An Exploration of Prompt-Based Zero-Shot Relation Extraction Method

Jun Zhao[1*], Yuan Hu[1*], Nuo Xu[1], Tao Gui[1†], Qi Zhang[1†], Yunwen Chen[2], Xiang Gao[2]
[1] School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[2] DataGrand Information Technology (Shanghai) Co., Ltd., Shanghai, China
{zhaoj19,yuanhu20,tgui,qz}@fudan.edu.cn
xun22@m.fudan.edu.cn
{chenyunwen,gaoxiang}@datagrand.com

Abstract

Zero-shot relation extraction is an important method for dealing with the newly emerging relations in the real world which lacks labeled data. However, the mainstream two-tower zero-shot methods usually rely on large-scale and in-domain labeled data of predefined relations. In this work, we view zero-shot relation extraction as a semantic matching task optimized by prompt-tuning, which still maintains superior generalization performance when the labeled data of predefined relations are extremely scarce. To maximize the efficiency of data exploitation, instead of directly fine-tuning, we introduce a prompt-tuning technique to elicit the existing relational knowledge in pre-trained language model (PLMs). In addition, very few relation descriptions are exposed to the model during training, which we argue is the performance bottleneck of two-tower methods. To break through the bottleneck, we model the semantic interaction between relational instances and their descriptions directly during encoding. Experiment results on two academic datasets show that (1) our method outperforms the previous state-of-the-art method by a large margin with different samples of predefined relations; (2) this advantage will be further amplified in the low-resource scenario.

1 Introduction

Relation extraction (RE) aims to extract the relation between entity pairs from unstructured text. The extracted relation facts can benefit various downstream applications such as knowledge graph completion (Wang et al., 2014), web search (Xiong et al., 2017) and dialog systems (Madotto et al., 2018). However, many effective RE methods (Wu and He, 2019; Du et al., 2018) work within predefined relation sets. They failed to deal with a real-world environment where new relations will emerge after the training phase. These fast-growing new relations make it impossible for us to gather labeled training data for all of them. To recognize the newly emerging relations lacking labeled data, zero-shot RE is of the utmost practical interest.

Despite the great potential of zero-shot RE in real-world applications, there have been relatively few studies focusing on this challenging task. To enable models to predict unseen relations, previous works usually model zero-shot relation extraction as a well-designed task form. Levy et al. (2017) consider relation extraction as a machine reading comprehension. They first associate a few question templates for each relation and then determine which relation satisfies the given sentence and question by model prediction. However, a reasonable and effective question template usually needs careful design, which cannot meet the extraction needs of rapidly growing new relations (Chen and Li, 2021). Therefore, instead of manually constructing question templates, subsequent works (Obamuyide and Vlachos, 2018; Chen and Li, 2021) take advantage of the readily available textual description to represent the new relations, and formulate zero-shot RE as a semantic matching task achieving superior results.

However, current methods usually require a large number of in-domain labeled data of predefined relations to train the model parameters. The learned relational knowledge is mainly from labeled data

* Equal contribution.
† Corresponding authors.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

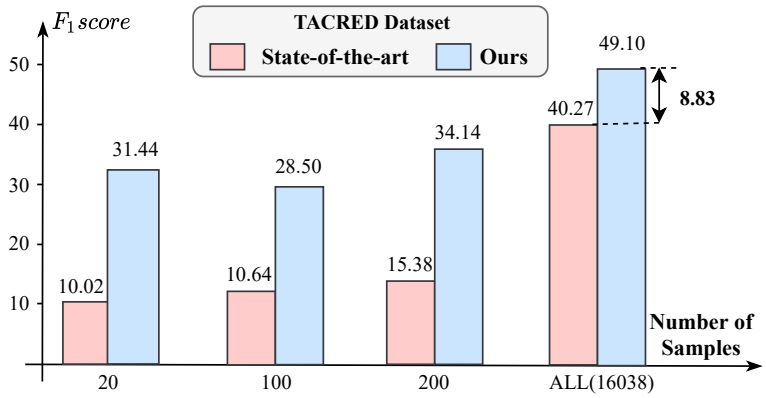

Figure 1: When shifting to some special domains (e.g. medicine, finance) where large-scale labeled data are not available, the performance of these methods on new relations decreases significantly. By inducing the knowledge in the pre-trained language model, our method can approach the results of previous state-of-the-art method ZS-BERT (Chen and Li, 2021) using only 200 labeled data. When using all data, our method improves the F1 score by 8.83%.

itself. As a result, when shifting to some special domains where large-scale labeled data are not available, the performance of these methods on new relations decreases significantly. An experimental illustration is shown in Figure 1. Fortunately, pre-trained language models (PLMs) such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018), can learn a wealth of linguistic (Peters et al., 2018), local syntactic (Hewitt and Manning, 2019) and long-range semantic (Jawahar et al., 2019) from large-scale corpora by self-supervised learning. An interesting question is whether we can reduce the dependence on labeled data of predefined relations with the help of knowledge in PLMs?

To answer this question, in this work, we propose a prompt-based zero-shot RE method. Different from previous methods, in which the learned relational knowledge mainly comes from the labeled data of predefined relations, we leverage prompt to stimulate the rich knowledge distributed in PLMs to reduce dependence on these labeled data. Specifically, we model zero-shot RE as a semantic matching task between relational instance and description. In order to induce the knowledge in PLMs, we fuse the original input with the prompt template to formulate a cloze-style task. Then, we count the probability distribution of the model output and take the words with significant differences between classes as label words. In addition, each predefined relation corresponds to **many** instances and **one** description. The significant quantity gap makes the two-tower methods unable to effectively model the semantics of relation description. Therefore, we directly model the semantic interaction between instances and descriptions during training. Based on the reformulated input and these selected label words, we optimize a semantic matching model, which predicts whether the relation and the textual description match. Experimental results show that our method has very significant advantages when the large-scale labeled data of predefined relations are not available.

To summarize, the main contributions of our work are as follows: (1) We propose a prompt-based zero-shot relation extraction method, which maintains high generalization ability when using even one labeled data per predefined relations. (2) We design comprehensive experiments to analyze the impact of predefined relations and prompt composition on the generalization performance of the model in the low-resource scenario, which may enlighten the following work. (3) Experiment results on two academic datasets show that our method outperforms the previous state-of-the-art method by a large margin and this advantage will be further amplified in low resource scenarios.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

2 Related Work

2.1 Knowledge in Pretrained Language Model

Contextual word representations derived from pre-trained language models have recently been shown to provide significant improvements to the state of the art for a wide range of NLP tasks, motivating a growing body of research investigating what aspects of linguistic knowledge they are able to learn from unlabeled data. Peters et al. (2018) showed that different neural architectures (e.g., LSTM, CNN, and Transformers) can hierarchically structure linguistic information that varies with network depth. (Jawahar et al., 2019; Clark et al., 2019; Goldberg, 2019) show that such hierarchy exists as well for BERT models that are not trained using the standard language modeling objective. More recently, many studies (Tenney et al., 2019; Hewitt and Manning, 2019) probe the knowledge within PLMs from various perspectives and find that the existing models trained on language modeling and translation produce strong representations for syntactic phenomena. Together, these results suggest that pre-trained language models entail comprehensive linguistic knowledge, which accounts for its great performance on downstream tasks and proves its potential to represent the samples of zero-shot relation extraction tasks, which has limited training data.

2.2 Prompt-Based Optimization

Since the advent of prompt tuning, it has soon become the prevailing paradigm of natural language processing. Prompt tuning is based on language models that estimate the probability of text. It modifies the original input of downstream tasks to a prompt with unfilled positions, and predicts the output based on the slot-filling result by language models (Liu et al., 2021). This method has been proven to be helpful on various NLP tasks, including text classification (Han et al., 2021), entity typing (Ding et al., 2021), text generation (Li and Liang, 2021), and also multi-modal tasks (Tsimpoukelli et al., 2021). Current studies have made some attempts to derive knowledge from PLMs with prompts. Jiang et al. (2020) proposed mining-based and paraphrasing-based methods to automatically generate high-quality prompts, which boosted the performance of knowledge-driven tasks. Zhong et al. (2021) conducted a set of control experiments to disentangle the efforts of training data and pre-trained knowledge. Inspired by these works, compared with direct fine-tuning, using the limited labeled data to derive the existing relational knowledge in the pretrained model is a better choice.

3 Method

We reformulate the task of zero-shot relation extraction as a semantic matching task optimized by prompt-tuning. In this section, we will introduce our proposed method in detail. We start by defining the problem we will tackle. Then we introduce how we reformulate zero-shot relation extraction, our prompt design and the selection of label tokens. Finally, we introduce the strategy of making predictions with our model.

3.1 Problem Definition

For the zero-shot relation extraction task, we expect the model $\mathcal{M}$ to predict the right relation of two annotated entities within the text, where the candidate relations are unseen during training.

Formally, let $R_s = \{r_s^1, \ldots, r_s^n\}$ denotes the set of predefined relations. Each relation in R_s has a corresponding textual description, composing the set of relation descriptions $D_s = \{d_s^1, \ldots, d_s^n\}$. In the train set $S_s = \{S_s^1, \ldots, S_s^N\}$, each sample $S_s^i = (x^i, r_s^i)$ consists of a relational instance x^i and its relation label $r_s^i \in R_s$, in which the relational instance x^i is a piece of text s^i with annotated entities e_1^i and e_2^i, namely $x^i = \langle s^i, e_1^i, e_2^i \rangle$. Similarly, the set of unseen relations for testing is denoted as $R_u = \{r_u^1, \ldots, r_u^m\}$, together with the corresponding description set $D_u = \{d_u^1, \ldots, d_u^m\}$. Note that all relations in R_u are unseen during training, i.e. $R_s \cap R_u = \varnothing$. The test set is denoted as $S_u = \{S_s^1, \ldots, S_u^M\}$, in which each test sample $S_u^j = (x^j, r_u^j)$.

3.2 Task Reformulation

In our work, we model zero-shot relation extraction as a semantic matching task where we need to recognize the semantic equivalence relations between relational instances and the description of their

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

Input		Label
Original Sample	**Prompt**	
Cloud Nothings was formed in Cleveland .	[CLS] [CT] Premise : *input text* [SEP] [CT] Hypothesis : *relation description* . Answer : [MASK] [SEP]	Place of Foundation
Reformulated Samples		
[CLS] [CT] Premise : Cloud Nothings was formed in Cleveland [SEP] [CT] Hypothesis : location where a group or organization was formed . Answer : [MASK] [SEP]		match
[CLS] [CT] Premise : Cloud Nothings was formed in Cleveland [SEP] [CT] Hypothesis : musical instrument that a person plays . Answer : [MASK][SEP]		not_match
[CLS] [CT] Premise : Cloud Nothings was formed in Cleveland [SEP] [CT] Hypothesis : league in which team or player plays or has played in . Answer : [MASK] [SEP]		not_match
[CLS] [CT] Premise : Cloud Nothings was formed in Cleveland [SEP] [CT] Hypothesis : heritage designation of a historical site . Answer : [MASK] [SEP]		not_match

Table 1: An example of the reformulation of zero-shot relation extraction task. Each original sample is paired with various descriptions to form new samples.

corresponding relation labels. Specifically, we pair each test sample with the description of every candidate relation, and label them with `match`/`not_match` to form semantic matching samples. And we set it to have half the probability of pairing the training sample with the non-corresponding relation description and half the probability of pairing it with the corresponding relation description. Therefore, the number of positive and negative semantic examples in the training set is roughly equal. As shown in Table 1, the pair is labeled as `match` only when the description matches the corresponding relation label of the relational instance.

Formally, taking the training sample $S_s^i = (x^i, r_s^i)$ for example, we can derive a semantic matching sample $\{(x^i, d_s^k, y^k)\}$ from it, where

$$y^k = \begin{cases} \texttt{match} & r_s^i = r_s^k \\ \texttt{not\_match} & otherwise, \end{cases} \tag{1}$$

We denote the newly derived train set for semantic matching as $S_s' = \{(x^i, d_s^k, y^{ik})\}_{i=1...N}$. Note that from each test sample we will derive m semantic matching samples. The test set is denoted as $S_u' = \{(x^j, d_u^l, y^{jl})\}_{j=1...M, l=1...m}$. In summary, the above efforts convert the original problem to a semantic matching task, which is basically a 2-classification task that we could handle.

Is the two-tower architecture suitable for this task? The state-of-the-art zero-shot methods (Obamuyide and Vlachos, 2018; Chen and Li, 2021) adopt a two-tower architecture to implement the above semantic matching model. However, encoding instances and descriptions in isolation is not a good choice. Assuming that we use 10 relations and 100 instances of each relation to train a two-tower model, there are 1000 different inputs for instance encoder and only 10 inputs for the description encoder. This significant gap makes it difficult for description encoder to learn semantics effectively. Different from the two-tower architecture, the proposed method directly models the semantic interaction between instances and description during encoding. We will show the significant improvement brought by this change in the experiments.

3.3 Model with Prompt Tuning

To model the semantic matching between relational instances and descriptions, we take advantage of pre-trained language models together with prompt tuning. Noticeably, for zero-shot relation extraction, the most critical issue during training is that very few relation descriptions are exposed to the model. Furthermore, all of the descriptions in the test set are unseen in training. Thus, the rich linguistic knowledge of PLM is necessary to ensure that the model understands the descriptions with limited training. Additionally, to tackle the discrepancy of PLM between the pre-training and fine-tuning stage, prompt tuning is necessary to reformulate downstream tasks as cloze-style tasks that BERT is good at. We believe that prompt tuning provides an effective way to fully export knowledge from pre-trained language models and also enables few-shot learning of the task. Due to the discussions, we build our model based on BERT, which learns the objectives by prompt tuning.

Prompt Design For each reformulated sample (x, d, y), we fill the original text of relational instance and the description into a prompt. We define the prompt x' for relational instance x and relation description d

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

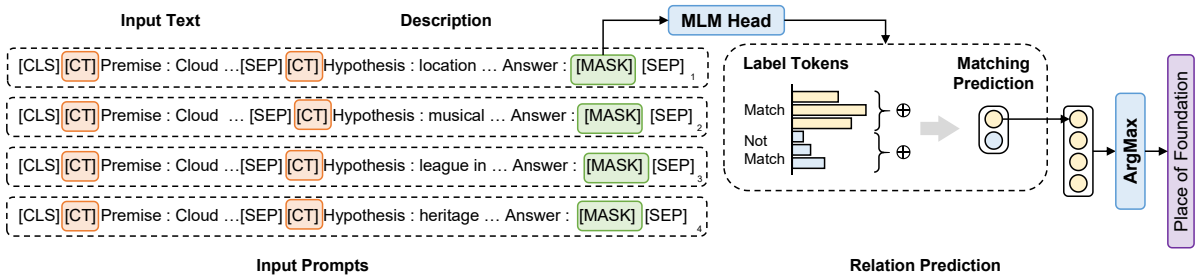

Figure 2: An overview of the process of relation prediction. The [MASK]ed positions within input prompts are firstly filled by language model, then the logits of label tokens are collected to predict the matching probability of input text and description. Lastly, we collect the matching probabilities for each pair and estimate the distribution of relational labels based on them.

as

$$\begin{aligned} x' = & \texttt{[CLS][CT]}\ s'\ \texttt{[SEP][CT]}\ d' \\ & \texttt{[MASK][SEP]} \\ \text{where}\ \ s' = & \text{Premise: } s \\ d' = & \text{Hypothesis: } d \text{ Answer:}, \end{aligned} \tag{2}$$

where s is the input text, which is the original text of x; s' and d' denotes the prompt-formulated input text and description respectively; [CT] denotes T different continuous tokens that make up the template. Examples of input prompts could be seen in Table 1. The design of prompt aims to fully utilize the ability of BERT as a rich knowledge base, and the introduction of continuous tokens in template aims to enhance the representation ability of the prompt, since these tokens could be optimized in the whole embedding space.

Label Token Selection Following the common settings of prompt tuning on classification task, we also determine label tokens for each category (namely match or not_match) for consequential prompt tuning. Basically, for the two categories, the probability distributions of masked language modeling should be different and distinguishable. Thus, retrieving label tokens is the process of capturing features that indicate the distribution associated with a certain category. We solve the problem by estimating the distributions and retrieving tokens that have the most significant difference of probability among distributions.

Formally, we partition the reformulated train set S'_s by category of label y. The matched and unmatched samples are denoted as $S_{sm} = \{(x, d, y) \in S'_s | y = \texttt{match}\}$ and $S_{sn} = \{(x, d, y) \in S'_s | y = \texttt{not\_match}\}$, respectively. The prompts of samples are then fed to BERT. For sample (x, d, y), the estimated distribution of the [MASK] token is calculated as

$$P(w|x, d) = \text{softmax}(W(\text{MLM}(x')) + b), \tag{3}$$

where w denotes every token in vocabulary, $P(w|x, d)$ indicates the estimated MLM distribution of the sample, MLM denotes the output embedding of [MASK] token, W and b denote trainable weights of linear projection.

The MLM distribution of categories is estimated by averaging the predicted distributions among samples in the category:

$$P_m(w) = \frac{1}{|S_{sm}|} \sum_{(x,d,y)}^{S_{sm}} P(w|x, d), \tag{4}$$

$$P_n(w) = \frac{1}{|S_{sn}|} \sum_{(x,d,y)}^{S_{sn}} P(w|x, d), \tag{5}$$

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

where $P_m(w)$ and $P_n(w)$ indicate the estimated MLM distribution of the category `match` and `not_match`, $|S_{sm}|$ and $|S_{sn}|$ denote the number of matched and unmatched samples, respectively.

Finally, for each category, the tokens with top-K possibility difference between the MLM distribution within and without the category are selected as the label tokens. The possibility difference of each word is divided by their estimated occurrence possibility to ensure fair comparison.

$$\{w_m^1, \ldots, w_m^K\} = \underset{w}{\text{topK}} \frac{P_m(w) - P_n(w)}{P_m(w) + P_n(w)}, \tag{6}$$

$$\{w_n^1, \ldots, w_n^K\} = \underset{w}{\text{topK}} \frac{P_n(w) - P_m(w)}{P_m(w) + P_n(w)}. \tag{7}$$

In Eq.6 and 7, K is the number of tokens selected for each category, $\{w_m^1, \ldots, w_m^K\}$ and $\{w_n^1, \ldots, w_n^K\}$ denote the selected label tokens of the category of {`match` and `not_match`} respectively.

3.4 Training and Inference

In this part, we introduce our strategy to derive relation predictions from the semantic matching model, along with the training objectives.

Similar to other prompt-based methods, the output possibilities of label tokens are collected to perform a 2-classification on label y. The possibility of categories is proportional to the production possibility of label words. As shown in Eq.8, in implementation, we achieve this by adding the output logits of label tokens and applying softmax on them:

$$P(y = c|x, d) = \text{softmax}\left(\sum_{k=1}^{K} \log P(w_c^k|x, d)\right), \tag{8}$$

where $c \in \{\texttt{match}, \texttt{not\_match}\}$.

The prediction of relation label for relational instance x_i is done by collecting the possibilities of match between x_i and the descriptions of every candidate relation $R^k \in R$. As in Eq. 9, the matching possibilities of x_i and all candidate relations are collected as logits and are put to a softmax function to predict the distribution of the relation label.

$$p^{ik} = P(y^{ik} = \texttt{match}|x^i, D^k), \tag{9}$$

$$P(r^k|x^i) = \frac{\exp(p^{ik})}{\sum\limits_{r^k \in R} \exp(p_{ik})}. \tag{10}$$

Lastly, the model is trained on cross-entropy loss L_{CE} to maximize the log-likelihood of all training samples.

$$L_{CE} = \sum_{i=1}^{N} \text{CrossEntropy}(r_s^i, \{p_s^{ik}\}_{k=1}^n). \tag{11}$$

As for making prediction on unseen samples, i. e. evaluating model on test sets, for each test sample S_u^j, the predicted relation distribution of relational instance x_j is illustrated in Eq. 12 and 13. We pick the relation with the highest possibility as the predicted result.

$$p^{jl} = P(y^{jl} = \texttt{match}|x^j, d_u^l), \tag{12}$$

$$P(R_u^l|x^j) = \frac{\exp(p^{jl})}{\sum\limits_{r_u^j \in R_u} \exp(p_{ik})}, \tag{13}$$

$$\hat{r}_u^j = \underset{l}{\text{argmax}} P(R_u^l|x^j). \tag{14}$$

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

Dataset	# Inst.	# relations	% N/A
FewRel	56000	80	-
TACRED	106264	42	79.5%

Table 2: Original statistics of datasets FewRel and TACRED. %N/A is the proption of label "no_relation" and "-" represents there is no N/A instances.

4 Experimental Setup

In this section, we describe the datasets for training and evaluating the proposed method. We also detail the baseline models for comparison. Finally, we clarify the implementation details and hyperparameter configuration of our method.

4.1 Datasets

Our main experiments are conducted on two relation extraction datasets: FewRel and TACRED. The original statistics of the two datasets are listed in Table 2.

FewRel (Han et al., 2018a). There are 80 relations included in FewRel, a high-quality RE dataset with 56,000 instances from Wikipedia. To be consistent with the previous state-of-the-art method, we rearrange the dataset. To be specific, we choose 65 relations as labeled set with predefined relation and select 15 relations as the unlabeled set with unseen relations.

TACRED (Zhang et al., 2017). TACRED is a human-annotated relation extraction dataset that contains 106,264 examples with 42 kinds of relations(including “no_relation”). The instances of special class "no_relation" is removed, and we use the remaining 21,773 instances for training and evaluation.

We also add a low-resource setting, which means the size of training data is small. Under the setting, the development set is provided, with about 5 examples per relation. As shown in Table 3 and Table 4, the three different values of n represents the number of data used for training are only 20, 100 and 200 respectively. For the setting, We randomly sample training data from each relation category roughly evenly. Note that when sampling 20 training data, the number of relation categories in the training set of both datasets is also reduced to 20. For both of the two datasets, we use the Macro-F1 score as the main metric to evaluate the model’s performance.

4.2 Compared Methods

To verify the effectiveness of our proposed method, we select the following models for comparison. The state-of-the-art method ZS-BERT (Chen and Li, 2021) adopted the two-tower architecture, this method encodes sentences and relation descriptions separately and uses nearest neighbor search as the matching function to obtain the prediction of unseen relations. When comparing with R-BERT (Wu and He, 2019) and Attentional Bi-LSTM (Zhou et al., 2016), two supervised relation extraction (SRE) models, we take the same way as ZS-BERT (Chen and Li, 2021) so that SRE models can carry out zero-shot prediction. Specifically, we change the last layer to a fully-connected layer with tanh activation function. Based on the input instance embedding and relation description’s embedding, the nearest neighbor search will be applied to generate the zero-shot prediction. We also compare our method with ESIM (Chen et al., 2017), a semantic matching model. To have a fair comparison, the strategy to generate relation predictions from the semantic matching model is the same as ours. Finally, we introduce BERT(CLS) (Devlin et al., 2019) to intuitively show the performance improvement brought by modeling the semantic interaction between instances and descriptions during encoding.

4.3 Implementation Details

We adopt BERT-base-cased as the encoder and all experiments are conducted using a NVIDIA GeForce RTX 3090 with 24GB memory. The number of continuous tokens is $t = 4$. We use AdamW for optimization, in which the initial learning rate is 3e-5. Taking into account the randomness of network initialization and random selection of n training instances, we run our experiment 5 times and the results

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

Method	FewRel(m=15)											
	n=20			n=100			n=200			n=all		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Att Bi-LSTM (Zhou et al., 2016)	14.19	13.88	14.03	15.75	19.8	17.55	20.83	26.00	23.13	38.13	32.05	34.82
ESIM (Chen et al., 2017)	0.60	5.45	1.08	0.90	6.56	1.58	7.66	7.38	7.52	36.97	32.51	34.60
R-BERT (Wu and He, 2019)	8.40	8.38	8.39	13.61	15.90	14.67	16.05	18.58	17.22	32.25	25.58	28.53
ZS-BERT (Chen and Li, 2021)	6.04	6.36	6.20	6.34	7.93	7.05	8.35	9.59	8.93	35.54	38.19	36.82
BERT(CLS) (Devlin et al., 2019)	**44.95**	33.65	38.49	49.99	47.20	48.55	**53.14**	52.13	52.62	**67.62**	59.12	63.09
Ours	44.94	**45.72**	**45.33**	**50.21**	**51.72**	**50.96**	52.49	**53.98**	**53.23**	64.48	**62.45**	**63.45**

Table 3: Main results on FewRel. The best results are bold. n is the number of provided training data and m represents unseen relations' number.

Method	TACRED(m=11)											
	n=20			n=100			n=200			n=all		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Att Bi-LSTM (Zhou et al., 2016)	14.33	11.38	12.68	13.73	10.64	11.99	15.68	21.70	18.20	25.20	20.17	22.41
ESIM (Chen et al., 2017)	9.09	0.15	0.29	8.54	9.41	8.96	1.52	9.15	2.61	26.99	18.38	21.87
R-BERT (Wu and He, 2019)	14.59	7.27	9.70	18.93	12.12	14.78	23.62	19.67	21.47	44.66	45.86	45.25
ZS-BERT (Chen and Li, 2021)	10.79	9.35	10.02	12.53	9.25	10.64	14.98	15.79	15.38	38.08	42.72	40.27
BERT(CLS) (Devlin et al., 2019)	25.53	19.78	22.29	9.34	10.55	9.91	**37.97**	**34.43**	**36.11**	**51.90**	44.71	48.03
Ours	**32.40**	**30.54**	**31.44**	**38.12**	**22.75**	**28.50**	34.56	33.73	34.14	51.85	**46.63**	**49.10**

Table 4: Main results on TACRED. The best results are bold. n is the number of provided training data and m represents unseen relations' number.

we report are the average results. Other results of compared methods are gotten when the parameters remain the same as its own published source code. We follow Soares et al. (2019) to augment each instance with four reserved word pieces to mark the begin and end of each entity. The relation descriptions of FewRel are obtained from (Han et al., 2018b) and TACRED's are obtained from the TAC-KBP relation ontology guidelines[2].

5 Results and Discussion

5.1 Main Results

The main results of our experiments on FewRel and TACRED are listed in Table 3 and Table 4. **First**, as can be seen, the method we propose steadily outperforms compared methods, and even the previous state-of-the-art method (Chen and Li, 2021) performs much worse than our method when targeting at different number of training instances. The reason is that the two-tower model which the previous state-of-the-art method (Chen and Li, 2021) encodes the input instances and candidate relations with large quantitative differences separately, and we argue that this modeling choice is insufficiently expressive for modeling the semantic matching between instances and relation descriptions. What's more, the simple matching function (ZS-BERT uses nearest neighbor search) is incapable of capturing the complicated interactions between input sentences and relation descriptions. Our proposed method yields rich interactions between the input instance and candidate relation description, as they are jointly encoded to obtain a final representation. At the layers of transformer, every word in the candidate relation description can attend to every word in the input instance, and vice-versa, so our proposed method can produce a candidate-sensitive input representation, which the ZS-BERT cannot. **Second**, it can be apparently found that the baseline's performance decreases significantly when the number of labeled data decreases, which indicates that large number of in-domain labeled data of predefined relations is a prerequisite for their good performance. While our method manage to derive the original knowledge in PLMs with prompt so that our method still performs well when the labeled data is scarce. For FewRel, our MACRO-F1 score reaches 45.33% training with 20 instances, which is better than the result of previous state-of-the-art using the complete

[2]https://tac.nist.gov/2015/KBP/ColdStart/guidelines/ TAC_KBP_2015_Slot_Descriptions_V1.0.pdf

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

Method	FewRel_TACRED n=all			TACRED_FewRel n=all		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Att Bi-LSTM (Zhou et al., 2016)	21.86	27.72	24.44	31.27	39.26	34.82
ESIM (Chen et al., 2017)	22.67	18.91	20.62	19.38	11.93	14.77
R-BERT (Wu and He, 2019)	23.10	28.49	23.98	15.31	14.70	15.00
ZS-BERT (Chen and Li, 2021)	35.90	29.78	32.55	17.69	11.81	14.16
Ours	**41.26**	**37.62**	**39.36**	**60.01**	**50.74**	**54.99**

Table 5: Results on two constructed cross-domain tasks.

Prompt	FewRel	TACRED
`[PRE]` Question : `[HYP]` . true or false ? Answer : `[MASK]`	63.11	47.19
`[PRE]` Question : `[HYP]` ? `[MASK]`	61.09	48.79
`[PRE]` Is `[HYP]` true ? Answer : `[MASK]`	**63.58**	47.72
Does `[HYP]` agree with `[PRE]` ? `[MASK]`	62.44	45.73
Ours	63.45	**49.10**

Table 6: Results on different prompts.

train dataset. Such results verify the strong ability of low-resource learning for our proposed method.

5.2 Cross Domain Analysis

Through the analysis of main results, we have concluded that large-scale labeled data of predefined relations is a prerequisite for the existing model to achieve good generalization performance on unseen relations. An ensuing question is: when we deal with the problem of a field that lacks labeled data, can we solve this problem by using labeled data with existing relations in common fields? To answer this question, we conducted experiments on two constructed cross-domain zero-shot relation extraction tasks.i.e.,: FewRel to TACRED and TACRED to FewRel. Specifically, pre-defined relations and their labeled instances come from the source domain training dataset, and we evaluate performance on the target domain testing dataset.

Table 5 shows the results. By comparing with the in-domain experimental results in the main experiment, we can find: the change of domain does increase the semantic gap between the pre-defined and unseen relations. As a result of that: For FewRel to TACRED, the experimental result of our method is reduced from 49.10% to 39.36%, and for TACRED to FewRel, the result is reduced from 63.45% to 54.99%. But our performance still outperforms compared methods, which shows the proposed method's generalization on unseen relations.

5.3 Influence of Pre-defined Relation Number

In this subsection, we study the effect of the number of seen predefined relations in the train dataset. And we conduct the experiment on FewRel. For FewRel, the original number of predefined relations is 65, we sample 33,17,9,5 classes from the original train dataset in turn, which correspond to 50%, 25%, 12.5%, 6.25% of the original classes represented by the scale on the horizontal axis in the figure. The results of Figure 3 prove that the number of pre-defined relations does matter. As the number decreases, the knowledge learned from the training set also decreases, which can weaken the model's generalization of unseen relations. So the performance of our proposed method also gets worse. Nevertheless, our method can still be said to perform well. For FewRel, When we reduce the number of predefined relation types to 5, our performance still outperforms the previous state-of-the-art, which can validate the effectiveness of our proposed method.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

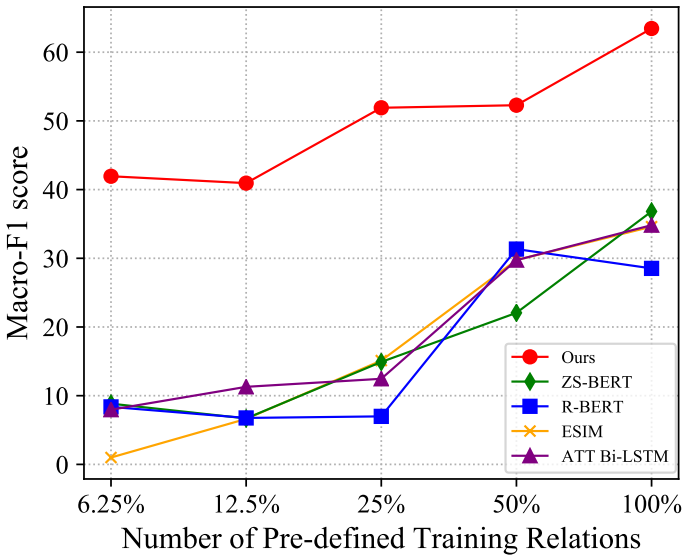

Figure 3: Model results with different number of predifined training relations on FewRel.

5.4 Analysis on Different Prompt Forms

To explore the impact of different forms of prompt on the performance of the proposed method, we conducted experiments on two datasets based on different prompts. Because the continuous tokens' position relative to `[HYP]` and `[PRE]` doesn't change, it is omitted from the table. As is shown in Table 6, inappropriate forms may lead to worse results, but on the other hand, a suitable prompt form can also improve model performance since it can help elicit the existing knowledge in PLMs. Among all the prompt forms, the form we have chosen is relatively well-behaved. Moreover, the prompt's performance is not necessarily the same as our intuition, in other words, the prompt we think good is not necessarily good for PLMs and we think the automatic generation of prompts is a promising research direction.

6 Conclusions

In this work, we introduce a prompt-based zero-shot relation extraction method, which still maintains superior generalization performance under low-resource settings. We clarify the limitations of the two-tower architecture in previous state-of-the-art methods, and directly model the interaction between instances and descriptions during encoding, which breaks the performance bottleneck of the previous model. The introduce of prompt-tuning effectively elicit the knowledge in PLMs and significantly reduces the dependence on predefined relations. We believe that these are the reasons why our method achieves excellent results. Experiment results on two academic datasets show that our method outperforms the previous state-of-the-art method by a large margin and this advantage will be further amplified in low resource scenarios.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

References

Chih-Yao Chen and Cheng-Te Li. 2021. Zs-bert: Towards zero-shot relation extraction with attribute representation learning. *arXiv preprint arXiv:2104.04697*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. 2021. Prompt-learning for fine-grained entity typing. *arXiv preprint arXiv:2108.10604*.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. *arXiv preprint arXiv:1809.00699*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018a. 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*, pages 4803–4809.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018b. 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*, pages 4803–4809, Brussels, Belgium, October-November. Association for Computational Linguistics.

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

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, August. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia, July. Association for Computational Linguistics.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium, November. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, October-November. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *arXiv preprint arXiv:2106.13884*.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1271–1279, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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*, pages 35–45, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online, June. Association for Computational Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.