to differentiate confusing relation types. We construct positive and negative samples as follows: Given an input X with relation type R, for positive samples, we randomly choose the samples with the same label to construct K positive samples. For negative samples, we use other samples in the batch. In particular, we use our similar relation filter to construct several hard negative samples that contain similar relation types. We add it to the negative samples to guide the model focusing on these confusing labels. Considering multiple positive samples, we employ supervised contrastive learning (SCL) (Khosla et al., 2020) to learn the robust representation. We define h_p and h_p^+ as the representation of input utterances and positive samples. We formulate SCL as follows:

$$\mathcal{L}_s = \frac{1}{N}\sum_{p=1}^{N} -\frac{1}{N_{y_p}-1}\sum_{q=1}^{N_{y_p}} \log \frac{e^{sim(h_p,h_q^+)/\tau}}{\sum_{k=1}^{N}\mathbf{1}_{p\neq k}e^{sim(h_p,h_k)/\tau}} \qquad (3)$$

where N is the total number of examples in the batch and N_{y_p} is the number of positive pairs in the batch. τ is a temperature hyperparameter and $sim(h_1, h_2)$ is cosine similarity $\frac{h_1^\top h_2}{||h_1||\cdot||h_2||}$. $\mathbf{1}$ is an indicator function.

We weight both pre-training objectives together as the final loss function by a hyperparameter α:

$$\mathcal{L}_{final} = \alpha\mathcal{L}_s + (1-\alpha)\mathcal{L}_{MLM} \qquad (4)$$

3 Experiments

3.1 Datasets and Implementation Details

We evaluate our approach on two Few-shot RE datasets: **CrossRE** (Bassignana and Plank, 2022): A manually-curated corpus contains 5265 sentences covering 6 domains with a unified label set of 17 relation types. To assess the domain adaptation of the model, we conduct experiments on CrossRE in single-source domain and multiple-source domain scenarios. **FewRel**: FewRel 1.0 (Han et al., 2018) is collected from Wikipedia articles, which contain 100 relations and 700 instances for each relation. FewRel 2.0 (Gao et al., 2019) contains a test set from the biomedical, which contains 25 relations and 100 instances for each relation.

Models	5-way-1-shot		5-way-5-shot		10-way-1-shot		10-way-5-shot		Avg.	
	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single
Proto-Bert*	67.70±0.5	52.2±0.7	80.71±1.0	64.65±0.8	58.65±0.9	39.86±1.2	76.82±1.1	50.82±00.8	70.97	51.83
HCRP*	70.47±1.0	60.34±0.9	85.05±0.3	70.68±1.5	59.17±0.5	47.53±0.6	78.51±1.0	60.70±0.9	73.30	60.06
IDA*	70.51±0.9	60.60±1.0	85.51±0.8	71.39±0.7	62.13±1.1	47.22±0.6	78.53±0.5	62.15±1.1	74.17	60.34
CP*	78.33±0.9	49.96±0.7	86.89±1.1	70.70±1.2	70.95±1.1	44.45±0.9	78.36±1.4	53.82±0.7	78.63	54.73
LPD*	81.90±0.8	62.35±0.5	86.87±1.4	75.39±0.5	69.81±1.7	48.39±1.2	78.65±0.5	63.36±0.9	79.31	62.37
ClearRE	**84.52**±0.7	**64.37**±1.1	**88.86**±0.3	**76.97**±0.6	**74.05**±0.4	**50.07**±1.4	**79.41**±0.8	**65.40**±0.9	**81.71**	**64.20**

Table 1: Accuracy (%) of cross-domain few-shot classification on CrossRE. We choose the music domain as the target domain. (* These works haven't been evaluated on CrossRE, the results are produced by our implementation.)

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Proto-Bert	40.12	51.50	26.45	36.93
BERT-PAIR	67.41	78.57	54.89	66.85
HCRP	76.34	83.03	63.77	72.94
IDA	76.30	84.71	67.87	75.84
CP	79.70	84.90	68.10	79.80
LPD	82.81±0.5	88.98±1.4	70.51±1.5	78.76±1.6
ClearRE	**84.68**±0.4	**91.60**±0.7	**73.88**±0.6	**83.92**±1.0

Table 2: Accuracy (%) of cross-domain few-shot classification on the FewRel2.0 test set.

Methods	**Fewrel2.0**		**CrossRE**	
	1-shot	5-shot	single	multi
ClearRE	**79.78**	**87.76**	**64.20**	**81.71**
w/o Context Prompt	77.65	86.09	61.57	80.00
w/o Label Prompt	77.42	85.37	61.66	80.24
w/o Pre-training	73.32	81.92	57.74	75.25
–w/o Prompt MLM	75.58	84.03	60.30	78.51
–w/o Contrastive Discrimination	75.04	83.58	59.14	77.48

Table 3: The ablation study results (average accuracy %) for Fewrel2.0 and CrossRE.

We follow the official split to use 64 relations of Fewrel 1.0 for training, 16 for validation and use FewRel 2.0 for testing to evaluate the domain adaptation of few-shot models. More implementation details are shown in Appendix B.

We compare our method with the following baseline methods: Proto-BERT (Snell et al., 2017), BERT-PAIR (Gao et al., 2019), CP (Peng et al., 2020b), HCRP (Han et al., 2021), Improved Domain Adaption (IDA) (Yuan et al., 2023), LPD (Zhang and Lu, 2022). **Proto-BERT** is a prototypical network with BERT-base (Devlin et al., 2019) serving as the backbone. **BERT-PAIR** is a method that measures the similarity of a sentence pair. **CP** pretrains Proto-BERT using a contrastive pre-training approach that divides sentences into positive pairs and negative pairs. **HCRP** equips Proto-BERT with a hybrid attention module and a task adaptive focal loss. **Improved Domain Adaption (IDA)** proposes an encoder learned by optimizing a representation loss and an adversarial loss to extract the relation of sentences in the source and target domain. **LPD** introduces a label prompt dropout training approach that is adaptable to cross-domain tasks.

3.2 Results and Analysis

Main Results: Table 1 and 2 report the main results compared with other baselines in cross-domain few-shot RE tasks. For CrossRE, our approach outperforms all baseline models in CrossRE, achieving an average improvement of at least 2.40% and 1.83% for multiple and single source domain scenarios respectively. Results demonstrate the effectiveness of ClearRE in different cross-domain scenarios. Meanwhile, our method has an average improvement of 3.25% over the previous sota LPD on Fewrel 2.0, indicating that ClearRE has better adaptability in a new target domain by focusing on context and distinguishing different relations. To further demonstrate the effectiveness of our method, we conducted an additional in-domain experiment, which is shown in Appendix C.

Ablation Studies: We construct ablation experiments on FewRel2.0 and CrossRE datasets to investigate the contribution of each component in our approach. We implement a w/o context prompt and a w/o label prompt experiment by removing the corresponding prompts in the training process. Table 3 indicates that the absence of any prompt weakens the ability of the model to distinguish relations, ultimately leading to a decrease in performance. We also conduct experiments to validate the effectiveness of the pre-training tasks (w/o Pre-training). Results show that removing any task leads to a significant decline in model performance, which demonstrates both the effectiveness and necessity of pre-training. Appendix D shows more detailed ablation experiments.

3.3 Comparison with Large Language Models

With the emergence of LLMs, the increasing number of traditional RE tasks are being solved using LLMs and achieving excellent results. Considering that they have been exposed to a sufficient amount of diverse data during the training

Model	5-way 1-shot
ChatGPT	80.44
ChatGPT+our prompt design	82.86
ClearRE	84.68

Table 4: Comparision with Large Language Models with our prompt design.

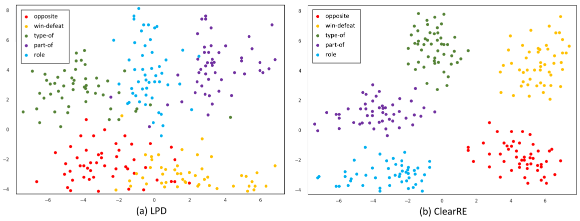

Figure 3: the visualization results for 250 samples from 5 labels in a 5-way-1-shot scenario on CrossRE dataset.

process, we did not directly evaluate cross-domain tasks using these models. However, we believe that conducting comparative experiments for the few-shot settings is necessary. We conducted the comparison experiments with ChatGPT. We tested two types of prompts under the 5-way 1-shot setting. The first type is the conventional Few-shot RE prompt: *for sentence i, the relation of 'head entity' and tail entity' is 'relation X'. Based on the 5 example above, which one is the relation of 'test head entity' and 'test tail entity' in sentence 'test sentence' in the relation list: relation X1, X2, X3, X4, X5?*. The second type is our designed prompt with Label Description, Contextual prompt and Label prompt. We conducted testing on FewRel 2.0 and present the results in Table 4. It's worth noting that generative models are sensitive to prompts when performing discriminative tasks. Our testing results may exhibit some minor deviations. This experiment mainly reflects the differences between models and settings. Under the few-shot setting, there is still some gap between the performance of LLMs and supervised small models. However, our prompt design helps ChatGPT achieve better results.

3.4 Visualization

As shown in Figure 3, we conduct a visualization study and compare the results of LPD (a) and ClearRE (b) to verify the effectiveness of our method in clearing up confusion. We choose 5 similar relations by our filters and collect the vector representations of the test samples along with their labels during the process of model forwarding. The t-SNE toolkit (Van der Maaten and Hinton, 2008) is used to map the high-dimensional feature space onto a two-dimensional plane, allowing for the measurement of sample similarity based on these representations. The results show that ClearRE makes samples of the same relations more compact while increasing the distance between different relations, demonstrating the improvement of the ability to distinguish confusing relations.

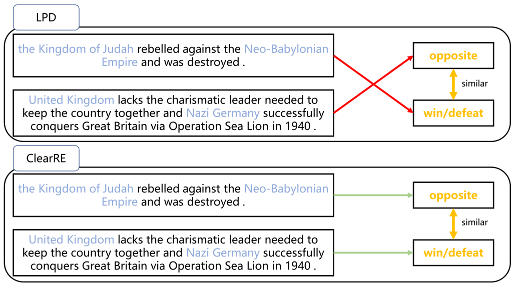

Figure 4: A case from CrossRE. The red lines in the figure represent incorrect predictions, while the green lines indicate correct predictions.

4 Case Study

To further verify the effectiveness of our method, we randomly sample 50 instances from the output and choose the most representative case in Figure 4. "win/defeat" and "opposite" are similar relations chosen by our similar filters. "win/defeat" means something that is physically or idealistically opposite, contrary, against or inverse of something else. "opposite" means someone or something who has won or lost a competition, an award, a war. Both two sentences are related to a war scenario. The baseline model confuses these two similar relations while our method solves this tricky problem. This case suggests that ClearRE can distinguish confusing relations in an entirely new domain.

5 Conclusion

In this paper, we propose a relation-aware contextual prompt learning approach with pre-training for cross-domain few-shot RE. We design a novel prompt containing context information and confusing relations. Two pre-training tasks further enable the model to adapt to the prompt format and learn to distinguish confusing relations under different conditions. Extensive experiments and analyses demonstrate the effectiveness of our method.

6 Limitations

Some limitations exist in our work. Our effectiveness is only examined on the task of relation extraction, while whether this method is able to generalize to other information extraction tasks, such as named entity recognition (NER) and event detection (ED), is not yet explored in this paper. In addition, our work only discusses the effectiveness of the current prompt design for clearing up confusion, but the effect of the formal transformation of the prompt on the effectiveness is also not discussed in detail in the paper.

References

Elisa Bassignana and Barbara Plank. 2022. CrossRE: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qianglong Chen, Feng-Lin Li, Guohai Xu, Ming Yan, Ji Zhang, and Yin Zhang. 2022. Dictbert: Dictionary description knowledge enhanced language model pre-training via contrastive learning.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using bart. *arXiv preprint arXiv:2106.01760*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Guanting Dong, Tingfeng Hui, Zhuoma GongQue, Jinxu Zhao, Daichi Guo, Gang Zhao, Keqing He, and Weiran Xu. 2023a. Demonsf: A multi-task demonstration-based generative framework for noisy slot filling task. *arXiv preprint arXiv:2310.10169*.

Guanting Dong, Zechen Wang, Liwen Wang, Daichi Guo, Dayuan Fu, Yuxiang Wu, Chen Zeng, Xuefeng Li, Tingfeng Hui, Keqing He, Xinyue Cui, Qixiang Gao, and Weiran Xu. 2023b. A prototypical semantic decoupling method via joint contrastive learning for few-shot named entity recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Guanting Dong, Zechen Wang, Jinxu Zhao, Gang Zhao, Daichi Guo, Dayuan Fu, Tingfeng Hui, Chen Zeng, Keqing He, Xuefeng Li, Liwen Wang, Xinyue Cui, and Weiran Xu. 2023c. A multi-task semantic decomposition framework with task-specific pre-training for few-shot ner. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 430–440, New York, NY, USA. Association for Computing Machinery.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1).

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of EMNLP-IJCNLP*.

Jiale Han, Bo Cheng, and Wei Lu. 2021. Exploring task difficulty for few-shot relation extraction. In *Proceedings of the 2021 Conference on EMNLP*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Kai He, Yucheng Huang, Rui Mao, Tieliang Gong, Chen Li, and Erik Cambria. 2023. Virtual prompt pre-training for prototype-based few-shot relation extraction. *Expert Systems with Applications*, 213:118927.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. *arXiv preprint arXiv:2109.07506*.

Xuefeng Li, Hao Lei, Liwen Wang, Guanting Dong, Jinzheng Zhao, Jiachi Liu, Weiran Xu, and Chunyun Zhang. 2022. A robust contrastive alignment method for multi-domain text classification. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7827–7831.

Xuefeng Li, Liwen Wang, Guanting Dong, Keqing He, Jinzheng Zhao, Hao Lei, Jiachi Liu, and Weiran Xu. 2023. Generative zero-shot prompt learning for cross-domain slot filling with inverse prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 825–834, Toronto, Canada. Association for Computational Linguistics.

Fangchao Liu, Hongyu Lin, Xianpei Han, Boxi Cao, and Le Sun. 2022. Pre-training to match for unified low-shot relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Jiachi Liu, Liwen Wang, Guanting Dong, Xiaoshuai Song, Zechen Wang, Zhengyang Wang, Shanglin Lei, Jinzheng Zhao, Keqing He, Bo Xiao, et al. 2023. Towards robust and generalizable training: An empirical study of noisy slot filling for input perturbations. *arXiv preprint arXiv:2310.03518*.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A simple neural attentive meta-learner. In *International Conference on Learning Representations*.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020a. Learning from context or names? an empirical study on neural relation extraction. *arXiv preprint arXiv:2010.01923*.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020b. An empirical study of multi-task learning on BERT for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on EMNLP-IJCNLP*, pages 3982–3992.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Livio Baldini Soares, Nicholas Fitzgerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 2895–2905.

Mingyang Sun, Qixiang Gao, Yutao Mou, Guanting Dong, Ruifang Liu, and Wenbin Guo. 2023. Improving few-shot performance of dst model through multitask to better serve language-impaired people. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Van-Hien Tran, Van-Thuy Phi, Hiroyuki Shindo, and Yuji Matsumoto. 2019. Relation classification using segment-level attention-based CNN and dependency-based RNN. 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 2793–2798, Minneapolis, Minnesota. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Di Wu, Yiren Chen, Liang Ding, and Dacheng Tao. 2021. Bridging the gap between clean data training and real-world inference for spoken language understanding. *arXiv preprint arXiv:2104.06393*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.

Zhongju Yuan, Zhenkun Wang, and Genghui Li. 2023. Cross-domain few-shot relation extraction via representation learning and domain adaptation. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Peiyuan Zhang and Wei Lu. 2022. Better few-shot relation extraction with label prompt dropout. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL*. Association for Computational Linguistics.

A Training and Inference

After the pre-training stage, we initialize the BERT encoder with pre-trained parameters and fed the entire input instance X_{all} into the encoder:

$$h_T = [Encoder(X_{all})_{e1}, Encoder(X_{all})_{e2}] \quad (5)$$

where h_T stands for the representation of the instance formed by concatenating the representations of the entity markers [E1] and [E2].

We calculate the class prototype $r \in \mathbb{R}^{N_C \times H}$ by averaging the relation representations of the N support instances of each class, where N_C indicates the number of classes. H indicates the input hidden dimension.

During the training stage, we adopt a Cross-Entropy Loss function as follows:

$$\mathcal{L} = -\sum_{k=1}^{N_C} \log \frac{\exp(r_k^\top h_q)}{\sum_{k'=1}^{N_C} \exp(r_{k'}^\top h_q)} \quad (6)$$

where r_k donates the class prototype of class k. h_q is the representation of the query instance. During inference, we choose the relation y as the prediction by finding the closest class prototype to the query sentence's relation representation:

$$\hat{y} = \underset{k=0 \ldots N_C}{arg\,max}\, r_k^\top h_q \quad (7)$$

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Proto-Bert	89.13	94.38	82.77	90.05
BERT-PAIR	88.32	93.22	80.63	87.02
HCRP	96.42	97.96	93.97	96.46
CP	95.10	97.10	91.20	94.70
LPD	98.17±0.0	98.29±0.2	96.66±0.0	96.75±0.2
ClearRE	**98.33**±0.4	**98.57**±0.4	**97.35**±0.1	**97.10**±0.3

Table 5: Accuracy (%) of in-domain few-shot classification on the FewRel1.0 test set.

B Implementation Details

During the experiment, we used bert-base as our backbone to keep it consistent with the baseline we compared against for the sake of fairness. For generative models like BART, considering its probabilistic generation approach, our contrastive learning method cannot be applicable. We implemented our model with PyTorch 1.8.1. We use the Adam optimizer and set the learning rate to 3e-5 and 2e-5 for pre-training and training, respectively. We set batch size to 1024 and 4 for pre-training and training, respectively. We used the same dataset for both pre-training and training without introducing external knowledge. The reason we used the same dataset for pre-training as in the training phase is to demonstrate that ClearRE can achieve good results even without additional datasets. This meets the requirements of low-resource scenarios and serves as another advantage over the baseline methods. In the pre-training stage, we set α as 0.6. During inference, we randomly sample 10,000 episodes from the N-way-K-shot support set and a query instance to evaluate our model. Following previous works (Han et al., 2018; Gao et al., 2019), we set N to 5 and 10, and K to 1 and 5. All experiments are repeated five times with different random seeds for both training and testing on 3090Ti GPU.

C In-domain experiments

To further demonstrate the effectiveness of our method, we conducted additional experiments on in-domain tasks. We conducted experiments on Fewrel 1.0 datasets, and the results are shown in Table 5. The results demonstrate that our method also achieves competitive performance in in-domain tasks, indicating that our approach has a universal capability to enhance the ability of the model to make accurate predictions.

D Detailed ablations experiments

We conduct detailed ablation experiments and show the result in Table 6 to demonstrate the effectiveness of our method. Firstly, we remove two filters separately to test their effectiveness. The absence of any type of filter would have a slight negative impact on the model's performance. For the label prompt, we separately remove the similar relation part and the unsimilar relation part to assess their respective effects. The results indicate that removing similar relations leads to a more pronounced decline in the model's performance, which is attributed to the model lacking cues for easily confused relations during predictions. Finally, we validate the effectiveness of incorporating hard negative samples into the relation contrastive discrimination task. The decrease in results substantiates the effectiveness of our design.

Methods	Fewrel2.0 1-shot	Fewrel2.0 5-shot	CrossRE single	CrossRE multi
ClearRE	**79.78**	**87.76**	**64.20**	**81.71**
w/o Semantic Filter	79.07	87.20	63.47	81.22
w/o Feature Filter	78.97	87.11	63.55	80.96
w/o Label Prompt	77.42	85.37	61.66	80.24
-w/o Similar Relation	77.85	85.77	61.98	80.58
-w/o Unsimilar Relation	78.20	86.01	62.21	80.83
w/o Contrastive Discrimination	75.04	83.58	59.14	77.48
-w/o hard negative sample	75.85	84.05	60.76	79.33

Table 6: The ablation study results (average accuracy %) for Fewrel2.0 and CrossRE.

Number of N_{random}	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
1	84.68	91.60	73.88	83.92
2	84.75	91.53	73.59	83.97
3	84.62	91.58	73.85	83.90
4	84.70	91.55	73.90	83.83

Table 7: Evaluation on the number of N_{random}

Value of k	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
4	84.54	91.46	73.76	83.77
5	84.68	91.60	73.88	83.92
6	84.68	91.71	73.88	83.91
7	84.74	91.55	73.93	83.92
8	84.61	91.67	73.85	83.86

Table 8: Evaluation on the value of k

E hyperparameters selection

For non-similar instances, we introduce these into the prompt to avoid the model over-avoiding similar relations and making radical predictions. We conducted experiments about the quantity of non-similar instances on FewRel 2.0 and the results indicated that increasing the number of non-similar instances did not have a significant impact on the model's performance. Therefore, we retained only one randomly selected non-similar instance. We supplement the experimental results as shown in Table 7.

We have also conducted experiments related to k on FewRel 2.0, and the results indicate that increasing k does not have a significant impact on the model's performance, which is shown in Table 8. If we set it larger, there will be more relations in the intersection. However, for almost all relations, the number of truly similar relations is limited. Increasing the value of k does not provide more useful information for prompts about confusion.