ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448452

ADAPTIVE PROMPT CONSTRUCTION METHOD FOR RELATION EXTRACTION

Zhenbin Chen[1,2], *Zhixin Li*[1,2,*], *Ying Huang*[1,2], *Zhenjun Tang*[1,2]

[1]Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education,
Guangxi Normal University, Guilin 541004, China
[2]Guangxi Key Lab of Multi-source Information Mining and Security,
Guangxi Normal University, Guilin 541004, China

ABSTRACT

Prompt learning was proposed to solve the problem of inconsistency between the upstream and downstream tasks and has achieved State-Of-The-Art (SOTA) results in various Natural Language Processing (NLP) tasks. However, Relation Extraction (RE) is more complex than other text classification tasks, which makes it more difficult to design a suitable prompt template for each dataset manually. To solve this issue, we propose a **A**daptive **P**rompt **C**onstruction method (**APC**) for relation extraction. Our method entails obtaining context-aware prompt tokens by extracting and generating trigger words associated with the entities. Furthermore, to alleviate the issue of instability in the prompt-tuning framework during training, we introduce a novel joint contrastive loss to optimize our model. Our method not only effectively reduces the human effort used for prompt template construction, but also achieves better performance in RE. We conduct the experiment on four public RE datasets, which demonstrate the proposed method outperforms the existing SOTA results in both datasets and experimental settings.

Index Terms— Relation Extraction, Pretrained Language Model, Prompt Learning, Contrastive Learning

1. INTRODUCTION

Relation extraction (RE) is a significant task that supports various NLP tasks. With the advancements made in Pretrained Language Models (PLMs), the majority of relation extraction methods are currently based on the fine-tuning paradigm [1, 2]. However, within the fine-tuning paradigm, the gap between downstream task objectives and pre-training task objectives severely constrains the performance of the tasks.

Recently, prompt-tuning was proposed to address these issues [3, 4]. The core idea of prompt-tuning is to reformulate the task to a cloze-style task and use PLMs as a predictor, which can bridge the gap between the pre-training and fine-tuning in the training objective. Studies on prompt-based RE methods have gained significant attention and demonstrated superiority over traditional fine-tuning methods [5, 3], such as PTR[5], which introduced sub-prompt to construct prompt templates and designed logic rules for extracting relations. However, the manual prompt templates used for prompt-based RE require manual design, and finding the optimal one is difficult. Therefore, some studies have proposed automatic constructed prompt templates, such as AutoPrompt[6]. But Autoprompt requires a lot of computation to retrieve the useful prompt tokens, which makes it hard to generalize. Consequently, a more efficient approach to automated prompt generation becomes imperative. Above all, despite the promising results obtained from the use of prompt-based RE methods, there are still some challenges that need to be addressed. Firstly, the designing of an appropriate prompt template for each dataset in different domains is a time-consuming task and we need a more automated approach for it. Secondly, the context related to the entity pair and relation label of the sentence should not be overlooked when constructing prompts as they play an important role in RE. For example, as shown in Fig 1, the terms *"president"* and *"city"* has indicate the entities' type and limit the scope of the

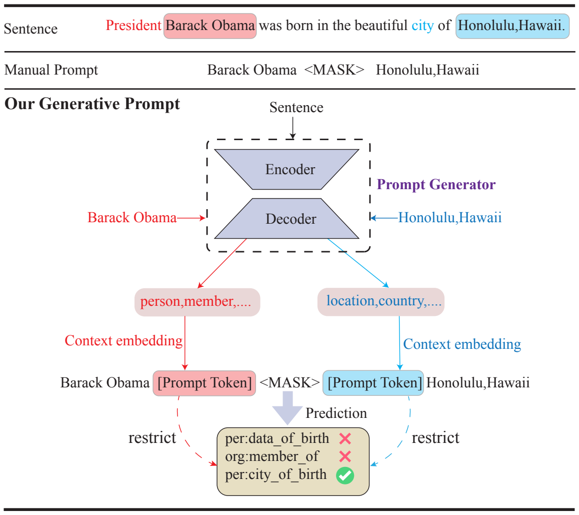

Fig. 1: The core idea of APC.

[*] Zhixin Li is the corresponding author (lizx@gxnu.edu.cn).

relation when we extract the relation between *Barack Obama* and *Honolulu, Hawaii* in the sentence *"The president Barack Obama was born in the beautiful city Honolulu, Hawaii."*. Pre-trained language models (PLMs) serve as classifiers in prompt-tuning, and researching how to make them aware of entities and contextual information is crucial for utilizing prompt-tuning for RE.

In this paper, we propose a novel approach, the **A**daptive **P**rompt **C**onstruction method (**APC**), for generating context-aware prompt tokens for RE. Our model uses contextual information to generate prompts for each instance, which effectively reduces the requirement for human effort. The contributions of the approach proposed in this paper can be summarized as follows:

- We propose to use the prompt-tuning for relation extraction, aims at bridging the gap between pre-trained language models and downstream classifiers.

- We design a context-aware prompt generator that can generate the effective prompt tokens based on the context, thereby enhancing the performance of relation extraction and alleviating the requirement for manual effort.

- We design an in-domain pre-training strategy and a joint contrastive loss function to enhance the model's domain adaptability and robustness.

Experimental results on four public RE datasets has demonstrated APC outperforms the existing SOTA result in both supervised and few-shot settings.

2. PROPOSED METHOD

The overall framework of our proposed model is depicted in Fig. 2.

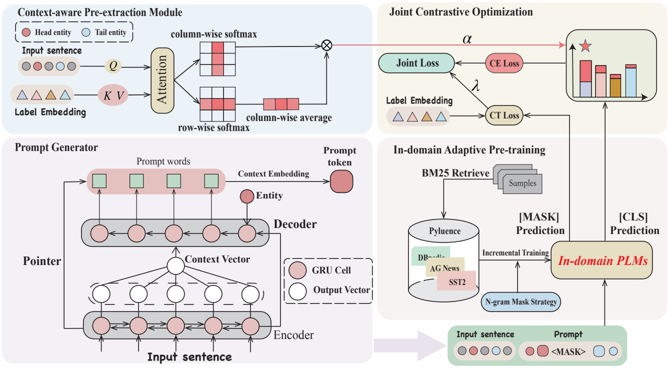

Fig. 2: The model architecture of APC. CE denotes the Cross-Entropy Loss and CT Loss denotes the Contractive Loss designed for APC.

2.1. Context-aware Prompt Generator

The context-aware prompt generator we proposed consists of two components: a context-aware pre-extractor and a prompt generator. Relying upon the text generation prowess of PLMs, prompt-based relation extraction is fraught with instability. Consequently, we design a context-aware pre-extractor to avoid significant deviations in the model's prediction.

The design of the pre-extractor is primarily inspired by Attention Over Attention method [7], as illustrated in the blue region of Fig.2. We employ scale-dot attention to derive the attention matrix M between the text sequence $\mathcal{X} = \{w_1, w_2, \ldots, w_{n-1}, w_n\}$ and the set of relation labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_{m-1}, y_m\}$ as shown in Equation 1.

$$M(i,j) = {w_i}^T \cdot y_j,\ w_i \in \mathcal{X},\ y_j \in \mathcal{Y} \tag{1}$$

Then, utilizing column-wise softmax and row-wise softmax operations on M, we can obtain the context-to-label relevance matrix $M_{c2l} \in \mathbb{R}^{|x|\times|\mathcal{Y}|}$ and the label-to-context relevance matrix $M_{l2c} \in \mathbb{R}^{|x|\times|\mathcal{Y}|}$. By performing column-wise averaging on M_{l2c} and then matrix multiplication of the resulet with M_{c2l}, the result is subsequently normalized to obtain the pre-extracted probability distribution $\alpha \in \mathbb{R}^{|\mathcal{X}|}$.

In order to efficiently attain valuable prompt templates, we propose a pre-trained prompt generator employing an encoder-decoder architecture. The application of the encoder-decoder framework is pervasive in text summarization tasks[8] due to its dual capacity for text extraction and generation. Thus, we can extract or generate the valuable contextual words from the input via prompt generator. And we term these words as "triggers". These triggers are subsequently embedded into the prompt tokens. We employ a single-layer bidirectional GRU as the encoder. The given textual embeddings, denoted as $\mathcal{X} = \{w_1, w_2, \ldots, w_{n-1}, w_n\}$, are sequentially input into the encoder one by one, yielding the encoder's output $H = GRU_{enc}(\mathcal{X})$. Subsequently, we utilize the embedding of the head entity T_{sub} and tail entity T_{obj} with Equation 2 and 3 as the start token, and input them into the decoder with $H = \{h_1, h_2, \cdots, h_i, \cdots, h_n\}$, where s and o denote the starting indices of the entities, and i and j denote the lengths of the entities.

$$T_{sub} = \sum_{p=s}^{s+i} \varphi(w_p) \cdot Emb(w_p) \tag{2}$$

$$T_{obj} = \sum_{q=o}^{o+j} \varphi(w_q) \cdot Emb(w_q) \tag{3}$$

We employ a single-layer unidirectional GRU as the decoder. At each step t of the decoding stage, the decoder computes the current time step's attention distribution a^t based on the decoder state s_t and the encoder's output H.

$$e_i^t = v^T \tanh\left(W_h h_i + W_s s_t + b_{\text{attn}}\right) \tag{4}$$

$$a^t = \operatorname{softmax}\left(e^t\right) \tag{5}$$

where v, W_h, W_s and b_{attn} are learnable parameters. The attention distribution is then utilized to compute a weighted sum of the encoder hidden states, known as the context vector $c_t = \sum_i a_i^t h_i$. Subsequently, the decoder's output at the t th time step can be computed by the following equation:

$$P_{vocab}(G_*^t) = \operatorname{softmax}\left(V'\left(V\left[s_t, c_t\right] + b\right) + b'\right) \tag{6}$$

where V, V' , b and b' are learnable parameters. $P_{vocab}(G_*^t)$ denotes the probability distribution over the vocabulary, and the word G_*^t predicted at step t has the highest probability value within the distribution, where $* \in \{sub, obj\}$ corresponds to different starting tokens (T_{sub} or T_{obj}).

Furthermore, we introduce the pointer mechanism[9] to promise the valuable contextual information such as the words "*president*" and "*city*" in Fig 1 will not be discarded during generation stage. The generation probability $p_{gen} \in [0, 1]$ can be obtained by:

$$p_{\text{gen}} = \sigma\left(w_c^T c_t + w_s^T s_t + w_x^T x_t + b_{\text{pt}}\right) \tag{7}$$

$$P'_{\text{vocab}}(G_*^{'t}) = p_{\text{gen}} P_{\text{vocab}}(G_*^t) + (1 - p_{\text{gen}}) \sum_{i:w_i = G_*^t} a^t \tag{8}$$

where σ is the sigmoid function. Note that if G_*^t is a special token (such as "[UNK]","[PAD]", etc.), then the $P_{vocab}(G_*^t)$ will be set as zero. If G_*^t does not appear in the input, then $\sum_{i:w_i=w'} a^t$ will be set as zero. And $G_*^{'t}$ are the generated trigger words.

After obtaining all the triggers, we embed them into prompt tokens $PT_* = \sum_{i=1}^{u} \varphi(G_*^i) \cdot Emb(G_*^i)$. And the prompt tokens will be concatenated with the input and fed into the PLMs for RE.

2.2. In-domain Adaptive Pre-trainning

The existing work [10] has demonstrated that the performance of prompt-tuning decreases when the pre-trained corpus and the training datasets belong to different domains. In this paper, we propose an in-domain adaptive pre-training strategy to inject the domain-relevance knowledge to the PLMs. We extract a maximum of 100 sample data instances from the training data. Subsequently, we retrieve relevant corpus content from a large-scale corpus (SST2 , DBPedia , AGNews) related to the current dataset and employ this corpus to perform incremental training on the PLMs. Inspired by the search-based pre-training methods [11], we chose BM25 score [12] as the metric for measuring text relevance. Additionally, we employ Pyluence[1] for retrieving and utilize N-gram masking strategy [13, 14] for incremental training.

2.3. Joint Contrastive Loss

To enhance model convergence, we design a Joint Contrastive Loss to optimize our model. We employ the hidden layer vector of the [CLS] token from the last layer of the PLMs to determine the ultimate distribution of relation predictions y_{plm}. This distribution is then combined with the pre-extracted distribution α to obtain the final predictions, denoted as $\hat{y} = \operatorname{softmax}(\alpha + y_{plm})$. Subsequently, we can compute the cross-entropy loss $\mathcal{L}_{CE}$ for RE:

$$\mathcal{L}_{CE} = \text{Cross-Entropy}(y_i, \hat{y}) \tag{9}$$

Furthermore, to ensure the consistency of the model's prediction, we think that the prediction vectors of the [MASK] token y_{pre} should also be closer to the correct relation labels embeddings y^+ and farther away from incorrect relation labels embeddings y^-. Thus, we can compute the contrastive loss $\mathcal{L}_{CT}$.

$$\mathcal{L}_{CT} = \frac{1}{N} \ln \frac{\exp(s(y_{pre}, y^+)/\tau)}{\exp(s(y_{pre}, y^+)/\tau) + \sum \exp(s(y_{pre}, y^-)/\tau)} \tag{10}$$

Finally, we use the Kullback-Leibler (KL) divergence to compute a weighted sum of the loss functions, resulting in the ultimate Joint Contrastive Loss function:

$$\mathcal{L} = \mathcal{L}_{CE} + (1 + \text{KL}(\alpha, y_{plm})) \cdot \mathcal{L}_{CT} \tag{11}$$

3. EXPERIMENTS

3.1. Datasets and Evaluation Metric

We conduct experiments on the following four public RE datasets to verify the effectiveness of our model: SemEval 2010 Task 8 (SemEval) [15], TACRED [16], TACRED-Revisit [17], Re-TACRED[18] and adopted the F1 score as the evaluation metric.

3.2. Comparison With State-of-the-Art Methods

3.2.1. Supervised setting

As we can see in Table 1, our APC model exhibits a promising improvement in supervised setting. SpanBERT[19], KnowBERT[20], LUKE[21], MTB[22] and GDPNet[23] are the methods both based on the fine-tuning paradigm. Despite they have well-designed downstream classifiers and knowledge injection, their performance lags behind the methods based on the prompt-tuning paradigm such as PTR[5] due to the existing gap between the objectives of the upstream and downstream tasks. Downstream relation classifier can not benefit significantly from pre-training, and fine-tuning relies on the scale and quality of annotated data, which severely limits methods based on the fine-tuning paradigm. KnowPrompt[4] is a strong competitor, but compare to it, we directly use contextual information as prior knowledge instead of just considering prior probabilities of entities. This prevents the establishment of shortcuts between relation categories and entities, thereby enabling us to achieve better performance.

[1] https://lucene.apache.org/pylucene/

Table 1: The F1 scores(%) of RE on different datasets in supervised setting. The best results are **bold**.

Model	Extra Data	SemEval	TACRED	TACREV	ReTACRED
Fine-tuning-[Roberta]	w/o	87.6	68.7	76.0	84.9
R-BERT[24]	w/o	89.3	69.4	-	-
SpanBERT[19]	w/	-	70.8	78.0	85.3
KnowBERT[20]	w/	89.1	71.5	79.3	89.1
LUKE[21]	w/	-	72.7	80.6	-
MTB[22]	w/	89.5	70.1	-	-
GDPNet[23]	w/o	-	71.5	79.3	-
RE-DPM[25]	w/o	89.9	71.5	79.3	-
SPOT[26]	w/o	89.4	-	-	89.4
RELA[27]	w/o	89.6	71.2	79.7	-
PTR-[Roberta][5]	w/o	89.9	72.4	81.4	90.9
KnowPrompt[4]	w/o	90.1	72.4	81.7	91.1
APC(ours)	**w/**	**90.3**	**72.7**	**82.7**	**91.4**

3.2.2. *Few-shot setting*

According to the result shown in Table 2, APC retains its advantages in few-shot setting. Moreover, in few-shot setting, methods like AdaPrompt[10] and PTR[5], based on the prompt-tuning paradigm, exhibits more notable improvements than those based on the fine-tuning paradigms. Specifically, our approach has led to an average increase of 16.6%, 13.2%, 11.1%, and 15.3% in F1 scores across the four datasets compared to vanilla fine-tuning methods, which indicates that our method exhibits a more pronounced advantage in few-shot settings.

Table 2: The F1 scores(%) of RE on different datasets in few-shot setting. The best results are **bold**.

Dataset	Method	K=8	K=16	K=32	Mean
Semeval	Fine-tuning	41.3	65.2	80.1	62.2
	GDPNet[23]	42	67.5	81.2	63.6
	AdaPrompt[10]	-	-	-	-
	PTR[5]	**70.5**	**81.3**	84.2	78.4
	APC(ours)	70.2	80.9	**85.3**	**78.8**
TACRED	Fine-tuning	12.2	21.5	28	20.6
	GDPNet[23]	11.8	22.5	28.8	21.1
	AdaPrompt[10]	-	-	-	-
	PTR[5]	28.1	30.7	32.1	30.3
	APC(ours)	**31.9**	**34.5**	**35.0**	**33.8**
TACREV	Fine-tuning	13.5	22.3	28.2	21.4
	GDPNet[23]	12.3	23.8	29.1	21.8
	AdaPrompt[10]	25.2	27.3	30.8	27.8
	PTR[5]	28.7	31.4	32.4	30.8
	APC(ours)	**30.7**	**34.1**	**35.7**	**33.5**
Re-TACRED	Fine-tuning	28.5	49.5	56	44.7
	GDPNet[23]	29.0	50.0	56.5	45.2
	AdaPrompt[10]	-	-	-	-
	PTR[5]	51.5	56.2	62.1	56.6
	APC(ours)	**53.9**	**60.6**	**65.5**	**60.0**

3.3. Ablation Study

To evaluate our proposed modules, we carry out ablation study and report the experimental results on SemEval and TACREV in Table 3, which demonstrates the effective of the proposed modules. The first row of Table 3 showes the performance of the baseline methods, which only uses the [MASK] token as the prompt template for RE. As shown in Table 3, APC is 1.6% and 1.2% higher than the baseline and the context-aware prompt tokens bring the most significant improvement. It is worth noting that domain-specific pre-training did not lead to a significant improvement on the SemEval dataset. The main reasons for this are (1) the challenging nature of relation type classification in the SemEval dataset, such as ”Cause-Effect(e1, e2),” and (2) the lack of text containing relevant domain knowledge in the general corpora. However, it still had a certain positive impact on the model’s performance.

Table 3: The F1 scores(%) of the ablation study in supervised setting. Legend : **CPG**: Context-aware Prompt Genetor; **IAP**: In-domain Adaptive Pretraining; **JCL**: Joint Contrastive Loss.

Module			Datasets	
CPG	IAP	JCL	SemEval	TACREV
			88.7	81.5
√			89.7 (+1.0)	82.3 (+0.8)
	√		88.7 (+0.0)	81.7(+0.2)
		√	89.1 (+0.4)	81.9(+0.4)
√	√		89.9 (+1.2)	82.5 (+1.0)
√		√	90.1 (+1.4)	82.5(+1.0)
	√	√	89.5 (+0.8)	82.1 (+0.6)
√	√	√	**90.3** (+1.6)	**82.7** (+1.2)

4. CONCLUSIONS

To overcome the limitations of traditional RE methods based on fine-tuning paradigm, we propose a novel prompt-tuning based RE method (APC). Our proposed APC has a prompt generator which generates more effective prompt tokens from entity and context information. These context-aware prompt tokens can better provide the priori knowledge for pre-trained language models, improving the relation extraction capability. Furthermore, we propose the in-domain pre-training strategy and joint comparison loss functions that can further improve the performance of APC. According to the experimental results on four public datasets, our method has more advantages in both supervised and few-shot setting. In future work, we will introduce the external knowledge to give PLMs stronger knowledge prior to further improve the model inference.

5. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (Nos. 62276073, 61966004, 62272111), Guangxi Natural Science Foundation (No. 2019GXNSFDA2 45018), Innovation Project of Guangxi Graduate Education (No. YCBZ2023055), Guangxi “Bagui Scholar” Teams for Innovation and Research Project.

6. REFERENCES

[1] C. Alt, M. Hübner, and L. Hennig, "Fine-tuning pretrained transformer language models to distantly supervised relation extraction," in *ACL*, 2019, pp. 1388–1398.

[2] Z. Li, Y. Sun, J. Zhu, et al., "Improve relation extraction with dual attention-guided graph convolutional networks," *Neural Computing and Applications*, vol. 33, pp. 1773–1784, 2021.

[3] Y. Chia, L. Bing, S. Poria, et al., "Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction," in *ACL*, 2022, pp. 45–57.

[4] X. Chen, N. Zhang, X. Xie, et al., "Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction," in *WWW*, 2022, pp. 2778–2788.

[5] X. Han, W. Zhao, N. Ding, et al., "Ptr: Prompt tuning with rules for text classification," *arXiv preprint arXiv:2105.11259*, 2021.

[6] T. Shin, Y. Razeghi, R. L. Logan IV, et al., "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," in *EMNLP*, 2020, pp. 4222–4235.

[7] Y. Cui, Z. Chen, S. Wei, et al., "Attention-over-attention neural networks for reading comprehension," in *ACL*, 2017, pp. 593–602.

[8] Z. Li, Z. Peng, S. Tang, et al., "Text summarization method based on double attention pointer network," *IEEE Access*, vol. 8, pp. 11279–11288, 2020.

[9] A. See, P. Liu, and C. Manning, "Get to the point: Summarization with pointer-generator networks," in *ACL*, 2017, pp. 1073–1083.

[10] Y. Chen, Y. Liu, L. Dong, et al., "Adaprompt: Adaptive model training for prompt-based nlp," *arXiv preprint arXiv:2202.04824*, 2022.

[11] S. Wang, Y. Xu, Y. Fang, et al., "Training data is more valuable than you think: A simple and effective method by retrieving from training data," in *ACL*, 2022, pp. 3170–3179.

[12] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[13] Y. Zeng, Z. Li, Z. Tang, et al., "Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis," *Expert Systems with Applications*, vol. 213, pp. 119240, 2023.

[14] Y. Zeng, Z. Li, Z. Chen, et al., "Aspect-level sentiment analysis based on semantic heterogeneous graph convolutional network," *Frontiers of Computer Science*, vol. 17, no. 6, pp. 176340, 2023.

[15] I. Hendrickx, S. N. Kim, Z. Kozareva, et al., "Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," *arXiv preprint arXiv:1911.10422*, 2019.

[16] Y. Zhang, V. Zhong, D. Chen, et al., "Position-aware attention and supervised data improve slot filling," in *EMNLP*, 2017, pp. 35–45.

[17] C. Alt, A. Gabryszak, and L. Hennig, "Tacred revisited: A thorough evaluation of the tacred relation extraction task," in *ACL*, 2020, pp. 1558–1569.

[18] G. Stoica, E. A. Platanios, and B. Póczos, "Re-tacred: Addressing shortcomings of the tacred dataset," in *AAAI*, 2021, vol. 35, pp. 13843–13850.

[19] M. Joshi, D. Chen, Y. Liu, et al., "Spanbert: Improving pre-training by representing and predicting spans," *TACL*, vol. 8, pp. 64–77, 2020.

[20] M.E. Peters, M. Neumann, R.L. Logan IV, et al., "Knowledge enhanced contextual word representations," in *EMNLP*, 2019, pp. 43–54.

[21] I. Yamada, A. Asai, H. Shindo, et al., "LUKE: Deep contextualized entity representations with entity-aware self-attention," in *EMNLP*, 2020, pp. 6442–6454.

[22] L. B. Soares, N. FitzGerald, J. Ling, et al., "Matching the blanks: Distributional similarity for relation learning," in *ACL*, 2019, pp. 2895–2905.

[23] F. Xue, A. Sun, H. Zhang, et al., "Gdpnet: Refining latent multi-view graph for relation extraction," in *AAAI*, 2021, vol. 35, pp. 14194–14202.

[24] S. Wu and Y. He, "Enriching pre-trained language model with entity information for relation classification," in *CIKM*, 2019, pp. 2361–2364.

[25] Y. Tian, Y. Song, and F. Xia, "Improving relation extraction through syntax-induced pre-training with dependency masking," in *ACL*, 2022, pp. 1875–1886.

[26] J. Li, Y. Katsis, T. Baldwin, et al., "Spot: Knowledge-enhanced language representations for information extraction," in *CIKM*, 2022, pp. 1124–1134.

[27] B. Li, D. Yu, W. Ye, et al., "Sequence generation with label augmentation for relation extraction," in *AAAI*, 2023, vol. 37, pp. 13043–13050.