LabelPrompt: Effective prompt-based learning for relation classification

Wenjie Zhang WENJIE.ZHANG@STU.JIANGNAN.EDU.CN
Xiaoning Song X.SONG@JIANGNAN.EDU.CN
Zhenhua Feng FENGZHENHUA@JIANGNAN.EDU.CN
Tianyang Xu TIANYANG_XU@163.COM
Xiaojun Wu WU_XIAOJUN@JIANGNAN.EDU.CN
School of Artificial Intelligence and Computer Science, Jiangnan University

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Recently, prompt-based learning has become popular in many Natural Language Processing (NLP) tasks by converting the task into a cloze-style one to smooth out the differences between Pre-trained Language Models (PLMs) and the current task. However, as for relation classification, it is challenging to associate the natural language word that fills in the mask token with relation labels due to the rich semantic information in textual label, *e.g.* "*org:founded_by*". To address this challenge, this paper presents a novel prompt-based learning method, namely LabelPrompt, for the relation classification task. It is an extraordinary intuitive approach motivated by "GIVE MODEL CHOICES!". Specifically, we first define additional tokens to represent the relation labels, which regard these tokens as the verbalizer with semantic initialisation and explicitly construct them with a prompt template method. Then, we address the inconsistency between predicted relations and given entities by implementing an entity-aware module that employs contrastive learning. Last, we conduct an attention query strategy to differentiate prompt tokens and sequence tokens. These strategies effectively improve the adaptation capability of prompt-based learning, especially when only a small labelled dataset is available. Extensive experimental results obtained on several bench-marking datasets demonstrate the superiority of our method, particularly in the few-shot scenario. Our code can be found at https://github.com/xerrors/Labelprompt.

Keywords: relation classification, prompt learning, few-shot.

1. Introduction

Pretrained Language Models (PLMs) have excelled in various Natural Language Processing (NLP) tasks (Xu et al., 2021; Zhong and Chen, 2021), such as Relation Classification (RC), Named Entity Recognition (NER), etc. The strong performance of PLMs stems from their ability to learn rich knowledge representations from large unlabelled corpora through self-supervised pretraining objectives such as Masked Language Modelling (MLM). However, a key challenge is the mismatch between the PLMs' pretraining objective and downstream tasks, resulting in sub-optimal knowledge transfer (Liu et al., 2021).

To address this, prompt-based learning reformulates tasks into the MLM format used during pretraining via prompt templates and verbalizers (Lester et al., 2021). This reduces the objective mismatch and enables effective PLM use. However, manually engineering

prompts is expensive (Liu et al., 2021). Recent work explores automatically generating prompts (Lester et al., 2021; Liu et al., 2022b) or learning continuous representations (Li and Liang, 2021) to reduce engineering costs. Overall, prompt-based learning is promising but current limitations warrant further research.

Recent studies have applied prompts to text classification like relation classification (Meng et al., 2023). As shown in Table 1, where given entity pairs, the goal is to predict their relation (*e.g.*, "*per:employee_of*"). While prompt-based learning provides an appealing approach to using PLMs for RC, applying prompts here also poses challenges.

First, a key challenge in applying prompts to relation classification is that the target semantic relation labels contain rich, multi-word information not directly representable within the pretrained vocabulary. Existing verbalizer techniques struggle to effectively map the model's masked outputs to these complex labels when reformulating the task as masked language modelling. Some approaches address this by searching external data for label-related words or defining new learnable tokens mapped to labels (Chen et al., 2022). However, these methods have yet to achieve optimal results or require more labelled data to tune the new tokens. So the first challenge is developing training methodologies that can enhance prompt-based model performance for relation classification under limited labelled data availability.

Second, another core challenge is that relation classification requires modelling the correspondence between a given entity pair and their relation, beyond just identifying relations at the sentence level. Existing prompt approaches predominantly predict relations without explicitly capturing entity-relation triplets (s, r, o) (Stoica et al., 2021). For example, a model may incorrectly predict the relation between the wrong entity pair in the sentence, failing to discern the given entities when inferring relations. Therefore, an additional challenge is enhancing the model's capacity to accurately capture entity-relation correspondence in the prompt-based framework for relation classification.

To mitigate the above limitations of prompt-tuning for relation classification, we develop a novel LabelPrompt approach with an entity-aware module. This approach is a simple yet effective method that bridges the gap between pre-training objectives and the relation classification task by enhancing the model's intuitive understanding of relation labels.

In terms of the studies in prompt templates, some studies have shown that it is essential to provide the PLMs with supplementary task-relevant information via the input. Building on this, we propose an intuitive technique that offers the model selectable options. Specifically, we develop a prompt template approach that embeds relation class labels into the input example before the model is ingested. We enlarging the vocabulary space with custom label tokens that are absent in natural language (*e.g.*, "*[C1]*"). These tokens' embeddings carry semantic information derived from relation label texts. These tokens are also be constructed into prompt templates on the input side in a more intuitive way to improve prompt-based method performance. Furthermore, to insulate additional prompt tokens from impacting sentence semantics, we redesign the attention query strategy that use distinct query projections for prompt and sentence token pairs. This manages prompt-sentence interactions, and allows label prompt tokens to provide task-specific information without compromising sentence semantics encoded in the PLMs.

Then, to enhance entity perception during model training, we implement an entity-aware module to assess the correlation between entities and their relations. As shown

in Figure 1, we adopt a contrastive learning-inspired approach to construct positive and negative samples, where the former consists of given entities and their predicted relation, while the latter contains randomly selected tokens and ground-truth relations within the sentence. This effectively constrains the relation and entities with semantic information.

Finally, extensive evaluations on popular benchmark datasets demonstrate that our proposed approach yields significant performance improvements compared to fine-tuning and other prompt-based methods. We attribute this enhancement to our method's ability to effectively constrict the model's search space and guide optimisation towards the correct direction.

In summary, the main contributions of the proposed LabelPrompt method include:

- We introduce a prompt-based method that explicitly exploits relation features and significantly improves the performance of prompt-based learning in few-shot scenarios.
- We implement an entity-aware module that employs contrastive learning to enhance entity awareness during the inference stage.
- We design an attention query strategy within self-attention layers to differentiate between prompt and sentence tokens and improve the performance of prompt-based learning.

2. Related Work

2.1. Pre-training and Fine-tuning

Pre-trained language models are trained in large-scale unlabelled text data to learn robust general-purpose features, such as lexical, syntactic, semantic, and word representation (Liu et al., 2022b). Many large language models, such as GPT (Ouyang et al., 2022), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), have been proposed, greatly promoting the development of the research in natural language processing. By fine-tuning the parameters of PLMs with additional specific modules and task-specific objective functions, the models can be adapted to most downstream tasks. Generally, the pre-training and fine-tuning paradigm become the foundation for many NLP tasks, such as named entity recognition (Wu et al., 2021; Jehangir et al., 2023), relation classification (Yamada et al., 2020; Wadhwa et al., 2023), and question answering (Abdel-Nabi et al., 2023), etc.

2.2. Prompt-based Learning

Although fine-tuning has achieved significant success in many fields, there remains a large gap between the objectives of pre-training and fine-tuning tasks (Liu et al., 2021). Typically, a PLM is trained with Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) tasks (Devlin et al., 2019), which differs from downstream NLP tasks like classification.

Prompt-based learning reformulates a downstream task to the original PLM training task. This method enhances PLM performance by providing additional information in cloze-style (Lester et al., 2021). Many studies have shown that prompt templates containing semantics can effectively activate knowledge in the model. Petroni et al. (2019) treated

Table 1: Two examples of relation classification.

Sentence	Relation
Mark Fisher writes for the **Dayton Daily News**.	per:employee_of
He has a sense of **humor** about his **reaction** to that day.	no_relation

language models as knowledge bases, introducing the LAMA (LAnguage Model Analysis) probe dataset. This dataset uses hand-crafted query sentences to extract knowledge from PLMs, performing comparably to direct PLM use. However, manual template design is time-consuming, costly (Shin et al., 2021), and often requires expert knowledge. Further, Shin et al. (2020) introduced automatic prompt generation based on gradient search. Some researchers have replaced manual templates with learnable task-specific prefixes as continuous prompts (Li and Liang, 2021; Lester et al., 2021; Wang et al., 2023).

Answer mapping is a crucial aspect of prompt-based learning. Since the masked output may not align with the textual label (Zhou et al., 2023). A verbaliser, also known as a projection method, is needed to align model output with the answer space. Just like prompt templates, verbalisers can be designed manually or through discrete/continuous search methods. The quality of a verbaliser significantly impacts model performance.

In conclusion, prompt-based learning is widely considered a highly effective method for utilising pre-trained language models, and it is a subject worthy of further study.

2.3. Relation Classification

Relation Classification (RC) is a sub-task of Information Extraction (IE) that aims to identify the semantic relation between two entities in a given text. As shown in Table 1, RC is a crucial task for many natural language understanding applications, including question answering, knowledge base construction, and text summarization (Hendrickx et al., 2010). In typical methods, PLMs, such as BERT, has provided universal language representations that are useful for RC tasks (Yamada et al., 2020; Liu et al., 2022a). However, these methods often require significant amounts of labelled data and complex, task-specific neural modules, and tend to underperform in the few-shot scenario.

Recently, there have been studies exploring the use of prompt-based learning to leverage the knowledge contained in PLMs for RC (Feng et al., 2024). The LAMA dataset is a probe for analysing factual and commonsense knowledge in PLMs, each comprising a set of facts. Han et al. (2021) applied logic rules to construct prompts with several sub-prompts, effectively using prior knowledge from relation classification. Chen et al. (2022) proposed a prompt-based learning approach called KnowPrompt that incorporates abundant semantics and prior knowledge in relation labels into relation classification. FPC (Yang and Song, 2022) attempts to introduce an auxiliary prompt-based fine-tuning task into the classification model, to enable the model to grasp the semantics of relation labels.

However, none of these methods addressed the issue of inconsistency between prompt output and relation label texts. Inspired by KnowPrompt, we propose a new prompt-based learning approach, namely LabelPrompt, that explicitly constructs prompt templates using label tokens with supervised information corresponding to the labels at the model's output.

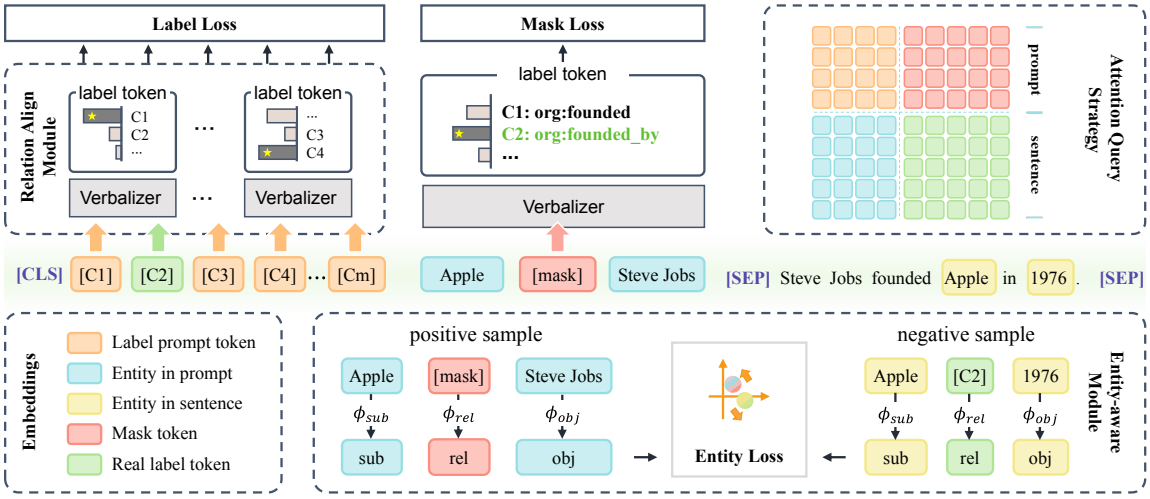

Figure 1: An overview of the proposed method. We input the tokens, including the label tokens, into the RoBERTa for global feature representation and contextual awareness. The masked token used to predict the relation. The relation-align module aligns the encoded label prompt tokens with their corresponding classes. The Entity-Aware Module constructs samples using triplet info and pure prompt template tokens to constrain entity correspondence. The attention query strategy facilitates independent prompt and sentence token streams.

3. The Proposed LabelPrompt Method

Relation Classification (RC) can be denoted as $\mathcal{T} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X}$ is the set of instances and $\mathcal{Y}$ is the set of relation labels. For each instance $\mathbf{x} \in \mathcal{X}$, it contains a sentence $x = \{x_1, x_2, \ldots x_n\}$, two entities (e_s, e_o) mentioned in x , and its label relation $y_x \in \mathcal{Y}$. The goal of relation classification is to predict the relation $y \in \mathcal{Y}$ between given subject entity e_s and object entity e_o in sentence x. If there is no relation between the entities, it will be considered as a special type of relation $\epsilon \in \mathcal{Y}$, *i.e.* "*no_relation*", which exists in the set of relations.

We will elaborate on the major components and training schemes of our approach in the following parts. As illustrated in Figure 1, the input to the model is composed of two primary components, the prompt tokens and the sentence tokens. We can further separate the prompt tokens into label prompt tokens and typical prompt tokens.

3.1. Label Prompt

3.1.1. Label Prompt Tokens

For each relation label $y \in \mathcal{Y}$, the label text is a composite phrase rather than a single word, *e.g.* "*org:top_members/employees*" and etc. We define a label space by creating m label tokens $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$ that extend the PLM's vocabulary, where m is the number of relations $|\mathcal{Y}|$. For better training initialisation, these tokens are initialised with their

semantic label texts as:

$$C_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \text{Embedding}_{\mathcal{M}}(t_j), \tag{1}$$

where C_i is the embedding for label token c_i, $\text{Embedding}_{\mathcal{M}}$ stands for the word embedding layer of $\mathcal{M}$, and t_j is denoted as the j-th sub-text of the decomposed label text $T_i = \{t_1, t_2, \ldots, t_{N_i}\}$ for relation i.

3.1.2. Label Prompt Template

The core idea of prompt-based learning is to convert the task to MLM using prompt templates (additional text with prompt tokens to construct the input $\mathbf{x}_{prompt}$ with a specific token `[MASK]`. However, previous work (Han et al., 2021; Chen et al., 2022) might fall short in terms of efficiency, because the newly added label tokens were not present during the pre-training phase of the model.

To overcome this, we introduce an intuitive approach of directly incorporating label tokens into the input sequences through our proposed Label Prompt Template. This template contains both label prompt tokens and typical prompt tokens. We represent the template as $T(\mathbf{x})$ and the input x will be converted to the input sequence as follows:

$$T(\mathbf{x}) = \{c_1, c_2, \ldots, c_m, [\text{SEP}], e_s, [\text{MASK}], e_o\}, \tag{2}$$

$$x_{prompt} = \{[\text{CLS}], T(\mathbf{x}), [\text{SEP}], x, [\text{SEP}]\}, \tag{3}$$

where $c_i \in \mathcal{C}$ is the label token, e_s and e_o are the subject and object entities. `[CLS]`, `[SEP]` and `[MASK]` are predefined in $\mathcal{M}$.

3.1.3. Label Prompt Learning

In this section, we introduce our key training objective. First, let $\mathbf{h} = \text{Encoder}_{\mathcal{M}}(x_{prompt})$ be the encoder output for input x_{prompt}. Define h_{mask}, h_{c_i}, h_{sub}, and h_{obj} as the vectors for `[MASK]`, label token c_i, subject entity, and object entity, respectively.

Next, we define a verbaliser V_ϕ as an answer-mapping method that maps a hidden vector h to a relation label:

$$V_\phi(h) = \mathbf{E}_{label} \cdot (W_v \cdot h + b), \tag{4}$$

where $\mathbf{E}$ is the embedding layer, and $\mathbf{E}_{label}$ represents the labels' embedding, W_v and b are learnable parameters. Then the probability that the `[MASK]` token maps to label c_i is:

$$p\left(V_\phi\left(h\right) = c_i | x_{prompt}\right) = \frac{\exp(C_i \cdot (W_v \cdot h + b))}{\sum_{j=1}^{m} \exp(C_j \cdot (W_v \cdot h + b))}, \tag{5}$$

where C_i is the embedding for c_i and $m = |\mathcal{Y}|$.

Last, we define the cross-entropy loss between the predicted probability $p(y_x|\mathbf{x}) = p(V_\phi(h_{mask})$ and the ground truth y:

$$\mathcal{L} = -\sum_{i=1}^{|\mathcal{Y}|} y_i \log p\left(V_\phi(h_{mask}) = c_i | x_{prompt}\right), \tag{6}$$

the label token with the highest probability is the predicted relation between the entities.

3.2. Relation-Align Module

In this section, we introduce the relation-align module, which aligns the label tokens with the model's decision layer to maintain their features. In relation extraction, input label tokens can lose their distinctiveness and inherent features in the encoder layers, which reduces their effectiveness in aiding the prediction of relations.

To address this, we also add a verbalizer behind each label token's vector to reinforce its semantic meaning, similar to the mask loss. For each label token, we compute its probability score $p(c_i|\mathbf{x}, c_i) = p(\tilde{c}_i = \mathcal{C}|x_{prompt})$ of classifying to itself via the verbalizer V_ϕ, where $\tilde{c}_i$ denotes $V_\phi(h_{c_i})$.

The cross-entropy losses are averaged to obtain the label loss:

$$\mathcal{L}_{label} = -\frac{1}{m}\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m} y_x \log p(c_j|\mathbf{x}, c_i). \tag{7}$$

3.3. Entity-Aware Module

In this section, we introduce an entity-aware module aimed at addressing the issue of a PLM can detect a relation in a sentence, but it may not exist between the two given entities. Inspired by TransE's equation $s + r = o$ relating entities and relations (Bordes et al., 2013), we propose the entity-aware module. It uses the distance $d(s, r, o) = \|s + r - o\|_2$ to connect entities and relations. Furthermore, the dimensions of the encoder outputs for entities and relations are reduced to eliminate redundant features, ϕ_* are trainable parameters:

$$s = \phi_{sub} \cdot h_{sub} \quad o = \phi_{obj} \cdot h_{obj} \quad r = \phi_{rel} \cdot h_{mask}, \tag{8}$$

We contrast the distance $d(s, r, o)$ for positive examples against negative examples (s', r, o'), where s' and o' are randomly sampled spans from the same sentence. Different from previous approaches, we found contrasting within-sentence negative examples was most effective. Thus, the entity-aware loss is:

$$\mathcal{L}_{entity} = -\log\sigma\left(\gamma - d(s, r, o)\right) - \log\sigma\left(d(s', r, o') - \gamma\right), \tag{9}$$

where σ is the sigmoid function and γ is the margin.

3.4. Attention Query Strategy

As shown in Equ. (3), $x_{prompt} = \{[\text{CLS}], T(\mathbf{x}), [\text{SEP}], x, [\text{SEP}]\}$, the x_{prompt} contains both prompt tokens $T(\mathbf{x})$ and sentence tokens x. However, the extral prompt tokens in x_{prompt} can potentially alter the semantics of the origin sentence x, thereby impairing the efficacy of information extraction.

To address this issue, we implement an attention query strategy during the encoding process. We modified the query matrix in the transformer block, specifically, this strategy involves using distinct query matrices for different pairs of tokens in the attention layer. By doing so, we aim to minimize the influence of prompt tokens on the semantics of the sentence tokens. Further details and visual representation of this strategy are provided in Figure 1. For token pair (x_i, x_j), the attention score can be formulated as follows:

$$\text{Attn}(x_i, x_j) = \frac{Q^* x_i \cdot (K x_j)^T}{\sqrt{d}} \cdot V x_j, \tag{10}$$

where Q^* is the query mapping matrix. We select Q_{pp}, Q_{ps}, Q_{sp}, and Q_{ss} for prompt-prompt, prompt-sentence, sentence-prompt, and sentence-sentence pairs, respectively. For example, when x_i is prompt token and x_j is origin sentence token, the $Q^* = Q_{ps}$. These query mapping matrices was initialised with pre-trained $Q*$ enables strong few-shot performance.

3.5. Objective Function

We propose three losses to optimise the model parameters: mask loss for relation prediction, label loss for semantic consistency of label prompt tokens, and entity loss for improving entity awareness.

Our objective is to minimise the final loss $\mathcal{L}$ which is the weighted sum of those losses:

$$\mathcal{L} = \mathcal{L}_{mask} + \alpha_1 \mathcal{L}_{label} + \alpha_2 \mathcal{L}_{entity}, \tag{11}$$

the weights α_1 and α_2 are set to 1 and 0.04 based on extensive experiments.

4. Experiment

4.1. Datasets

In this section, as shown in Table 2, we evaluate the proposed method on four popular relation classification datasets, including TACRED (Zhang et al., 2017), TACREV (Alt et al., 2020), ReTACRED (Stoica et al., 2021), and SemEval (Hendrickx et al., 2010) with the widely used micro F_1 metric.

TACRED is one of the most enormous and widely used crowd-sourced datasets in relation extraction. It is created by combining available human annotations from the TAC KBP challenges and crowd-sourcing, and it contains 42 relation types (including "*no_relation*").

TACREV is a modified version of TACRED. It corrects the label errors in the validation and test sets while leaving the training set unchanged.

ReTACRED is a new, completely re-annotated version of the TACRED dataset. The dataset fixes many labeling errors in the TACRED dataset and refactors the training, validation, and test sets. Thus, ReTACRED has the highest annotation quality, and we use this dataset for a wide range of experiments.

SemEval is a traditional dataset in relation classification. It contains 9 symmetric relations and one special relation "*Other*".

4.2. Baselines

We use $\text{RoBERTa}_{\text{large}}$ (Liu et al., 2019) as the PLM for all our experiments. And our method is compared with the current state-of-the-art approaches for relation classification. Note that, for a fair comparison, we fine-tune all parameters during training. As list in Table 3, these selected baselines cover different aspects of relation classification.

For fine-tuning pre-trained models, KnowBERT (Chen et al., 2022) embeds multiple Knowledge Bases (KBs) into PLM to leverage structured, human-curated knowledge. LUKE (Yamada et al., 2020) predicts randomly masked words and entities in a large entity-annotated corpus retrieved from Wikipedia.

For prompt-based learning models, we select representative works of prompt-based learning models. PTR (Han et al., 2021) uses logic rules to create prompts with sub-prompts that

Table 2: Statistics of different datasets for relation classification.

Dataset	Train	Dev	Test	Relation
TACRED (Zhang et al., 2017)	68,124	22,631	15,509	42
TACREV (Alt et al., 2020)	68,124	22,631	15,509	42
ReTACRED (Stoica et al., 2021)	58,465	19,584	13,418	40
SemEval (Hendrickx et al., 2010)	6,507	1,493	2,717	19

Table 3: F_1 scores (%) on the TACRED, TACREV, ReTACRED and SemEval. The best results are marked **bold**. The result gaps (+/−) are compared to our baseline, KnowPrompt.

Scenario	Methods	TACRED	TACREV	ReTACRED	SemEval
full	KnowBERT (Chen et al., 2022)	71.5	79.3	-	89.1
	LUKE (Yamada et al., 2020)	72.7	80.6	90.3	-
	PTR (Han et al., 2021)	72.4	81.4	90.9	89.9
	KnowPrompt (Chen et al., 2022)	72.4	82.4	91.3	90.2
	FPC (Yang and Song, 2022)	72.9	**82.9**	91.3	90.4
	CCPrefix (Li et al., 2024)	72.6	**82.9**	91.2	90.6
	LabelPrompt (ours)	**73.1**(+0.7)	82.5(+0.1)	**91.6**(+0.3)	**91.3**(+1.1)
32-shot	fine-tuning	28.0	28.2	56.0	80.1
	PTR (Han et al., 2021)	32.1	32.4	62.1	84.2
	KnowPrompt (Chen et al., 2022)	36.5	34.7	65.0	**84.8**
	FPC (Yang and Song, 2022)	35.8	35.5	65.3	-
	CCPrefix (Li et al., 2024)	**37.6**	34.0	65.2	-
	LabelPrompt (ours)	35.4(−1.1)	**36.8**(+2.1)	**66.7**(+1.7)	84.6(−0.2)
16-shot	fine-tuning	21.5	22.3	49.5	65.2
	PTR (Han et al., 2021)	30.7	31.4	56.2	81.3
	KnowPrompt (Chen et al., 2022)	**35.4**	33.1	63.3	**82.9**
	FPC (Yang and Song, 2022)	34.7	34.3	60.4	-
	CCPrefix (Li et al., 2024)	33.4	33.0	61.4	-
	LabelPrompt (ours)	34.4(−1.0)	**35.4**(+2.3)	**63.8**(+0.5)	81.7(−1.2)
8-shot	fine-tuning	12.2	13.5	28.5	41.3
	PTR (Han et al., 2021)	28.1	28.7	51.5	70.5
	KnowPrompt (Chen et al., 2022)	32.0	32.1	55.3	74.3
	FPC (Yang and Song, 2022)	33.6	33.1	57.9	-
	CCPrefix (Li et al., 2024)	30.1	29.8	54.4	-
	LabelPrompt (ours)	**33.9**(+1.9)	**34.8**(+2.7)	**62.1**(+6.8)	**77.0**(+2.7)

encode the prior knowledge of each class. KnowPrompt (Chen et al., 2022) injects latent knowledge in relation labels into the prompts, thus injecting the knowledge into relation labels. FPC (Yang and Song, 2022) incorporates relation prompt learning and a prompt learning curriculum, adapting the model to increasingly difficult tasks. CCPrefix (Li et al., 2024) leverage instance-dependent soft prefixes derived from fact-counterfactual pairs to address verbalizer ambiguity.

Table 4: The F_1 scores (%) of ablation experiments in ReTACRED.

Method	**8-shot**	**16-shot**	**32-shot**	**full**
LabelPrompt	**62.1**	**63.8**	**66.7**	**91.6**
w/o label prompt tokens	57.6 ↓4.5	61.6 ↓2.2	64.7 ↓2.0	91.5 ↓0.1
w/o entity-aware module	58.9 ↓3.2	61.7 ↓2.1	64.7 ↓2.0	91.2 ↓0.4
w/o relation-align module	59.4 ↓2.7	62.2 ↓1.6	64.8 ↓1.9	91.1 ↓0.5
w/o attention query strategy	61.8 ↓0.3	62.7 ↓1.1	65.4 ↓1.3	90.9 ↓0.7
w/ use mask tokens	58.6 ↓3.5	61.5 ↓2.3	63.0 ↓3.7	90.9 ↓0.7
w/ use learnable tokens	58.4 ↓3.7	62.7 ↓1.1	64.5 ↓2.2	91.0 ↓0.6

4.3. Results

We first evaluate and analyse the proposed LabelPrompt method by comparing it with both fine-tuning and prompt-based methods. The experimental results are reported in Table 3.

Few-Shot. LabelPrompt outperforms other methods in most few-shot cases, especially on TACREV and ReTACRED which have more accurate annotations. This shows prompt-learning utilises PLMs' knowledge for low-resource tasks, unlike fine-tuning requiring sufficient data. Compared to state-of-the-art prompt methods, our LabelPrompt also achieves 1.5-4.2% gains on ReTACRED and TACREV. The explicit label tokens reduce target search space early in training.

Full-Data. LabelPrompt also surpasses other methods. It improves 0.7-1.1% over our baseline KnowPrompt on TACRED and SemEval. Even when compared to state-of-the-art methods, we get better performance on TACRED, ReTACRED, and SemEval.

Thus, prompt-learning excels in both few-shot and full-data tasks by leveraging PLMs' knowledge with a simple design.

4.4. Ablation Study

In this section, we evaluate the performance of each component of LabelPrompt in both few-shot and full-data settings. The results are presented in Table 4.

Impact of Label Prompt Tokens: Our prompt-based method relies heavily on label prompt tokens, which strongly influence performance. First, we remove all label prompt tokens from the input and exclude the label loss $\mathcal{L}_{label}$ during training. As Table 4 shows, removing label prompts substantially hurts few-shot performance, indicating their importance for efficient convergence. We also evaluate two variations: 1) replacing label tokens with [`MASK`] and 2) substituting them with random learnable prompt embeddings. As Table 4 shows, it reveals both hurt few-shot accuracy, proving label tokens work due to their semantic correlation with labels, not just their length. Moreover, extra random tokens do not help, as the model struggles to fit them given limited data and performs worse with full data.

Impact of The Entity-Aware Module: Our proposed Entity-Aware Module improves relation classification by helping the model focus on relations between specific entity pairs. As Table 4 shows, removing this module decreases accuracy across all settings, especially for full data, which demonstrate the benefits of including this component in our model.

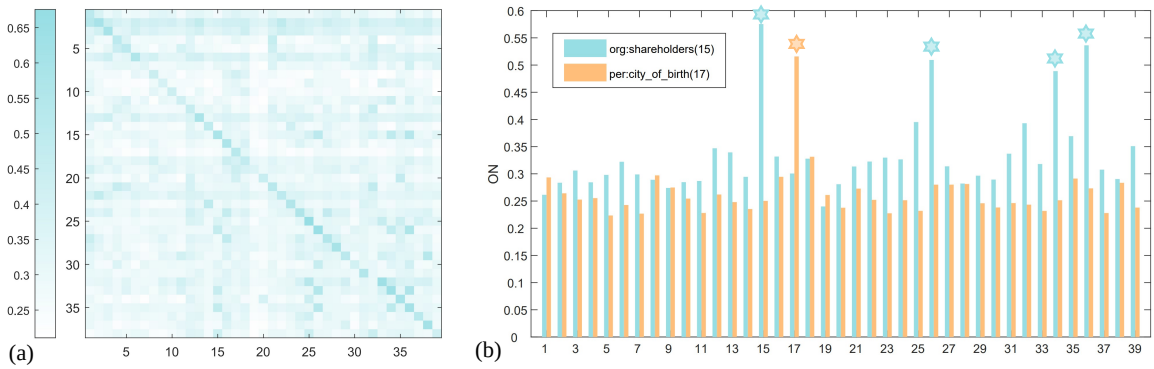

Figure 2: The left subplot shows the ON between the mask token and label tokens for different relations. The right subplot compares the data from rows 15 and 17 of the left subplot in two colours. For example, the blue histogram represents the ON between the mask token and label tokens when the true label is "*org:shareholders*".

Impact of Attention Query Strategy: Label prompt tokens are not defined during pre-training of the PLM. Although initialised with label text semantics, these tokens can still alter sentence semantics when the model processes the sequence. As Table 4 shows, removing the attention query strategy from our model degrades performance across all scenarios.

5. Analysis

In this section, we will conduct a detailed analysis of why label prompt tokens are useful for the relation classification task.

Recent work shows the Feed-Forward Network (FFN) in PLMs may store knowledge. Prompts can activate specific FFN neurons for given inputs, explaining why prompt learning is effective for PLMs. Su et al. (2022) considered each FFN node as a neuron, finding the activation layer corresponds to specific model behaviours.

As shown in Fig 3, we consider the activation output after the first dense layer as the activated neuron sequence $\mathbf{S}$:

$$\mathbf{S} = \{s_{c_1}, s_{c_2}, \ldots, s_{c_m}, s_M\}, \tag{12}$$

where s_M is the mask token sequence.

We define $\mathrm{ON}(s_a, s_b)$ to calculate the Overlapping rate of activated Neurons (ON) between s_a and s_b:

$$\mathrm{ON}(s_a, s_b) = \frac{\sum_{k=1}^{4d} \zeta(s_{a,k} > 0 \wedge s_{b,k} > 0)}{\sum_{k=1}^{4d} \zeta(s_{a,k} > 0) + \sum_{k=1}^{4d} \zeta(s_{b,k} > 0)}, \tag{13}$$

where $\zeta(c) = 1$ if c is true, else 0.

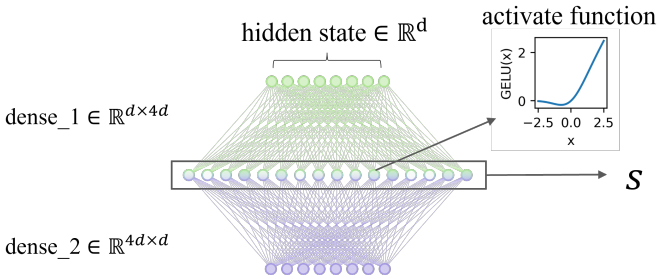

Figure 3: Illustration for activated value sequence s.

The ON quantifies the association between model behaviors. We average ON across samples for each relation to reduce noise. And the result R is plotted in Figure 2 (a), each point $r(i,j) \in R$ is the ON between [MASK] and token c_j when the ground truth label is c_i, which can be calculated as follows:

$$r(i,j) = \frac{1}{|\mathcal{C}_i|}\sum_{u=1}^{|\mathcal{C}_i|} \mathrm{ON}(s_j^{(u)}, s_M^{(u)}) \qquad \text{s.t.} \quad gt = c_i. \tag{14}$$

The higher $r(i,j)$ along the diagonal in Figure 2 (a) shows [MASK] overlaps more with the true label prompt. This demonstrates the label prompt activates pertinent knowledge.

Figure 2 (b) shows ON for "*org:shareholders*" and "*per:city_of_birth*". The orange histogram peaks at token 17 for "*org:shareholders*", while "*per:city_of_birth*" has a high overlap with multiple city-related tokens, this aligns with our expectation.

In summary, label tokens activate knowledge related to the true relation, guiding the model's behaviour. The visualisation and analysis verifies the effectiveness of label prompts.

6. Conclusion

In this paper, we proposed LabelPrompt, a strategy to effectively apply label information for the relation classification task based on prompt learning and prompt engineering. The method focused more on the additional label tokens and correctly predicted the relation class of the entities in a sentence. The experimental results demonstrated that the label prompt tokens are effective in both the few-shot and full-data scenarios.

In future work, we plan to apply the idea of LabelPrompt to LLM-Based information extraction tasks. In LLM, e.g. ChatGPT and LLaMA, the input sequence can be quite long (suck as 4k or even 128k), so the "label prompt tokens" can be upgraded to "label prompt guidelines", providing a more detailed explanation for each class.

Acknowledgments

This work was supported in part by Major Project of the National Social Science Foundation of China (No. 21&ZD166), National Natural Science Foundation of China (61876072) and Natural Science Foundation of Jiangsu Province (No. BK20221535).

References

Heba Abdel-Nabi, Arafat Awajan, and Mostafa Z Ali. Deep learning-based question answering: a survey. *Knowledge and Information Systems*, 65(4):1399–1485, 2023.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.142. URL https://aclanthology.org/2020.acl-main.142.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, pages 2778–2788, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. URL https://www.aclweb.org/anthology/N19-1423.pdf.

Kai Feng, Lan Huang, Kangping Wang, Wei Wei, and Rui Zhang. Prompt-based learning framework for zero-shot cross-lingual text classification. *Engineering Applications of Artificial Intelligence*, 133:108481, 2024. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2024.108481. URL https://www.sciencedirect.com/science/article/pii/S0952197624006390.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*, 2021.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/S10-1006.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. A survey on named entity recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017, 2023. ISSN 2949-7191. doi: https://doi.org/10.1016/j.nlp.2023.100017. URL https://www.sciencedirect.com/science/article/pii/S2949719123000146.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. 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)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

Yang Li, Canran Xu, Guodong Long, Tao Shen, Chongyang Tao, and Jing Jiang. CCPrefix: Counterfactual contrastive prefix-tuning for many-class classification. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2977–2988, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.181.

Fangchao Liu, Hongyu Lin, Xianpei Han, Boxi Cao, and Le Sun. Pre-training to match for unified low-shot relation extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5785–5795, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.397. URL https://aclanthology.org/2022.acl-long.397.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.8. URL https://aclanthology.org/2022.acl-short.8.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Shiao Meng, Xuming Hu, Aiwei Liu, Shuang Li, Fukun Ma, Yawen Yang, and Lijie Wen. RAPL: A relation-aware prototype learning approach for few-shot document-level relation extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5208–5226, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.316. URL https://aclanthology.org/2023.emnlp-main.316.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

F. Petroni, T. Rocktschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel. Language models as knowledge bases? In *2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.608. URL https://aclanthology.org/2021.emnlp-main.608.

T. Shin, Y. Razeghi, Irl Logan, E. Wallace, and S. Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

George Stoica, Emmanouil Antonios Platanios, and Barnabas Poczos. Re-tacred: Addressing shortcomings of the tacred dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13843–13850, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17631.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, et al. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, 2022.

Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.868. URL https://aclanthology.org/2023.acl-long.868.

Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, Enze Shi, Yi Pan, Tuo Zhang, Dajiang Zhu, Xiang Li, Xi Jiang, Bao Ge, Yixuan Yuan, Dinggang Shen, Tianming Liu, and Shu Zhang. Review of large vision models and visual prompt engineering. *Meta-Radiology*, 1(3):100047, 2023. ISSN 2950-1628. doi: https://doi.org/10.1016/j.metrad.2023.100047. URL https://www.sciencedirect.com/science/article/pii/S2950162823000474.

Shuang Wu, Xiaoning Song, and Zhenhua Feng. MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition. 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), pages 1529–1539, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.121. URL https://aclanthology.org/2021.acl-long.121.

Haoran Xu, Benjamin Van Durme, and Kenton Murray. BERT, mBERT, or BiBERT? a study on contextualized embeddings for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6663–6675, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.534. URL https://aclanthology.org/2021.emnlp-main.534.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 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, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.523. URL https://aclanthology.org/2020.emnlp-main.523.

Sicheng Yang and Dandan Song. FPC: Fine-tuning with prompt curriculum for relation extraction. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *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 (Volume 1: Long Papers)*, pages 1065–1077, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.78.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1004. URL https://aclanthology.org/D17-1004.

Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.5. URL https://aclanthology.org/2021.naacl-main.5.

Yulin Zhou, Yiren Zhao, Ilia Shumailov, Robert Mullins, and Yarin Gal. Revisiting automated prompting: Are we actually doing better? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1822–1832, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.155. URL https://aclanthology.org/2023.acl-short.155.