

FPC: Fine-tuning with Prompt Curriculum for Relation Extraction

Sicheng Yang, Dandan Song*

School of Computer Science and Technology,
Southeast Academy of Information Technology,
Beijing Engineering Research Center of High Volume Language Information
Processing and Cloud Computing Applications,
Beijing Institute of Technology, Beijing, China
yangsicheng@bit.edu.cn, sdd@bit.edu.cn

Abstract

The current classification methods for relation extraction (RE) generally utilize pre-trained language models (PLMs) and have achieved superior results. However, such methods directly treat relation labels as class numbers, therefore they ignore the semantics of relation labels. Recently, prompt-based fine-tuning has been proposed and attracted much attention. This kind of methods insert templates into the input and convert the classification task to a (masked) language modeling problem. With this inspiration, we propose a novel method Fine-tuning with Prompt Curriculum (FPC) for RE, with two distinctive characteristics: the relation prompt learning, introducing an auxiliary prompt-based fine-tuning task to make the model capture the semantics of relation labels; the prompt learning curriculum, a fine-tuning procedure including an increasingly difficult task to adapt the model to the difficult multi-task setting. We have conducted extensive experiments on four widely used RE benchmarks under fully supervised and low-resource settings. The experimental results show that FPC can significantly outperform the existing methods and obtain the new state-of-the-art results.

1 Introduction

As one of the essential tasks in natural language processing (NLP), relation extraction (RE) intends to extract relational facts hidden in text. Figure 1 shows the typical RE setting: a sentence with two marked entities ("Tesla" and "Elon Musk") is input into a model to classify the relation (founded by) between the entities. Structured knowledge captured by RE can benefit many downstream applications such as knowledge graph completion (Bordes et al., 2013), dialogue systems (Madotto et al., 2018) and question answering (Bordes et al., 2014).

As the mainstream of RE, the classification methods extract semantic features from text to form

Sentence: Elon Musk is known for co-founding Tesla.
founded by

Figure 1: An example to show the typical RE setting.

relation representations (vectors). Then the representations are fed into a classifier to predict relation labels. The recent classification methods generally utilize pre-trained language models (PLMs) and have achieved promising results. This is because self-supervised learning on large-scale unlabeled data makes PLMs obtain rich knowledge, which is important for natural language understanding (Devlin et al., 2019; Liu et al., 2019) and generation (Raffel et al., 2020; Lewis et al., 2020). However, such methods directly treat relation labels as class numbers, hence they can not capture the semantics of relation labels.

On the contrary, the reformulation methods can improve the deficiency by intuitively transform RE into other tasks such as question answering (QA) (Levy et al., 2017). For example, some questions are designed based on relational semantics and a QA model is utilized to produce answers. Prompt-based fine-tuning (Schick and Schütze, 2021) is a new kind of reformulation method which is originated from GPT-3 (Brown et al., 2020) and has attracted much attention. This kind of methods insert templates into the input and convert the classification task to a (masked) language modeling problem. For example, in a binary sentiment classification task, we use a template $T(\cdot) = "\cdot$ It is [MASK]." and a set of label words $\mathcal{V} = \{"great", "terrible" \dots\}$. Each instance is modified by the template and then input into the PLM to produce the probability of the label words to fill the masked token(s). There is a mapping function (verbalizer) that links the label words to the specific classes $\mathcal{M} : \mathcal{V} \rightarrow \mathcal{Y}$. Therefore the probability distribution over \mathcal{Y} can be formalized with the probability distribution over \mathcal{V} .

*Corresponding author

Inspired by this, we propose a novel method Fine-tuning with Prompt Curriculum (FPC) for RE, with the following two distinctive characteristics:

The relation prompt learning introduces an auxiliary prompt-based fine-tuning task to the classification model, aiming to make the model capture the semantics of relation labels. We manually design a template with language words and consecutive mask tokens ([MASK]), which can "enquire" the relation expressed by the input. The words of relation labels are directly used with a little modification to form the prediction targets for the mask tokens. We insert the template into each instance to bring a cloze-style auxiliary task to the model. Provided the new input, the model is fine-tuned to classify relation labels and fill the mask tokens with the target word tokens through masked language modeling (MLM) simultaneously.

The prompt learning curriculum is a fine-tuning procedure including an increasingly difficult task. This task-level curriculum helps the model to build the connections between class numbers and the prediction targets of the cloze-style auxiliary task. We design an "easy" sub-task where a part of instances directly shows the prediction targets. All instances are divided into two types: "mask" and "unmask". While "mask" instances are in the original input format as described above, "unmask" instances are formed by replacing the mask tokens with the corresponding prediction targets. During fine-tuning, the proportion of "mask" instances gradually increases, which should be low at the beginning and become 100% before the end. As the number of instances showing the prediction targets decreases, the sub-task gradually becomes "harder" and finally turns into the target task, which adapts the model to the multi-task setting.

In summary, the contributions of our work are concluded as follows:

(1) We propose a novel method Fine-tuning with Prompt Curriculum (FPC) for RE, which enables the model to capture the semantics of relation labels through a cloze-style auxiliary task introduced by the relation prompt learning.

(2) We design the prompt learning curriculum to adapt the model to the multi-task setting with an increasingly difficult task.

(3) We conduct extensive experiments on four widely used RE datasets under fully supervised and low-resource settings. The results show that FPC significantly outperforms the existing methods and

achieve the new state-of-the-art results¹.

2 Related Work

2.1 Relation Extraction

We can divide the recent RE methods into two classes: classification and reformulation. The early classification methods (Zhang et al., 2017; Zhang et al., 2018) construct complicated models to capture semantic features. In recent years, fine-tuning PLMs (Devlin et al., 2019; Liu et al., 2019) can achieve remarkable results since PLMs have acquired rich knowledge from large-scale unlabeled data. The following studies focus on designing effective pre-training objectives such as span-level modeling (Joshi et al., 2020) and contrastive learning (Soares et al., 2019; Peng et al., 2020) to further improve PLMs. Because entity information is important for comprehending relational semantics, a series of methods (Zhang et al., 2019; Peters et al., 2019; Yamada et al., 2020) integrate entity embedding into PLMs. The reformulation methods can leverage the recent advances or datasets of other tasks to boost RE. Such methods intuitively transform RE into other targets like question answering (Levy et al., 2017; Li et al., 2019), natural language inference (Sainz et al., 2021) and translation (Paolini et al., 2021; Wang et al., 2021a).

2.2 Prompt-based Fine-tuning

Fueled by the emergence of GPT-3 (Brown et al., 2020), prompt-based fine-tuning has drawn much attention. This kind of approaches can bridge the gap between pre-training and fine-tuning and effectively stimulate knowledge distributed in PLMs. A series of prompt-based studies on knowledge probing (Trinh and Le, 2018; Petroni et al., 2019; Davison et al., 2019), text classification (Schick and Schütze, 2021; Liu et al., 2021b), relation extraction (Han et al., 2021; Chen et al., 2022) and entity typing (Ding et al., 2021) have achieved promising results. To avoid the cumbersome process of prompt construction, the following methods (Schick et al., 2020; Shin et al., 2020; Gao et al., 2021) focus on searching and generating prompts automatically. Some studies (Li and Liang, 2021; Qin and Eisner, 2021; Lester et al., 2021) propose to tune continuous prompts and fix the entire PLM parameters, which is effective for large-scale PLMs with billions of parameters.

¹Our experimental implementation is available at <https://github.com/yangsc98/FPC>

2.3 Curriculum Learning

Inspired by the meaningful learning order of human, curriculum learning (CL) (Bengio et al., 2009) aims to train a model with "easy" data or sub-task whose difficulty is gradually increasing. The training process finally adapts the model to "hard" data or task, aiming to train better and faster (Wang et al., 2021b). CL methods can be divided into two classes: data-level and task-level. In the field of NLP, CL has been widely used for machine translation. The data-level CL studies (Platanios et al., 2019; Liu et al., 2020; Zhou et al., 2020) assess data difficulty and model competence to input instances in an easy-to-hard order during training. Utilizing the similar setting can also improve other tasks including RE (Park and Kim, 2021). The task-level CL methods (Guo et al., 2020; Liu et al., 2021a) propose to get non-autoregressive translation models by fine-tuning general translation models with increasingly difficult input format.

3 Method

This section presents the common way to fine-tune PLMs for RE and describes our proposed method Fine-tuning with Prompt Curriculum (FPC).

3.1 Fine-tuning PLMs for RE

A RE dataset can be denoted as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, in which \mathcal{X} is the instance set and \mathcal{Y} is the relation label set. Each instance $x \in \mathcal{X}$ consists of a token sequence $\{w_1, w_2, \dots, w_{|x|}\}$ and the spans of two marked entities. The target is to predict the relation label $y \in \mathcal{Y}$ between the entities.

The token sequence is first converted to the input sequence according to the utilized PLM like $\{[\text{CLS}], w_1, w_2, \dots, w_{|x|}, [\text{SEP}]\}$. Following the general setting (Soares et al., 2019), entity markers are used to index the positions of the entities. We insert special tokens such as "[E]" and "[/E]" into the sequence at the start and end of the entity spans. If the annotation of entity type is provided, type markers can be used by fusing entity type information into the markers.

The PLM encodes the input sequence into the output sequence $\{h_{[\text{CLS}]}, h_1, h_2, \dots, h_{|x|}, h_{[\text{SEP}]}\}$. The output vectors of the two start markers are concatenated to form the relation representation which is fed into a classifier to output the probability distribution over the label set \mathcal{Y} . The fine-tuning process is optimized with a cross-entropy loss denoted as L_{cls} .

3.2 Relation Prompt Learning

The relation prompt learning introduces a cloze-style auxiliary task with the idea of prompt-based fine-tuning, in order to make the model capture the semantics of relation labels.

As shown in Figure 2, we manually design templates with language words and mask tokens. The hard encoding templates are declarative sentences which can "enquire" the relation expressed by the input. There are consecutive mask tokens at the end of the templates which should be filled with words describing the relation expressed by the instance. The same guide words are placed at the start and end of the templates, so we only need to modify the content in the middle. The mentions and types of the entities should be copied to the corresponding positions of [Ent] and [Typ] in the templates. These two designed templates are denoted as "E" and "ET" respectively according to the included entity information.

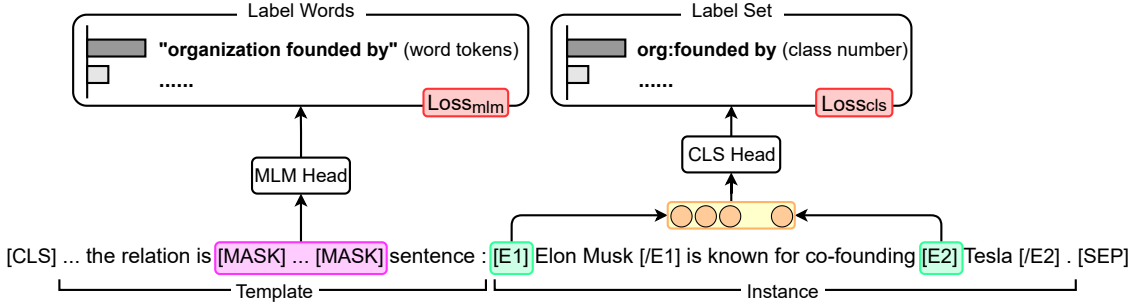
To make the model capture relational semantics, the label words (prediction targets) should be meaningful words describing relations. The words of relation labels are exactly suitable, hence we directly use them with a slight modification to construct the label words. RE datasets generally present relation labels in a hierarchical structure. We remove the punctuations and restore the abbreviations in relation labels and tokenize the labels into token sequences to get the label words. For example, the relation label "org:founded_by" is converted to the token sequence {"organization", "founded", "by"} which is used as the label words. Because relation labels have different lengths and can be tokenized into different number of tokens, we use the same dummy token to pad the label words. Therefore the label words have the same length after tokenizing, which makes the number and positions of the mask tokens fixed in the templates.

Figure 2 illustrates the overview of the relation prompt learning. We insert the template into each instance and choose the corresponding label words in order to bring the cloze-style auxiliary task to the model. We fine-tune the model to classify relation labels and fill the mask tokens with the correct label words at the same time. Through learning to predict the label words, the model can capture the semantics of relation labels and build the connection between the label words and class numbers.

The loss functions of classification L_{cls} and MLM L_{mlm} are applied for the fine-tuning pro-

Template (E): "In this sentence, the relation between [Ent1] and [Ent2] is [MASK] ... [MASK] sentence:"
 Template (ET): "In this sentence, the relation between [Ent1] ([Typ1]) and [Ent2] ([Typ2]) is [MASK] ... [MASK] sentence:"

(a) Prompt Templates



(b) Relation Prompt Learning

Figure 2: (a) shows the manually designed templates. The same guide words "In this sentence," and "sentence:" are added at the start and end of the templates. The mentions and types of the entities need to be copied to the corresponding positions of [Ent] and [Typ]. (b) illustrates the overview of the relation prompt learning.

cess. $Loss_{mlm}$ is defined on the masked positions and other positions do not join in the calculation. We formalize the total loss of fine-tuning as Equation (1) in which α is a hyperparameter to control the weights of the two objectives.

$$Loss_{total} = (1 - \alpha) * Loss_{cls} + \alpha * Loss_{mlm} \quad (1)$$

Compared with other prompt-based fine-tuning methods, our proposed method only needs a little manual labor.

3.3 Prompt Learning Curriculum

It is a common problem for multi-task learning that auxiliary tasks do not always benefit the target task. If the relation prompt learning is directly introduced, the same problem will arise. The reason is that it is difficult for the model to connect classification target with MLM target, therefore the model can not effectively learn the two objectives simultaneously.

The prompt learning curriculum is proposed to address this problem. This task-level curriculum is a fine-tuning procedure which can adapt the model to the multi-task setting with an increasingly hard sub-task. We define an "easy" sub-task in which a part of instances directly shows the prediction targets of the cloze-style auxiliary task.

As shown in Figure 3, all instances are divided into two types denoted as "mask" and "unmask". The "mask" format is the original format described above: consecutive mask tokens are placed at the

end of the template. In the "unmask" format, the mask tokens are replaced with the corresponding label words. Provided the two kinds of instances, the model predicts the label words for the specific positions where may be mask tokens or the prediction targets, therefore the fine-tuning objective is always the same.

Each instance is originally in the "mask" format, which can be converted to the "unmask" format according to a probability, hence it is easy to control the ratio between "mask" and "unmask" instances by adjusting this probability. In our setting, the proportion of "mask" instances P_{mask} gradually increases during fine-tuning, which should be low at the beginning and become 100% before the end. The sub-task gradually becomes "harder" and finally turns into the target task as the number of "unmask" instances decreases, which can adapt the model to the multi-task setting.

Figure 3 illustrates an example of the proposed prompt learning curriculum. Specifically we fix P_{mask} in each fine-tuning epoch, hence the difficulty of the sub-task is fixed in a epoch. P_{mask} is low in the first epoch and gradually increases in the subsequent epochs. Finally all instances are in the "mask" format, which makes the model handle the test scenario.

To some extent, the prompt learning curriculum can transfer the knowledge of "unmask" instances to the model. Through observing and predicting the label words shown in "unmask" instances, the model can know the range of the label words and

"mask" format :
 ... the relation is [MASK] ... [MASK] sentence : ...
 "unmask" format :
 ... the relation is organization founded by sentence : ...

(a) "mask" and "unmask" formats

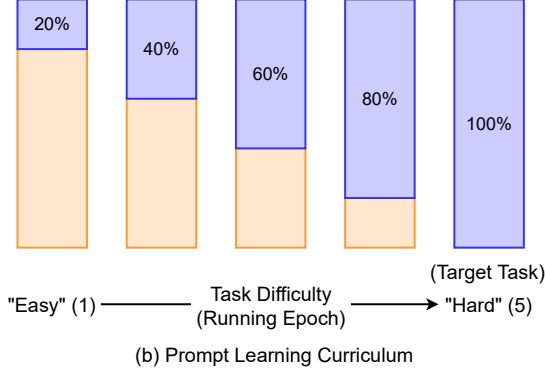


Figure 3: (a) shows the "mask" and "unmask" formats of instances. (b) illustrates an example of the prompt learning curriculum.

easily capture the connection between the label words and class numbers. Therefore our proposed curriculum can improve the performance.

4 Experiments

4.1 Datasets

We have conducted experiments on four widely used RE datasets, including TACRED (Zhang et al., 2017), TACREV (Alt et al., 2020), Re-TACRED (Stoica et al., 2021) and SemEval 2010 Task 8 (SemEval) (Hendrickx et al., 2010). We present more details about the datasets in Table 1 and use micro F_1 scores as the primary metric for evaluation.

TACRED is one of the largest RE datasets. It annotates subject and object entities with their type and contains 42 relations (including "no_relation").

TACREV relabels the incorrect instances in the original TACRED development and test sets, while the training set remains unchanged.

Re-TACRED re-annotates the full dataset of TACRED to rectify mislabeled instances and refines some relation descriptions.

SemEval annotates first and second entities and contains 9 relations with two directions and one special relation "Other". We follow the data split provided by OpenNRE (Han et al., 2019).

4.2 Baseline Models

We compare FPC with the competitive RE models which can be divided into 3 types: the classifica-

Dataset	#Train	#Dev	#Test	#Rel
TACRED	68,124	22,631	15,509	42
TACREV	68,124	22,631	15,509	42
Re-TACRED	58,465	19,584	13,418	40
SemEval	6,507	1,493	2,717	19

Table 1: Statistics of the used datasets.

tion methods, the reformulation methods and the prompt-based fine-tuning methods.

Fine-tuning vanilla PLMs can achieve promising results for RE and we use **RoBERTa_{LARGE}** without adding entity markers as our baseline.

GDPNet (Xue et al., 2021) captures relations of tokens with a latent multi-view graph, which is refined to select vital words for prediction.

SpanBERT (Joshi et al., 2020) extends the MLM pre-training objective to masked contiguous spans with random lengths.

MTB (Soares et al., 2019) is pre-trained on entity linked text, with the new task to decide whether two sampled sentences share the same entities.

KnowBERT (Peters et al., 2019) is pre-trained jointly with an entity linker to incorporate entity embeddings to update word representations.

LUKE (Yamada et al., 2020) treats words and entities as independent tokens and directly models the relations between entities.

TYP Marker (Zhou and Chen, 2021) adopts the specific punctuations and the words of entity types to construct type markers.

RECENT (Lyu and Chen, 2021) exploits entity types to restrict candidate relations and uses a specific classifier for each pair of entity types.

TANL (Paolini et al., 2021) frames RE as a translation task between augmented natural languages and decodes the output text to make predictions.

NLI (Sainz et al., 2021) transforms RE into a textual entailment problem by designing hypotheses based on relational semantic.

PTR (Han et al., 2021) manually designs some essential sub-prompts and composes them into final prompts by applying logic rules.

KnowPrompt (Chen et al., 2022) proposes to inject semantic knowledge for the construction of learnable virtual type words and answer words.

4.3 Implementation Details

We implement FPC based on the vanilla PLM RoBERTa_{LARGE} provided by Transformers (Wolf et al., 2020). We set most hyperparameters following previous works and conduct experiments

Model	PLM Size	Extra Data	TACRED	TACREV	Re-TACRED	SemEval
Classification Methods						
Fine-tuning	RoBERTa _{LARGE}	w/o	68.7	76.0	84.9	87.6
GDPNet	BERT _{LARGE}	w/o	70.5	80.2	-	-
SpanBERT	BERT _{LARGE}	w/o	70.8	78.0	85.3	-
MTB	BERT _{LARGE}	w/	71.5	-	-	89.5
KnowBERT	BERT _{BASE}	w/	71.5	79.3	-	89.1
LUKE	RoBERTa _{LARGE}	w/	72.7	80.6	90.3	-
TYP Marker	RoBERTa _{LARGE}	w/o	74.6	83.2	91.1	-
RECENT	BERT _{LARGE} (multiple)	w/o	75.2	-	-	-
Reformulation Methods						
TANL	T5 _{BASE}	w/o	71.9	-	-	-
NLI	DeBERTa v2 _{XLARGE}	w/	73.9	-	-	-
Prompt-based Fine-tuning Methods						
PTR	RoBERTa _{LARGE}	w/o	72.4	81.4	90.9	89.9
KnowPrompt	RoBERTa _{LARGE}	w/o	72.4	82.4	91.3	90.2
Our Proposed Method						
FPC _E	RoBERTa _{LARGE}	w/o	72.9	82.9	91.3	90.4
FPC _{ET}	RoBERTa _{LARGE}	w/o	76.2	84.9	91.6	\

Table 2: Experimental results of F_1 scores (%) on the test sets of the RE benchmarks and the best results are bold. We report the original or reproduced results from the papers of the baselines and benchmarks. In the "PLM Size" column, we use the frequently-used PLMs to report the PLM configurations of these models for better comparison. In the "Extra Data" column, "w/o" means that only use the data of the benchmarks, while "w/" means that extra data or knowledge bases are utilized. \ marks the unavailable results since entity type information is not provided.

under fully supervised and low-resource settings. AdamW (Loshchilov and Hutter, 2018) is adopted as the optimizer. We conduct all experiments on one NVIDIA Tesla V100 GPU and select the best model checkpoint according to the performance on the development set. For all results, we report the median score of 5 runs with different random seeds. We provide further details of our experiments in Appendix A.

4.4 Results of Fully Supervised RE

Table 2 demonstrates the overall experimental results of our proposed FPC and the compared baselines under fully supervised setting.

The performance of RoBERTa is generally lower than other models. The reason is that simply fine-tuning can not completely cover the knowledge required for RE.

Since the model design of GDPNet and the pre-training objectives of MTB and SpanBERT are really effective, these models can obtain task-specific knowledge for RE and attain higher performance.

However, KnowBERT and LUKE can obviously outperform these models. The reason is that they design specific architectures to integrate entity information from knowledge bases into the models.

Reformulation methods such as TANL and NLI

can obtain promising performance. However, such methods usually need abundant effort for task design and extra usage of time and memory.

KnowPrompt and PTR are able to achieve competitive or higher performance. They can inject relational knowledge into the models by constructing prompts. These prompt-based fine-tuning methods can effectively stimulate the rich knowledge hidden in the PLMs as well.

TYP Marker designs the effective type markers. RECENT builds the restriction between relations and entity types and uses multiple models to handle different pairs of entity types. These models can attain apparent improvements, which illustrates the effectiveness of their designs.

As shown in Figure 2, we design two templates for the relation prompt learning and report the results of FPC using them marked as "E" and "ET" respectively. FPC_E and FPC_{ET} can significantly outperform these compared baselines. FPC_{ET} can achieve the new state-of-the-art results with the more informative template. This demonstrates the effectiveness of our designs: the relation prompt learning makes the model capture the semantics of relation labels and the prompt learning curriculum guides the model to build the connection between the two learning objectives.

Model	TACRED			TACREV			Re-TACRED		
	K=8	K=16	K=32	K=8	K=16	K=32	K=8	K=16	K=32
Fine-tuning	12.2	21.5	28.0	13.5	22.3	28.2	28.5	49.5	56.0
GDPNet	11.8	22.5	28.8	12.3	23.8	29.1	29.0	50.0	56.5
TYP Marker	28.9	32.0	32.4	27.6	31.2	32.0	44.8	54.1	60.0
PTR	28.1	30.7	32.1	28.7	31.4	32.4	51.5	56.2	62.1
KnowPrompt	<u>32.0</u>	35.4	36.5	<u>32.1</u>	<u>33.1</u>	<u>34.7</u>	<u>55.3</u>	63.3	<u>65.0</u>
FPC _{ET}	33.6	<u>34.7</u>	<u>35.8</u>	33.1	34.3	35.5	57.9	<u>60.4</u>	65.3

Table 3: Experimental results of low-resource RE. We sample 5 different data subsets and report the mean score on these data subsets for each result. The best results are bold and the second best results are underlined.

Model	TACRED	TACREV	Re-TACRED	SemEval
showing no entity type words				
ENT Marker	71.4	81.2	90.5	89.8 [†]
FPC _E (TEMP)	72.1	81.9	91.0	90.2
FPC _E (RPL)	72.2	82.0	91.3	90.4
FPC _E (RPL + PLC)	72.9	82.9	91.3	90.4
showing entity type words				
TYP Marker	74.6	83.2	91.1	\
FPC _{ET} (TEMP)	75.3	83.9	91.4	\
FPC _{ET} (RPL)	75.4	84.0	91.6	\
FPC _{ET} (RPL + PLC)	76.2	84.9	91.6	\

Table 4: Experimental results of the ablation study. † marks our reproduced results of the baseline.

4.5 Results of Low-Resource RE

We conduct experiments of low-resource RE following the setting of LM-BFF (Gao et al., 2021; Han et al., 2021; Chen et al., 2022). We randomly sample K training instances and K development instances per class from the original dataset and evaluate the model on the whole test set. In practice K is set to $\{8, 16, 32\}$. We sample 5 different data subsets based on a fixed set of seeds and report the mean score on these data subsets for each result.

The experimental results under low-resource setting are shown in Table 3. TYP marker, PTR and KnowPrompt obtain higher results than other baselines by utilizing entity information. This indicates that entity information is critical for RE, especially under low-resource setting.

FPC can obtain the best results when the number of instances is small ($K=8$) and the competitive or best performance if more instances are provided ($K=16,32$). In practice, we find that the relation prompt learning is the main contributor for the high results, which shows that capturing the semantics of relation labels is effective for low-resource RE. The prompt learning curriculum can improve the results if the amount of instances is more ($K=32$), which indicates that the prompt learning curriculum needs more instances to show the guide effect.

5 Analysis

5.1 Ablation Study

We present a thorough ablation study to show the effects of our designs. FPC is mainly compared with **Ent Marker** and **TYP Marker** (Zhou and Chen, 2021). This work utilizes the specific punctuations as entity markers and further inserts the words of entity type to construct type markers.

Table 4 reports the experimental results of the ablation study, from which we can know that:

The words of entity mentions and types can provide entity information and the model can utilize the clues to make predictions. Hence further showing entity type words can boost the results.

FPC(TEMP): We insert the templates "E" and "ET" into the input to get the results. The evidently improved performance shows that introducing entity information in the templates is more helpful than using the type markers. The model can utilize this kind of relation-oriented knowledge better if it is presented directly and orderly in the templates.

FPC(RPL): We introduce the relation prompt learning based on the templates to attain the results. While the model achieves obviously higher results on Re-TACRED and SemEval, the results of TACRED and TACREV are slightly improved. This is because the mislabeled instances of Re-TACRED

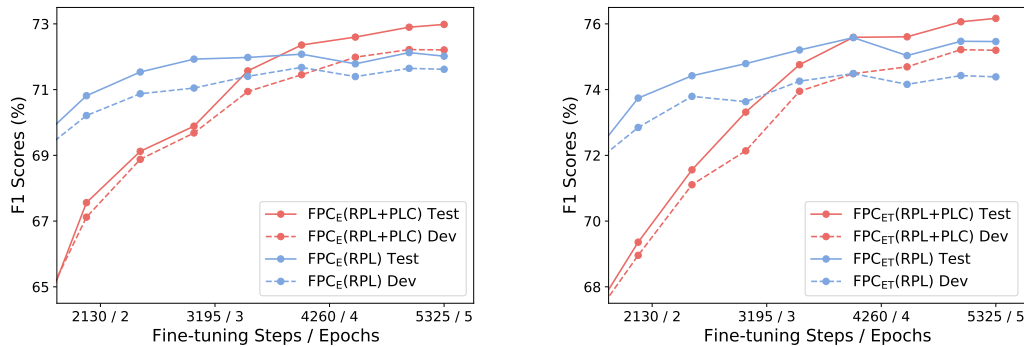


Figure 4: Results of different model checkpoints on TACRED development and test sets.

and SemEval are less and these datasets are easy for our model. When handling the other two hard datasets, the model can not successfully build the connection between the targets of classification and MLM. Therefore the prompt learning curriculum is proposed to improve the performance.

FPC(RPL+PLC): We fine-tune the model according to the prompt learning curriculum to obtain the results. Our model attains remarkable improvement on TACRED and TACREV and similar results on Re-TACRED and SemEval. By learning the sub-task with increasing difficulty, the model can easily connect classification target with MLM target and adapt to the multi-task setting, which is more effective on hard datasets. The superior results show the effectiveness of the prompt learning curriculum.

5.2 Influence of Template

We find that the templates have a great influence on the results. The reason is that they can provide entity information which is crucial for RE. To study the importance of different entity information, we design two new templates shown as below.

Template (S) :

the relation is [MASK] ... [MASK]

Template (T) :

the relation between [Typ1] and [Typ2] is [MASK] ... [MASK]

We conduct experiments of FPC with different templates and the results are shown in Table 5. The model obtains better performance by observing the words of entity mentions and types in the templates and type information can contribute to higher improvement. We argue that entity information can make the model build the restriction between relations and entity types whose effectiveness is shown by RECENT.

Model	TACRED	TACREV	Re-TACRED	SemEval
FPC _S	72.4	82.6	91.0	90.2
FPC _E	72.9	82.9	91.3	90.4
FPC _T	75.4	84.2	91.5	\
FPC _{ET}	76.2	84.9	91.6	\

Table 5: Experimental results of different templates.

5.3 Influence of Curriculum

To study the effect of the prompt learning curriculum, we evaluate different model checkpoints during fine-tuning on TACRED development and test sets. We report the average scores of 10 runs and the results are shown in Figure 4.

We introduce the relation prompt learning to the model and find that the results quickly reach the peaks and then randomly and slightly shake.

We further utilize the prompt learning curriculum to fine-tune the model and find that the model performance is gradually and stably improved after each epoch. Most best results are obtained at the end of fine-tuning and the final results are significantly improved. This indicates that the prompt learning curriculum can help the model to link the objectives of the multi-task setting and make full use of the datasets, hence our model can capture and utilize the semantics of relation labels.

Based on the setting of the relation prompt learning, we propose the prompt learning curriculum which is different from other existing curriculum learning methods. In order to better show the influence of the prompt learning curriculum, we design another curriculum learning method as our baseline to make a comparison.

We propose the increasing α curriculum with the similar idea: we increase the difficulty of the sub-task by changing the weights in the total loss func-

CL method	TACRED	TACREV	Re-TACRED	SemEval
FPC _E (RPL + Curriculum)				
I α C	72.4	82.1	91.3	90.4
PLC	72.9	82.9	91.3	90.4
FPC _{ET} (RPL + Curriculum)				
I α C	75.6	84.2	91.6	\
PLC	76.2	84.9	91.6	\

Table 6: Experimental results of different curriculum learning methods.

tion Equation (1). The weight of $Loss_{mlm}$ should gradually increase during fine-tuning, hence α should be low at the beginning and become the target value before the end.

Specifically we adopt the similar setting: α is fixed in each fine-tuning epoch. α is low in the first epoch, gradually increases in the following epochs and become the target value in the last epoch.

Table 6 shows experimental results of different curriculum learning methods. The increasing α curriculum can help the model to obtain better scores. The improvement of the prompt learning curriculum is higher overall, especially on the two hard datasets TACRED and TACREV. This shows that the prompt learning curriculum is more effective.

6 Conclusion

In this paper, we propose a novel method Fine-tuning with Prompt Curriculum (FPC) for RE. The relation prompt learning introduces the cloze-style auxiliary task, through which the model can capture the semantics of relation labels. The prompt learning curriculum makes the model adapt to the multi-task setting by learning the increasingly difficult sub-task, which makes the model build the connection between the targets of classification and MLM. Extensive experiments have been conducted on four popular RE benchmarks. The results show that FPC achieves the new state-of-the-art performance for fully supervised RE and the competitive or best performance for low-resource RE.

Acknowledgements

This work was supported by National Key Research and Development Program of China (Grant No. 2020YFC0833402), National Natural Science Foundation of China (Grant Nos. 61976021, U1811262), and Beijing Academy of Artificial Intelligence (BAAI).

References

- Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Taced revisited: A thorough evaluation of the taced relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.
- Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 615–620.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, pages 2778–2788.
- Joe Davison, Joshua Feldman, and Alexander M Rush. 2019. Commonsense knowledge mining from pre-trained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178.
- 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.
- 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*.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

- and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3816–3830.
- Junliang Guo, Xu Tan, Linli Xu, Tao Qin, Enhong Chen, and Tie-Yan Liu. 2020. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7839–7846.
- Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. Opennre: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 169–174.
- 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*.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. 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.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. 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.
- 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.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Xiang Lisa Li and Percy Liang. 2021. 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.
- Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350.
- Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2021a. Task-level curriculum learning for non-autoregressive neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3861–3867.
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.
- Xuebo Liu, Houtim Lai, Derek F Wong, and Lidia S Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Shengfei Lyu and Huanhuan Chen. 2021. Relation classification with entity type restriction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 390–395.
- 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.
- Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.
- Seongsik Park and Harksoo Kim. 2021. Improving sentence-level relation extraction through curriculum learning. *arXiv preprint arXiv:2107.09332*.
- Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672.

- Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.
- Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. 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 1162–1172.
- Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.
- Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212.
- Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578.
- Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.
- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.
- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.
- George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13843–13850.
- Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.
- Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2021a. Zero-shot information extraction as a unified text-to-triple translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1225–1238.
- Xin Wang, Yudong Chen, and Wenwu Zhu. 2021b. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.
- Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. 2021. Gdpnet: Refining latent multi-view graph for relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14194–14202.
- Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.
- Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215.
- 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.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *arXiv preprint arXiv:2102.01373*.

Yikai Zhou, Baosong Yang, Derek F Wong, Yu Wan, and Lidia S Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944.

A Further Implementation Details

This section presents more details about the fine-tuning procedures and hyperparameters. We report the used settings which result in the overall best performance.

We use the same punctuations "@" and "#" as entity markers following (Zhou and Chen, 2021). We warm up the learning rate over the first 10% steps and then linearly decay it. We set the weight decay to $1e - 5$ and clip gradients if their norms exceed 1.0. The maximum sequence length is set to 512 and none of the instances exceed it. Table 7 shows the other used hyperparameters.

Hyperparameter	Value
Fully Supervised RE	
learning rate	3e-5
fine-tuning epochs	5
curriculum epochs	5
batch size	32
Low-Resource RE	
learning rate	2e-5
fine-tuning epochs	30
curriculum epochs	20
batch size	16 ($K=8$) or 32 ($K=16,32$)

Table 7: The settings of the other used hyperparameters.

For the relation prompt learning, we set α in Equation (1) to 0.4 on TACRED, TACREV and Re-TACRED and 0.3 on SemEval under both fully supervised and low-resource settings.

For the prompt learning curriculum, the proportion of "mask" instances P_{mask} is controlled by the number of fine-tuning epochs. P_{mask} linearly increases during fine-tuning and finally become 100%. For the increasing α curriculum, α in Equation (1) linearly increases during fine-tuning and

we use the number of fine-tuning epochs to adjust α as well. Table 8 shows the detailed settings of the prompt learning curriculum and the increasing α curriculum.

The designed label words of the used datasets are shown in Table 9 and Table 10. Specifically we use the punctuation "-" to pad the label words and make them have the same length after tokenizing.

Epoch	1	2	3	4	5
$P_{mask}(PLC)$	20%	40%	60%	80%	100%
$\alpha(I\alpha C)$	0.08	0.16	0.24	0.32	0.40

Table 8: The settings of the prompt learning curriculum and the increasing α curriculum. For the increasing α curriculum, the values in the row of $\alpha(I\alpha C)$ should be changed if α is set to other values.

Relation Label	Label Words
Other	[other, relations]
Component-Whole(e2,e1)	[whole, component]
Instrument-Agency(e2,e1)	[agency, instrument]
Member-Collection(e1,e2)	[member, collection]
Cause-Effect(e2,e1)	[effect, cause]
Entity-Destination(e1,e2)	[entity, destination]
Content-Container(e1,e2)	[content, container]
Message-Topic(e1,e2)	[message, topic]
Product-Producer(e2,e1)	[producer, product]
Member-Collection(e2,e1)	[collection, member]
Entity-Origin(e1,e2)	[entity, origin]
Cause-Effect(e1,e2)	[cause, effect]
Component-Whole(e1,e2)	[component, whole]
Message-Topic(e2,e1)	[topic, message]
Product-Producer(e1,e2)	[product, producer]
Entity-Origin(e2,e1)	[origin, entity]
Content-Container(e2,e1)	[container, content]
Instrument-Agency(e1,e2)	[instrument, agency]
Entity-Destination(e2,e1)	[destination, entity]

Table 9: The designed label words for SemEval.

Relation Label	Label Words
no_relation	[no, relation, -, -, -, -]
org:alternate_names	[organization, alternate, names, -, -, -]
org:city_of_headquarters	[organization, city, of, headquarters, -, -]
org:country_of_headquarters	[organization, country, of, headquarters, -, -]
org:dissolved	[organization, date, of, dissolution, -, -]
org:founded	[organization, date, of, founding, -, -]
org:founded_by	[organization, founded, by, -, -, -]
org:member_of	[organization, member, of, -, -, -]
org:members	[organization, members, -, -, -, -]
org:number_of_employees/members	[organization, number, of, employees, members, -]
org:parents	[organization, parents, -, -, -, -]
org:political/religious_affiliation	[organization, political, religious, affiliation, -, -]
org:shareholders	[organization, shareholders, -, -, -, -]
org:stateorprovince_of_headquarters	[organization, state, or, province, of, headquarters]
org:subsidiaries	[organization, subsidiaries, -, -, -, -]
org:top_members/employees	[organization, top, members, employees, -, -]
org:website	[organization, website, -, -, -, -]
per:age	[person, age, -, -, -, -]
per:alternate_names	[person, alternate, names, -, -, -]
per:cause_of_death	[person, cause, of, death, -, -]
per:charges	[person, charges, -, -, -, -]
per:children	[person, children, -, -, -, -]
per:cities_of_residence	[person, city, of, residence, -, -]
per:city_of_birth	[person, city, of, birth, -, -]
per:city_of_death	[person, city, of, death, -, -]
per:countries_of_residence	[person, country, of, residence, -, -]
per:country_of_birth	[person, country, of, birth, -, -]
per:country_of_death	[person, country, of, death, -, -]
per:date_of_birth	[person, date, of, birth, -, -]
per:date_of_death	[person, date, of, death, -, -]
per:employee_of	[person, employee, or, member, of, -]
per:origin	[person, origin, -, -, -, -]
per:other_family	[person, other, family, -, -, -, -]
per:parents	[person, parents, -, -, -, -]
per:religion	[person, religion, -, -, -, -]
per:schools_attended	[person, schools, attended, -, -, -, -]
per:siblings	[person, siblings, -, -, -, -]
per:spouse	[person, spouse, -, -, -, -]
per:stateorprovince_of_birth	[person, state, or, province, of, birth]
per:stateorprovince_of_death	[person, state, or, province, of, death]
per:stateorprovinces_of_residence	[person, state, or, province, of, residence]
per:title	[person, title, -, -, -, -]
org:city_of_branch	[organization, city, of, branch, -, -]
org:country_of_branch	[organization, country, of, branch, -, -]
org:stateorprovince_of_branch	[organization, state, or, province, of, branch]
per:identity	[person, identity, -, -, -, -]

Table 10: The designed label words for TACRED, TACREV and Re-TACRED.