MatchPrompt: Prompt-based Open Relation Extraction with Semantic Consistency Guided Clustering

Jiaxin Wang[1,2] **, Lingling Zhang**[1,2*]**, Jun Liu**[1,2] **, Xi Liang**[1,2] **, Yujie Zhong**[1,2] **, Yaqiang Wu**[2,3,]
[1]Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering,
School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, China
[3]Lenovo Research, Beijing, China
jiaxinwangg@outlook.com, {zhanglling,liukeen}@xjtu.edu.cn
{lliangxii,yjzhong6}@gmail.com, wuyqe@lenovo.com

Abstract

Relation clustering is a general approach for open relation extraction (OpenRE). Current methods have two major problems. One is that their good performance relies on large amounts of labeled and pre-defined relational instances for pre-training, which are costly to acquire in reality. The other is that they only focus on learning a high-dimensional metric space to measure the similarity of novel relations and ignore the specific relational representations of clusters. In this work, we propose a new prompt-based framework named MatchPrompt, which can realize OpenRE with efficient knowledge transfer from only a few pre-defined relational instances as well as mine the specific meanings for cluster interpretability. To our best knowledge, we are the first to introduce a prompt-based framework for unlabeled clustering. Experimental results on different datasets show that MatchPrompt achieves the new SOTA results for OpenRE.

1 Introduction

Relation Extraction (RE) is one of the most essential technology for knowledge graph construction (Church and Bian, 2021; Mirtaheri, 2021). It aims to detect the relation between two entities in a sentence. For example, when given the entity pair *(San Diego, the United States)* in the sentence *San Diego is located in the United States*, a relation extraction model should predict the relation *located in* between these two entities. Despite traditional RE models having achieved great success (Li et al., 2021; Ye et al., 2022), they are designed based on pre-defined relations and are incapable of extracting novel relations in the real world. Therefore, Open Relation Extraction (OpenRE) has been proposed to extract relations without pre-defined types from the open-domain corpus. Methods for OpenRE can be divided into two categories based on the form of extracted novel relations. The former is called open information extraction (OIE) (Yates et al., 2007; Wang et al., 2021), which requires the relational information to appear explicitly and extracts relation-related phrases from sentences based on structure and syntactic analysis. The latter is unsupervised relation discovery, which generally clusters sentences into collections based on their semantic similarities, and regards each cluster as a relation. Considering that relations are usually implicit in sentences and it is difficult to align different phrases that have the same meaning, thus much attention has been paid to the latter category.

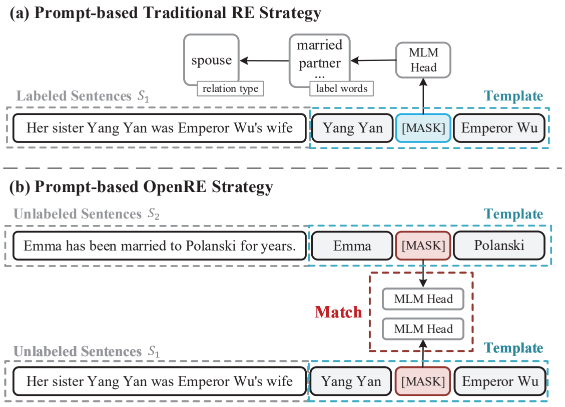

(a) Prompt-based Traditional RE Strategy
spouse
relation type
married
partner
...
label words
MLM
Head
Labeled Sentences S_1
Template
Her sister Yang Yan was Emperor Wu's wife
Yang Yan
[MASK]
Emperor Wu
(b) Prompt-based OpenRE Strategy
Unlabeled Sentences S_2
Template
Emma has been married to Polanski for years.
Emma
[MASK]
Polanski
Match
MLM Head
MLM Head
Unlabeled Sentences S_1
Template
Her sister Yang Yan was Emperor Wu's wife
Yang Yan
[MASK]
Emperor Wu

Figure 1: Comparison of prompt-based traditional RE strategy and OpenRE strategy.

Specifically, Hu et al. (2020) proposes an adaptive clustering method for novel relation discovery. Simon et al. (2019) introduces a discriminative training scheme to extract novel relations. However, these methods (Yao et al., 2011; Simon et al., 2019; ElSahar et al., 2017; Hu et al., 2020; Tran et al., 2020) are totally trained in unsupervised settings that ignore the transferable knowledge in existing pre-defined relations and have limited performance (Wu et al., 2019). To this end, Wu et al. (2019) and Zhao et al. (2021) learn transferable

[*] Corresponding author

knowledge from pre-defined relations for novel relation clustering and achieve relatively good performance. **But they rely on large amounts of labeled and pre-defined relational instances for pre-training which are costly to acquire in reality.** In addition, all these clustering-based methods have a major problem of poor interpretability, that is, **they only focus on learning a high-dimensional metric space to measure the similarity of novel relations and do not care about relation specifics** (e.g., words or phrases related to the relation). Actually, we prefer clusters to have specific relational meanings to satisfy the purpose of RE better. Therefore, it is expected to realize OpenRE with efficient knowledge transfer from a few pre-defined relational instances and to mine the specific meanings of clusters as much as possible.

Prompt-based learning (Schick and Schütze, 2021; Shin et al., 2020; Hu et al., 2021) has been proposed as a new paradigm with the potential to solve the above problems. As shown in Fig. 1(a), for traditional RE, a typical prompt-based method first wraps sentence S_1 into a template (e.g., "$< S_1 >$Yang Yan `[MASK]` Emperor Wu") and construct a set of label words (e.g., "married" and "partner"). Then it utilizes Pre-trained Language Models (PLMs) to predict the probability of each word filling the `[MASK]` token and maps the word with maximum probability back to relation categories. This process can achieve good performance with a few labeled data (Chen et al., 2022) and provides the natural conditions for transforming high-dimensional features into words with specific meanings. This motivated us to formalize OpenRE as a form of prompt-based learning.

In this paper, we propose a prompt-based framework, **MatchPrompt**, for OpenRE. Different from the prompt-based traditional RE paradigm, there are no pre-defined relations in the open domain, thus we cannot construct label words set for OpenRE directly. Inspired by the underlying assumptions of OpenRE are that relational instances can form clusters and instances from the same cluster should have similar semantics and share the same relation label (Liu et al., 2021a), as illustrated in Fig. 1(b), we devise a matching strategy to drive prompt learning in the open domain, i.e., if different entity pairs from two sentences express the same relation, the semantic features of `[MASK]` in their templates should be matching. Additionally, three modules are designed to enable MatchPrompt generate efficient relational representations during training. First, the pre-training and knowledge transfer module leverages a few pre-defined relational instances to activate templates to produce relation-aware representations in `[MASK]` tokens. Second, the semantic consistency guided clustering module selects reliable pseudo-labels for clustering loss computation of unlabeled instances. Third, the semantic consistency regularization module encourages MatchPrompt to maintain the robustness of feature predictions in `[MASK]` tokens. On the one hand, MatchPrompt can achieve better knowledge transfer with only a few pre-defined relational instances for novel relation clustering, which reduces the cost of pre-training with large amounts of pre-defined and labeled data. On the other hand, it obtains representations of novel relations by `[MASK]` tokens, which are easily converted into the specific meanings of relations, making clusters more comprehensible.

To summarize, the main contributions of our work are as follows: (1) We develop a novel prompt-based framework MatchPrompt which enables the model to learn clustering novel relational instances. To our best knowledge, we are the first to introduce a prompt-based framework for unlabeled clustering. (2) The proposed model can generate efficient representations for instances in the open domain. These representations can also provide a basis for judging the specific relational words of clusters. (3) Experimental results show that MatchPrompt has a competitive performance with only 10% of pre-defined relational instances for model pre-training compared to the current SOTA method, which requires 100% of these data for pre-training.

2 Background

Define a relational example as a tuple: $\langle \boldsymbol{w}, h, t \rangle$, where $\boldsymbol{w} = \{w_1, w_2, \cdots, w_n\}$ is a relational sentence with n tokens, $h = \{[w_a, w_b] | 1 \leq a \leq b \leq n\}$ and $t = \{[w_c, w_d] | 1 \leq c \leq d \leq n\}$ indicate the head entity and tail entity in $\boldsymbol{w}$, respectively. For one given $(\boldsymbol{w}, h, t)$, traditional RE defaults the relation between h and t as one of the pre-defined relations, so it predicts the type of relation label y between this entity pair based on the sentence information of $\boldsymbol{w}$. Given a pre-trained language model $\mathcal{M}$, it first converts the input sequence with special tokens $\{\texttt{[CLS]}, w_1, \cdots, w_n, \texttt{[SEP]}\}$, and then encodes the sequence into hidden vectors $\mathcal{A} = \{\boldsymbol{a}_{\texttt{[CLS]}}, \boldsymbol{a}_{w_1}, \cdots, \boldsymbol{a}_{w_n}, \boldsymbol{a}_{\texttt{[SEP]}}\}$. Con-

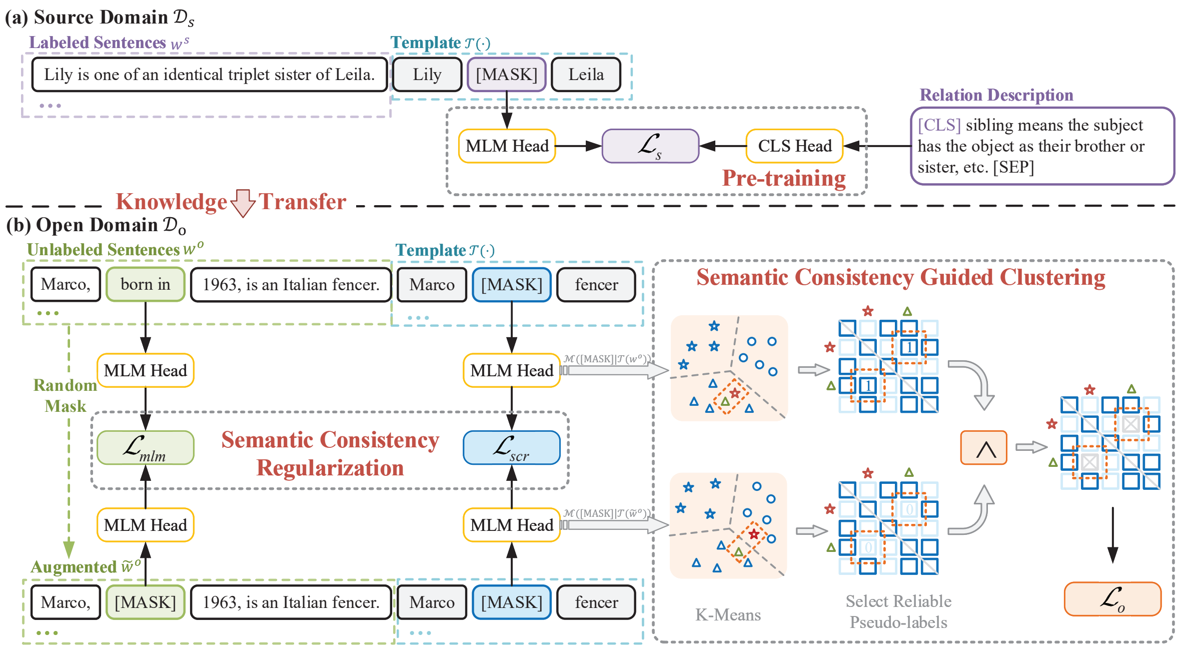

(a) Source Domain $\mathcal{D}_s$
Labeled Sentences w^s
Template $\mathcal{T}(\cdot)$
Lily is one of an identical triplet sister of Leila.
Lily
[MASK]
Leila
MLM Head
$\mathcal{L}_s$
CLS Head
Pre-training
Relation Description
[CLS] sibling means the subject has the object as their brother or sister, etc. [SEP]
Knowledge Transfer
(b) Open Domain $\mathcal{D}_o$
Unlabeled Sentences w^o
Template $\mathcal{T}(\cdot)$
Marco,
born in
1963, is an Italian fencer.
Marco
[MASK]
fencer
Random Mask
MLM Head
$\mathcal{L}_{mlm}$
Semantic Consistency Regularization
$\mathcal{L}_{scr}$
$\mathcal{M}([\text{MASK}]|\mathcal{T}(w^o))$
$\mathcal{M}([\text{MASK}]|\mathcal{T}(\tilde{w}^o))$
Augmented $\tilde{w}^o$
Marco,
[MASK]
1963, is an Italian fencer.
Marco
[MASK]
fencer
Semantic Consistency Guided Clustering
K-Means
Select Reliable Pseudo-labels
$\wedge$
$\mathcal{L}_o$

Figure 2: Illustration of our MatchPrompt framework.

ventional fine-tuning methods leverage the corresponding embedding of h and t in $\mathcal{A}$ to achieve relation classification. For prompt-based methods, they usually design a task-specific template $\mathcal{T}(\cdot)$ and a label words set $\mathcal{V}$ to coax $\mathcal{M}$ into producing a textual output related to the relation class. For example, they construct a label words set $\mathcal{V}$ and a generic template $\mathcal{T}(\cdot)$ = "h [MASK] t". Then, they warp $\boldsymbol{w}$ into:

$$\mathcal{T}(\boldsymbol{w})=\texttt{[CLS]}\,\boldsymbol{w}\,\texttt{[SEP]}\circ h\;\texttt{[MASK]}\;t\;\texttt{[SEP]}, \tag{1}$$

where $\circ$ is the string concatenation operation. When fed $\mathcal{T}(\boldsymbol{w})$ into $\mathcal{M}$, it will produce the hidden vector $\boldsymbol{a}_{\texttt{[MASK]}}$ at the [MASK] position. The probability $P_{\mathcal{M}}(\texttt{[MASK]}= v|\boldsymbol{w}_{prompt})$ for each label word $v \in \mathcal{V}$ that fills the mask position can reflect the probability of label y as:

$$P_{\mathcal{M}}(y|\boldsymbol{w}) = g(P_{\mathcal{M}}(\texttt{[MASK]} = v|\mathcal{T}(\boldsymbol{w}))|v\in\mathcal{V}_y), \tag{2}$$

where $g(\cdot)$ is an aggregation function and $\mathcal{V}_y \in \mathcal{V}$ is the label words set of y.

3 Methodology

3.1 Problem Definition

In contrast to traditional RE, OpenRE has no predefined relations and is commonly conceived as a clustering task that determines whether two relational examples refer to the same relation. Specifically, let $\mathcal{D}_o = \{\langle \boldsymbol{w}_i^o, h_i, t_i\rangle\}_{i=1}^N$ denotes the unlabeled open domain dataset with N novel relational instances. Our goal is to train $\mathcal{M}$ to cluster the instances with the same relations in $\mathcal{D}_o$ into a number of K classes automatically. Following Wu et al. (2019) and Zhao et al. (2021), we assume K is the prior knowledge. Also, we introduce a labeled source domain dataset $\mathcal{D}_s = \{(\langle \boldsymbol{w}_j^s, h_j, t_j\rangle, d_j, y_j), y_j \in Y^l\}_{j=1}^M$ with the predefined relation set Y^l for pre-training, where M and y_j represent the number of the labeled instances and the relation label index for j-th instance, respectively. d_j is the corresponding relation description of y_j. Note that the relation type spaces of $\mathcal{D}_o$ and $\mathcal{D}_s$ are disjoint.

3.2 Overview

As illustrated in Fig. 2, our MatchPrompt includes three components: (1) **Pre-training and Knowledge Transfer.** We pre-train $\mathcal{M}$ with the predefined relational instances in $\mathcal{D}_s$ to activate templates to produce relation-aware representations in [MASK] tokens. (2) **Semantic Consistency Guided Clustering.** We select reliable pseudo-labels in this component for clustering loss computation of unlabeled instances. (3) **Semantic Consistency Regularization.** We design a semantic consistency regularization loss to encourage $\mathcal{M}$ to maintain the robustness of representations produced by [MASK] tokens.

3.3 Pre-training and Knowledge Transfer

Incorporating prior knowledge from pre-defined relations can help the model to obtain a good initialization for clustering those novel relations. For instance, the pre-defined relation "Birthplace" will help the model to learn the novel relation "Deathplace" for their similar potential sentence structures and entity types. Therefore, to encourage templates to produce accurate relation-aware representations at `[MASK]` tokens, we pre-train it by using source domain dataset $\mathcal{D}_s$, that is, make the representation of `[MASK]` tokens closer to the representation of their relation descriptions. Specifically, we first wrap the sentence $\boldsymbol{w}_j$ with a template $\mathcal{T}(\cdot) =$ "h_j `[MASK]` t_j" and take the hidden state of the `[MASK]` token as the initial relational representation between entities h_j and t_j, denoted as $\boldsymbol{v}_j^s = \mathcal{M}(\texttt{[MASK]}|\mathcal{T}(\boldsymbol{w}_j^s))$. Then, we feed the relation description d_j into $\mathcal{M}$ and take the hidden state of the `[CLS]` token as the sentence-level representation (Chen and Li, 2021) for a relation, denoted as $\boldsymbol{v}_j^d = \mathcal{M}(\texttt{[CLS]}|d_j)$. We train $\mathcal{M}$ so that the semantics between $\boldsymbol{v}_j^s$ and $\boldsymbol{v}_j^d$ can be aligned. The objective function is:

$$\mathcal{L}_s = \frac{1}{M}\sum_{i,j=1}^{M} \max(0, \underset{y_i=y_j}{D}(\boldsymbol{v}_j^s, \boldsymbol{v}_i^d) - \max(\underset{y_i \neq y_j}{D}(\boldsymbol{v}_j^s, \boldsymbol{v}_i^d)) + \Delta_1), \tag{3}$$

where $y_i = y_j$ and $y_i \neq y_j$ indicate the positive pair and the negative pair, respectively. And $\Delta_1 > 0$ is a margin hyper-parameter to be tuned. The function $D(\cdot,\cdot)$ is the KL-divergence $D(\boldsymbol{v}_j^s, \boldsymbol{v}_i^d) = \sum v_j^s log \frac{v_j^s}{v_i^d}$ that measures the similarity of semantic features between $\boldsymbol{v}_j^s$ and $\boldsymbol{v}_j^d$.

3.4 Semantic Consistency Guided Clustering

After pre-training, $\mathcal{M}$ is used for the unsupervised clustering. we obtain $\boldsymbol{v}_i^o = \mathcal{M}(\texttt{[MASK]}|\mathcal{T}(\boldsymbol{w}_i^o))$, the embedding in `[MASK]` tokens, as the relation representations of the sentence $\boldsymbol{w}_i^o$ in $\mathcal{D}_o$, and utilize the K-Means algorithm (Arthur and Vassilvitskii, 2007) to generate the pseudo-label p_i for $\boldsymbol{v}_i^o$. As the basic assumptions of OpenRE expect the semantics of two samples in the same cluster (positive pair) to be similar while in different clusters (negative pair) be different, we set a pairwise pseudo-label (Zhao et al., 2021) as follows:

$$b_{ij} = \begin{cases} 1, \text{if } p_i = p_j, \\ 0, \text{if } p_i \neq p_j. \end{cases} \tag{4}$$

where $b_{ij} = 1$ and $b_{ij} = 0$ indicate positive pairs from the same cluster and negative pairs from different clusters, respectively. Apparently, we can use b_{ij} to train $\mathcal{M}$ to pull the semantic features of relational examples in a positive pair closer, while pushing those features in a negative pair away. However, the K-Means algorithm is sensitive to the representation of instances (Vijayaraghavan et al., 2017), which means that if the representations produced by $\mathcal{M}$ are inaccurate, it is easy to introduce many incorrect pseudo-labels, which in turn affects the quality of b_{ij}. Moreover, these incorrect pseudo-labels not only fail to provide useful information during training but also make $\mathcal{M}$ confident in the generated inaccurate relational representations, thereby causing error accumulation.

To alleviate this issue, we propose a semantic consistency guided clustering method. The core idea is that if $\mathcal{M}$ can produce an accurate and robust representation for a relational instance, then it can produce semantically consistency representation for the augmented sentence of this instance with high probability, and vice versa. This semantic consistency principle can help reduce incorrect pairwise pseudo-labels. Specifically, the details of our semantic consistency guided clustering method consist of three stages. First, for any $\boldsymbol{w}_i^o$ in $\mathcal{D}_o$, we obtain the augmented sentence $\widetilde{\boldsymbol{w}}_i^o$ by randomly replacing 15% of the tokens other than h_i and t_i in $\boldsymbol{w}_i$ with the special `[MASK]` token and warp $\widetilde{\boldsymbol{w}}_i^o$ with the template as before. Second, we use $\mathcal{M}$ to get the new relational representation $\mathcal{M}(\texttt{[MASK]}|\mathcal{T}(\widetilde{\boldsymbol{w}}_i^o))$ for $\widetilde{\boldsymbol{w}}_i^o$ as well as generate the new pseudo-label $\widetilde{p}_i$. Third, we recompute pairwise pseudo-label $\widetilde{b}_{ij}$ and rewrite Eq. 4 as:

$$\widetilde{b}_{ij} = \begin{cases} 1, \text{if } p_i = p_j \wedge \widetilde{p}_i = \widetilde{p}_j, \\ 0, \text{if } p_i \neq p_j \wedge \widetilde{p}_i \neq \widetilde{p}_j. \end{cases} \tag{5}$$

Then, the pairwise comparison loss is defined as follows:

$$\mathcal{L}_o = \begin{cases} D(\boldsymbol{v}_i^o, \boldsymbol{v}_j^o), & \text{if } \widetilde{b}_{ij} = 1, \\ \max(0, \Delta_2 - D(\boldsymbol{v}_i^o, \boldsymbol{v}_j^o)), & \text{if } \widetilde{b}_{ij} = 0. \end{cases} \tag{6}$$

where $\Delta_2 > 0$ is a margin hyper-parameter to be tuned and $D(\cdot,\cdot)$ is the KL-divergence for evaluating semantic similarity of pairwise examples. $\mathcal{M}$ can trained by making the pairwise semantic features to be similar and different in the $\widetilde{b}_{ij} = 1$ and $\widetilde{b}_{ij} = 0$ conditions, respectively.

3.5 Semantic Consistency Regularization

There are two reasons for us to introduce a semantic consistency regularization component for our framework. One is that even if the relational examples in $\mathcal{D}_o$ are unlabeled, they still contain potential structure information about relations. This can help $\mathcal{M}$ learn semantic representations with stronger generalization ability. The other is that pairwise pseudo-labels for these unlabeled examples are subject to change on the fly during training. In fact, for a sentence $\boldsymbol{w}_i^o$ and the augmented sentence $\widetilde{\boldsymbol{w}}_i^o$ generated by the random mask, if we do not enforce their semantic consistency, there will be semantic drift during training. Specifically, we propose a semantic consistency regularization loss at `[MASK]` token in template $\mathcal{T}(\cdot) =$ "h `[MASK]` t" to minimize the difference between relation representations produced by $\mathcal{T}(\boldsymbol{w}_i^o)$ and $\mathcal{T}(\widetilde{\boldsymbol{w}}_i^o)$, which is computed by the KL-divergence function $D(\cdot,\cdot)$:

$$\mathcal{L}_{scr} = \frac{1}{M}\sum_{i=1}^{M} D\big(\mathcal{M}\left(\texttt{[MASK]}|\mathcal{T}(\boldsymbol{w}_i^o)\right), \\ \mathcal{M}\left(\texttt{[MASK]}|\mathcal{T}(\widetilde{\boldsymbol{w}}_i^o)\right)\big). \tag{7}$$

In addition, we also use the masked language model loss $\mathcal{M}_{mlm}$ (Chen et al., 2021) to guarantee the consistency between original sentence $\boldsymbol{w}_i$ and augmented sentence $\widetilde{\boldsymbol{w}}_i^o$ as:

$$\mathcal{L}_{mlm} = \frac{1}{M}\sum_{i=1}^{M}\sum_{w_x \in W_i} \mathcal{M}_{mlm}\big((w_x|\boldsymbol{w}_i^o), \\ (w_x = \texttt{[MASK]}|\widetilde{\boldsymbol{w}}_i^o)\big), \tag{8}$$

where W_i is the random mask tokens set for $\boldsymbol{w}_i^o$.

3.6 Training Details

Our MatchPrompt has a two-stage optimization procedure. When training epoch $T \le \mu$, it is in the pre-training and knowledge transfer stage and is optimized with only $\mathcal{L}_s$ as:

$$\mathcal{L} = \mathcal{L}_s, \tag{9}$$

when training epoch $T > \mu$, the pre-training stage ends and MatchPrompt is optimized by all loss functions as:

$$\mathcal{L} = \mathcal{L}_o + \lambda_s \mathcal{L}_s + \lambda_{SC}(\mathcal{L}_{scr} + \mathcal{L}_{mlm}), \tag{10}$$

where λ_s and λ_{SC} control the trade-off of each objective function and μ is the epoch number of pre-training.

Dataset	Source Doamin	Open Domain
	#Rel (train_n/test_n)	#Rel (train_n/test_n)
FewRel	64 (38400/6400)	16 (9600/1600)
TACRED	31 (9865/1644)	10 (1038/174)

Table 1: Statistical and split of two datasets."Rel" refers to relation types, "train_n" refers to number of training instances and "test_n" refers to number of test instances.

4 Experiments

4.1 Experimental Settings

Datasets. As shown in Table 1, our experiments are conducted on two public datasets: FewRel[1] (Han et al., 2018) and TACRED[2] (Zhang et al., 2017). The details of datasets are in Appendix A.1.

Baselines and Evaluation metrics. Six SOTA OpenRE models are categorized into two groups: (1)Three models without knowledge transfer: RW-HAC (ElSahar et al., 2017), Etype+ (Tran et al., 2020), and SelfORE (Hu et al., 2020); (2) Three models with knowledge transfer: RSN (Wu et al., 2019), RSN-BERT (Wu et al., 2019), and the current SOTA method RoCORE (Zhao et al., 2021). Following these models, we use B^3 (Bagga and Baldwin, 1998), V-measure (Rosenberg and Hirschberg, 2007) and ARI (Hubert and Arabie, 1985) to measure the precision and recall for clustering, homogeneity and completeness of clusters, and the agreement between clusters and true distributions, respectively. The details of the baseline models are in Appendix A.2.

Implementations. For fair comparisons, our MatchPrompt utilizes the pre-trained language model BERT-base (Devlin et al., 2019) as previous baseline models. More implementation details please refer to Appendix A.3.

4.2 Overall Results

Table 2 represents comparative results on two datasets. We can make the following observations: **(1) Our MatchPrompt outperforms previous OpenRE models and achieves new state-of-the-art results in terms of F$_1$ scores.** When using full pre-defined relational instances for knowledge transfer, our model improves the SOTA B^3 F$_1$, V-

[1] The relation descriptions of FewRel are in https://github.com/thunlp/FewRel/fewrel.pdf

[2] The relation descriptions of TACRED are in https://nlp.stanford.edu/projects/tacred

Source percentage	Model	FewRel							TACRED						
		B^3			V-Measure			ARI	B^3			V-Measure			ARI
		Pre.	Rec.	F_1	Hom.	Comp.	F_1		Pre.	Rec.	F_1	Hom.	Comp.	F_1	
0%	RW-HAC*	25.6	49.2	33.7	39.1	48.5	43.3	25.0	42.6	63.3	50.9	46.9	59.7	52.6	28.1
	EType+*	23.8	48.5	31.9	36.4	46.3	40.8	24.9	30.2	80.3	43.9	26.0	60.7	36.4	14.3
	SelfORE*	67.2	68.5	67.8	77.9	78.8	78.3	64.7	57.6	51.0	54.1	63.0	60.8	61.9	44.7
	RSN	9.8	25.6	14.2	9.6	14.4	11.5	4.7	16.3	17.8	17.0	7.1	7.0	7.1	2.5
	RSN-BERT	30.3	53.2	38.6	43.4	54.3	48.3	31.3	24.0	32.2	27.5	21.4	26.3	23.6	14.2
	RoCORE	69.5	71.2	70.2	78.3	79.8	78.8	65.3	75.3	83.8	80.2	72.0	86.4	79.3	72.5
	MatchPrompt	**71.6**	**73.0**	**72.3**	**81.2**	**83.1**	**82.2**	**66.5**	**80.1**	**85.2**	**83.0**	**82.7**	**86.5**	**84.5**	**75.3**
1%	RSN	34.7	31.7	33.2	47.2	47.6	47.4	22.8	19.0	24.8	21.5	19.5	17.6	18.5	11.8
	RSN-BERT	28.4	52.5	36.9	39.7	51.6	44.9	27.3	35.9	30.2	32.7	33.7	31.2	32.4	28.7
	RoCORE	44.5	74.0	55.6	63.0	76.5	69.1	47.2	66.6	67.3	67.0	68.3	68.8	68.6	52.6
	MatchPrompt	**71.8**	**82.3**	**75.9**	**80.7**	**85.3**	**83.0**	**66.7**	**81.9**	**85.1**	**83.5**	**83.6**	**86.4**	**84.9**	**79.1**
5%	RSN	41.2	52.0	46.0	53.9	61.2	57.3	30.9	29.1	30.1	29.6	27.0	27.7	27.3	20.5
	RSN-BERT	39.3	62.6	48.3	52.7	61.6	56.8	42.4	32.7	51.9	40.1	28.2	41.8	33.6	30.2
	RoCORE	58.7	68.9	63.4	74.4	79.9	77.1	57.1	76.2	76.4	76.3	79.8	77.6	78.6	65.1
	MatchPrompt	**73.1**	**82.6**	**77.6**	**82.1**	**86.6**	**84.3**	**68.4**	**87.2**	**85.4**	**86.3**	**88.6**	**88.2**	**88.9**	**81.6**
10%	RSN	42.4	50.6	46.1	54.8	60.1	57.3	30.7	34.7	30.4	32.4	33.7	33.8	33.7	21.5
	RSN-BERT	39.4	73.8	51.3	56.4	73.1	63.6	44.9	30.3	55.9	39.6	30.8	42.2	35.6	30.3
	RoCORE	65.6	77.7	71.1	78.9	84.1	81.5	63.6	80.5	83.2	80.5	82.9	82.7	82.8	74.4
	MatchPrompt	**74.0**	**84.7**	**79.2**	**82.9**	**87.9**	**85.8**	**69.7**	**86.6**	**88.1**	**87.4**	**89.0**	**89.7**	**89.3**	**83.2**
100%	RSN*	48.6	74.2	58.9	64.4	78.7	70.8	45.3	62.8	63.4	63.1	62.4	66.3	64.3	45.9
	RSN-BERT*	58.5	**89.9**	70.9	69.6	**88.9**	78.1	53.2	79.5	87.8	83.4	84.9	87.0	85.9	75.6
	RoCORE*	75.2	84.6	79.6	83.8	88.3	86.0	70.9	87.1	84.9	86.0	**89.5**	88.1	88.8	82.1
	MatchPrompt	**75.6**	86.0	**80.3**	**84.2**	88.8	**86.5**	**71.2**	**88.5**	**87.9**	**88.2**	89.3	**89.5**	**89.4**	**83.5**

Table 2: Overall results of the compared models. "Source percentage" indicates the percentage of pre-defined relational instances in source domain used for pre-training knowledge transfer, e.g., 1% in FewRel refers to we randomly select $38400 * 1\% = 384$ instances (more details are in Appendix B.1). Results with $*$ are reported by Zhao et al. (2021). Note that we also convert the ARI into a percentage.

measure F_1 and ARI by $0.7\%/2.2\%$, $0.5\%/0.6\%$ and $0.3\%/1.4\%$ on FewRel/TACRED, respectively. When do not have knowledge transfer (source percentage $=$ 0%), our model improves the SOTA B^3 F_1, V-measure F_1 and ARI by $2.1\%/2.8\%$, $3.4\%/5.2\%$ and $1.2\%/2.8\%$ on FewRel/TACRED, respectively. On the one hand, this indicates that MatchPrompt can effectively generate relation-aware representations for examples in the open domain, which enables the derived clusters to have well-aligned relational semantics. On the other hand, it illustrates that MatchPrompt can remain effective without pre-training and knowledge transfer. **(2) Our MatchPrompt achieves competitive and robust performance with only a few pre-defined relational instances for knowledge transfer.** When MatchPrompt only with 10% data in the source domain for knowledge transfer, its performance on FewRel is almost close to the SOTA model using 100% data in the source domain, and its performance on B^3 F_1, V-measure F_1, ARI are even $1.4\%, 0.5\%, 1.1\%$ higher than SOTA model with 100% data in source domain on TACRED. This may be due to MatchPrompt bridging the task gap between representation learning and cluster optimization by eliminating the need to train additional classifiers like other models (e.g., SelfORE, RSN, RSN-BERT, RoCORE). It utilizes only a few data in the source domain as a drive to motivate templates to exploit the rich knowledge in the pre-trained language model, and generates better relation-aware representations for instances in the open domain.

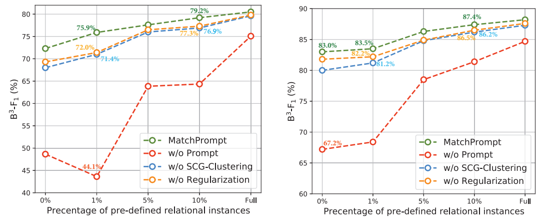
4.3 Ablation Studies

To further analyze MatchPrompt, we conduct ablation experiments to study the effectiveness of each component on two datasets. Specifically, MatchPrompt w/o Prompt does not use prompt learning to generate relational representations, and like previous works (Hu et al., 2020; Zhao et al., 2021), it concatenates the representations of head and tail entities produced by the language model to represent relational examples. MatchPrompt w/o SCG-Clustering is a model that replaces the proposed semantic consistency clustering method with

Dataset	Model	B^3F_1	V F_1	ARI
FewRel	**MatchPrompt**	**80.3**	**86.5**	**71.2**
	w/o Prompt	76.2	83.3	65.9
	Δ	4.1↓	3.2↓	5.3↓
	w/o SCG-Clustering	79.6	86.4	69.8
	Δ	0.7↓	0.1↓	0.4↓
	w/o Regularization	79.8	85.0	69.7
	Δ	0.5↓	1.5↓	0.5↓
TACRED	**MatchPrompt**	**88.2**	**89.4**	**83.5**
	w/o Prompt	84.7	85.7	79.7
	Δ	3.5↓	3.7↓	3.8↓
	w/o SCG-Clustering	87.3	88.0	82.7
	Δ	0.9↓	1.4↓	0.8↓
	w/o Regularization	87.6	88.2	82.8
	Δ	0.6↓	1.2↓	0.7↓

Table 3: Ablation results on FewRel and TACRED (Source percentage = 100%).

k-Means clustering. MatchPrompt w/o Regularization means removing the regularization loss of $\mathcal{L}_{scr}$ and $\mathcal{L}_{mlm}$ from the model. Table 3 represents the results with 100% source domain data and Fig. 3 further compares the performance when the percentage of data in the source domain differs. we can make the following observations: **(1) All components contribute to the improvement of model performance.** In both FewRel and TACRED, the combination of the above three terms can result in obvious performance improvements. **(2) Prompt term is essential to mine representations of unlabeled novel relations in all cases.** w/o Prompt reduces model performance by $3.5\% \sim 31.8\%$ on two datasets, especially when only a few source data are used for knowledge transfer. This indicates that the prompt is essential for effectively transferring knowledge from a few pre-defined relational instances and takes advantage of the rich knowledge in the pre-trained language model for novel relation clustering. **(3) SCG-Clustering term and Regularization term are important to MatchPrompt, especially with less data for pretraining.** w/o SCG-Clustering term, the model performance of B^3 F_1 drops by $2.3\%/1.2\%$ and $4.5\%/2.3\%$ with 10% and 1% percentage of source data on FewRel/TACRED, respectively. Similarly, without regularization term decreases the model performance on B^3 F_1 by $1.9\%/0.9\%$ and $3.9\%/1.3\%$ when using 10% and 1% percentage of source data on FewRel/TACRED, respectively. This illustrates that our clustering method indeed alleviates the negative impact of incorrect pseudo-labels for better clustering. Meanwhile, the regularization loss encourages the model to produce more consistent representations for similar instances in the open domain.



(a) FewRel (b) TACRED

Figure 3: Ablation results (B^3 F_1) on two datasets with the percentage of source domain data differs.

4.4 Generalized Relation Extraction

Generalized relation extraction means the model should be able to extract pre-defined and novel relations simultaneously. One advantage of MatchPrompt is that it aims to generate relation-aware representations for each given instance, and thus has the ability for generalized relation extraction. RoCORE (Zhao et al., 2021) designs a classifier that combines pre-defined and novel relation types together, which is the only existing method that can achieve generalized relation extraction. In this experiment, we randomly select pre-defined relations as $\{10, 20, 40\}$ for knowledge transfer [3] and always use all 16 novel relations to evaluate model performance on FewRel. Table 4 compares the results of these three cases. We can make the following observations: **(1) MatchPrompt works well on novel relations and sightly outperforms RoCORE on both pre-defined and novel relations.** Specifically, Matchprompt outperforms RoCORE by $12.9\% \sim 15.3\%$ on novel relation clustering in different cases. Meanwhile, the overall performance of MatchPrompt on both pre-defined and novel relations is relatively stable, while the performance of RoCORE is more sensitive to the number of pre-defined relations. **(2) MatchPrompt can narrow the clustering bias on pre-defined relations.** Using pre-defined relations for knowledge transfer will inevitably introduce clustering bias on these relations. The B^3 F_1 discrepancy between pre-defined and novel relations for RoCORE is 23.0% on average, while for MatchPrompt is 5.8% on average. The reason may be that MatchPrompt can achieve effective knowledge transfer and generate relation-aware representations in `[MASK]` tokens,

[3] For fair comparisons, we guarantee that the relation types randomly selected for different models are always the same.

Case Model	1: Pre/Nov(10/16)			2: Pre/Nov(20/16)			3: Pre/Nov(40/16)		
	Pre	Nov	All	Pre	Nov	All	Pre	Nov	All
RoCORE	**92.9**	63.1	73.9	**87.2**	64.7	77.3	**84.1**	67.3	**78.5**
MatchPrompt	90.6	**78.4**	**77.8**	84.5	**79.7**	**78.6**	80.7	**80.2**	78.4

Table 4: The results of generalized relation extraction on B^3 F_1. "Pre"refers to the average performance in pre-defined relations (Souce domain) while "Nov" refers to the average performance in novel relations (Open Domain). "All" indicates the average performance on both source domain and open domain testing data. For example, "Case 1: Pre/Nov(10/16)" means only using 10 pre-defined relations for pre-training, testing model performance with 10 pre-defined relations (Pre), 16 novel relations (Nov), and all 10+16=26 relations (All).

Open Relations	**Examples** (<S_1>. head entity `[MASK]` tail entity.)	**Intra**	**Inter**	**Top Predicted Relational Words**
Mother	She married Polyctor, son of Aegyptis and Caliadne. Polyctor `[MASK]` Caliadne.	98.0	36.1	parenting, partnered, marries
Spouse	Emma is married to Polanski. Emma `[MASK]` Polanski.	99.0	56.6	married, marriages, spouse
Child	She married Polyctor,son of Aegyptis and Caliadne. Caliadne `[MASK]` Polyctor.	96.0	37.8	parenting, trained, marriages
Constellation	NGC354 is a spiral galaxy in the constellation Pisces. NGC354 `[MASK]` Pisces.	100.0	99.0	borders, edges, constellation
Follows	He was re-relected in 1896, 1900, 1904, and 1908. 1908 `[MASK]` 1904.	80.3	84.2	##-1, ##nary, ##rricular
Next to body of water	Pie island is an island in lake superior ntario Canda. Pie island `[MASK]` lake superior.	96.1	80.4	##urbed, ##uaries
Part of	Herm is one of the channel islands in the english channel. Herm `[MASK]` channel islands.	21.1	15.4	in, ##under, complete
Original language of work	The film is remake of hindi movie "Suhaag". Suhaag `[MASK]` hindi.	98.9	99.0	in, spoke, balthazar
Military rank	Edwin was an American major general. Edwin `[MASK]` major general.	97.9	98.9	##uga, ##anga, ##ibar

Table 5: The clustering results and top predicted relational words for some novel relations on FewRel. "Intra" and "Inter" mean intra-relation consistency (%) and inter-relation separability (%), respectively.

which alleviates the clustering bias on pre-defined relations to a certain extent.

4.5 Discussion: which novel relations can be well recognized?

We evaluate the extraction result of novel relations on FewRel by the following two aspects.

Clustering results. For each novel relation, we set intra-relation consistency and inter-relation separability to evaluate the clustering results. A well-recognized novel relation requires both high intra-relation consistency and inter-relation separability. After clustering, instances from one relation may be assigned to different clusters. Specifically, we represent each relation by the cluster that has the highest coincidence degree with it, then calculate the percentage of these coincident instances to the total instances of the relation as an intra-relation consistency. Also, we use the percentage of these coincident instances to all instances in the current cluster as the inter-relation separability[4]. As shown in Table 5, we find that the clustering results of novel relations are related to two factors. **(1) Whether have prominent semantic features**. For example, *Constellation*, *Original language of work* and *Military rank* are both salience and have high intra-relation consistency ($> 97\%$) as well as high inter-relation separability ($> 98\%$). In contrast, *Part of* has little significance as its semantics covers a broad range, e.g., part of an island, part of a literary, or part of a sporting season. **(2) Whether similar with other novel relations.** We notice that intra-relation consistency ($98\%/96\%/99\%$) for *Mother/Child/Spouse* are high while inter-relation separability ($36.1\%/37.8\%/56.6\%$) are relatively low. This indicates that it is difficult for the model to distinguish them from each other. This may be

[4] Assume there are 100 instances of one relation, with 20 in cluster A and 80 in cluster B. Then, we use cluster B as the representative cluster of this relation. If the total instances of cluster B are 300, then the intra-relation consistency and the inter-relation separability of this relation are $(80/100) * 100\% = 80\%$ and $(80/300) * 100\% = 26.7\%$, respectively.

due to the approximate nature of these sentences, even the same sentences express different relations (simply swap the head and tail entities).

Top predicted relational words. One advantage of our MatchPrompt is that it can convert relation-aware representations to relational words for unlabeled instances in the open domain, providing better interpretability of clustering results compared to other models. To specifically quantify the recognition results of novel relations, we show the top predicted relational words in Table 5. Specifically, we get the words with the highest probability in `[MASK]` tokens and report the top three words with the highest frequency among 100 instances for each novel relation. We find that **the predictability is higher for relational words that appear in sentences.** For example, the semantics of the top predicted words "married","marriage" and " spouse" are highly related to novel relation *Spouse*. Furthermore, even though some of the proper nouns (e.g., "NGC345" and "Pisces") in sentences are incomprehensible, cue words like "galaxy" and "constellation" appear in the sentence can prompt our model to find these proper nouns related to *Constellation*. Conversely, predicting accurate words for *Follows* is difficult since there are no cue words appearing in sentences.

To sum up, clustering results and top predicted relational words are two independent indicators for evaluating novel relations. The former focuses on the consistency of high-dimensional relation-aware vectors, while the latter concerns the explicit semantics of relations conveyed by sentences.

5 Related Work

Open Relation Extraction. Relation Extraction (RE) is one of the most essential technology for knowledge graph construction. Traditional RE methods mainly focus on identifying the relations between two entities from pre-defined relation categories (Dong et al., 2021; Li et al., 2021; Ji et al., 2020) and are incapable of extracting novel relations emerging in the real world. Open relation extraction (OpenRE) has been explored to meet this needs. Previous work can be divided into tagging-based methods (Banko et al., 2007; Yates et al., 2007) and clustering-based methods (Yao et al., 2011; Simon et al., 2019; ElSahar et al., 2017). Tagging-based methods usually extract multiple phrases in sentences as relations and lack generality due to the sentences with the same relation may express differently (Fader et al., 2011). Comparatively, clustering-based methods have drawn more attention. Hu et al. (2020) proposes a self-supervised framework that leverages a large pre-trained language model for adaptive clustering. Liu et al. (2021a) revisit OpenRE from a causal view and formulates relation clustering by a structural causal model. Zhao et al. (2021) uses large amounts of pre-defined relational instances to learn a relational-oriented clustering model for novel relation clustering in the open world. In this paper, we focus on clustering-based methods.

Prompt-based Learning. Prompt-based learning is a new paradigm and has drawn more attention since the emergence of GPT-3 (Brown et al., 2020). Prompt-based methods have achieved outstanding performance in various NLP tasks (Chen et al., 2022; Zhu et al., 2022; Schick and Schütze, 2021). Specifically, a prompt contains a template and a verbalizer. Existing methods focusing on different kinds of verbalizers (Gao et al., 2021). For instance, human-written verbalizers are built by associating labeled words with pre-defined types manually (Schick and Schütze, 2021) and have less coverage; Automatically determined verbalizers use pre-defined types to search labeled words after optimizing on a lot of training data (Shin et al., 2020; Liu et al., 2021b; Hu et al., 2022). Different from the traditional prompt paradigm, there are no pre-defined types (relations) in the open domain, thus we cannot construct a verbalizer for OpenRE directly. In order to solve this problem, we devise a pairwise matching strategy to drive prompt learning in the open domain.

6 Conclusion

In this work, we propose a prompt-based open relation clustering framework, MatchPrompt, for OpenRE. The proposed model can generate efficient representations with efficient knowledge transfer from only a few pre-defined relational instances as well as mine the specific meanings of clusters. Meanwhile, these relation-specific representations provide interpretability for OpenRE clustering. Experimental results on two datasets show that MatchPrompt achieves the new SOTA results. Meanwhile, it has a competitive performance with only 10% of pre-defined relational instances for model pre-training compared to the current SOTA method, which requires 100% of these data for pre-training.

Limitations

We consider that our current method has the following two limitations. (1) we use pre-defined relational instances for pre-training knowledge transfer. During the experiment, we fixed the pre-training number as $\mu = 5$. However, this fixed number should be considered to be an adaptive value to better satisfy different data in the further. (2) For fair comparisons, we only use the BERT_base in our experiments like the previous baseline models. However, the impact of different pre-trained language models (e.g., BERT_large, RoBERTa_base, RoBERTa_large, and so on) on the performance of MatchPrompt should be explored in the future.

Acknowledgments

We sincerely thank all anonymous reviewers for their constructive comments. This work is supported by National Key Research and Development Program of China (2020AAA0108800), National Natural Science Foundation of China (62137002, 61937001, 62192781, 62176209, 62176207, 62106190, and 62250009), Innovative Research Group of the National Natural Science Foundation of China (61721002), Innovation Research Team of Ministry of Education (IRT_17R86), MoE-CMCC "Artifical Intelligence" Project (MCM20190701), Consulting research project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China", "LENOVO-XJTU" Intelligent Industry Joint Laboratory Project, CCF-Lenovo Blue Ocean Research Fund, Project of China Knowledge Centre for Engineering Science and Technology, Foundation of Key National Defense Science and Technology Laboratory (6142101210201), the Fundamental Research Funds for the Central Universities (xhj032021013-02, xzy022021048, xpt012022033).

References

David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 1027–1035.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, pages 563–566.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 2670–2676.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 3470–3479.

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 *WWW '22: The ACM Web Conference 2022*, pages 2778–2788.

Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. 2021. Revisiting self-training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 9125–9135.

Kenneth Church and Yuchen Bian. 2021. Data collection vs. knowledge graph completion: What is needed to improve coverage? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 6210–6215.

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, NAACL-HLT*, pages 4171–4186.

Manqing Dong, Chunguang Pan, and Zhipeng Luo. 2021. Mapre: An effective semantic mapping approach for low-resource relation extraction. In *Proceedings of the 2021 Conference on Empirical*

Methods in Natural Language Processing, EMNLP, pages 2694–2704.

Hady ElSahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frédérique Laforest. 2017. Unsupervised open relation extraction. In *The Semantic Web: ESWC 2017 Satellite Events - ESWC 2017 Satellite Events*, volume 10577, pages 12–16.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1535–1545. ACL.

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, ACL/IJCNLP 2021*, pages 3816–3830. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 2225–2240. Association for Computational Linguistics.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip S. Yu. 2020. Selfore: Self-supervised relational feature learning for open relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3673–3682.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Bin Ji, Jie Yu, Shasha Li, Jun Ma, Qingbo Wu, Yusong Tan, and Huijun Liu. 2020. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING*, pages 88–99.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*.

Xianming Li, Xiaotian Luo, Chenghao Dong, Daichuan Yang, Beidi Luan, and Zhen He. 2021. TDEER: an efficient translating decoding schema for joint extraction of entities and relations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 8055–8064.

Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. 2021a. Element intervention for open relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pages 4683–4693.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. GPT understands, too. *CoRR*, abs/2103.10385.

Mehrnoosh Mirtaheri. 2021. Relational learning to capture the dynamics and sparsity of knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pages 15724–15725.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2339–2352.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 4222–4235.

Étienne Simon, Vincent Guigue, and Benjamin Piwowarski. 2019. Unsupervised information extraction: Regularizing discriminative approaches with relation distribution losses. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 1378–1387. Association for Computational Linguistics.

Thy Thy Tran, Phong Le, and Sophia Ananiadou. 2020. Revisiting unsupervised relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 7498–7505.

Aravindan Vijayaraghavan, Abhratanu Dutta, and Alex Wang. 2017. Clustering stable instances of euclidean k-means. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6500–6509.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2021. Zero-shot information extraction as a unified text-to-triple translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 1225–1238. Association for Computational Linguistics.

Ruidong Wu, Yuan Yao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2019. Open relation extraction: Relational knowledge transfer from supervised data to unsupervised data. 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 219–228.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011*, pages 1456–1466. ACL.

Alexander Yates, Michele Banko, Matthew Broadhead, Michael J. Cafarella, Oren Etzioni, and Stephen Soderland. 2007. Textrunner: Open information extraction on the web. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA*, pages 25–26. The Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pages 4904–4917. Association for Computational Linguistics.

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, EMNLP*, pages 35–45.

Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. 2021. A relation-oriented clustering method for open relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 9707–9718.

Qi Zhu, Bing Li, Fei Mi, Xiaoyan Zhu, and Minlie Huang. 2022. Continual prompt tuning for dialog state tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics ACL 2022*, pages 1124–1137. Association for Computational Linguistics.

A Details of Experiment Setup

A.1 Datasets.

FewRel contains 70,000 relational examples on 100 relations, that is, each relation has 700 examples. However, only 80 relations were released. We follow the same setting in (Zhao et al., 2021; Wu et al., 2019) to utilize the original training set with 64 relations and the original validation set with 16 relations of FewRel as source domain and open domain, respectively. For each relation, we randomly select 100 instances for the test. **TACRED** contains 42 relations and similar to the setting of FewRel, we remove the instances labeled as "No Relation" and use the remaining 0-30 and 31-40 relation types as the source domain and the open domain, respectively. For each relation, we randomly select 15% of the instances for the test.

A.2 Baselines.

HAC with Re-weighted Word Embeddings (RW-HAC) (ElSahar et al., 2017). In OpenRE, RW-HAC is a method for clustering features. In order to construct relational features, the model utilizes weighted word embeddings as well as the type of entities. **Entity Based URE (Etype+)** (Tran et al., 2020). A method based exclusively on entity types known as Etype+ is a simple and effective model. Etype+ employs the link predictor and two additional regularisers. **Self-supervised Feature Learning for OpenRE (SelfORE)** (Hu et al., 2020). In SelfORE, self-supervised signals are exploited through the use of a large pre-trained language model for adaptive clustering based on contextualized relational features. **Relational Siamese Network (RSN)** (Wu et al., 2019). By learning similarity metrics from labeled data about pre-defined relations, RSN identifies novel relations from unlabeled data by transferring the relational knowledge. **RSN with BERT Embedding (RSN-BERT)** (Wu et al., 2019). RSN-BERT is regarded for comparison purposes as a variant of RSN in which the static word vector is replaced by the BERT embedding. **Relation-Oriented Open Relation Extraction (RoCORE)** (Zhao et al., 2021). To identify novel relations in the unlabeled data, RoCORE proposes a relation-oriented clustering model which learns the relation-oriented representation by using the readily available labeled data of pre-defined relations.

Hyper-parameters	value
Pre-trained Language Model	BERT_base_uncased
optimizer	Adam
learning rate	1e-4
batch size	100
pre-training epochs μ	5
loss coefficient $\mathcal{L}_s$ for FewRel	0.7
loss coefficient $\mathcal{L}_{SC}$ for FewRel	0.01
loss coefficient $\mathcal{L}_s$ for TACRED	0.5
loss coefficient $\mathcal{L}_{SC}$ for TACRED	0.01

Table 6: Hyper-parameter settings

A.3 Implementations.

For fair comparisons, our MatchPrompt utilizes the pre-trained language model BERT-base (Devlin et al., 2019) as previous baseline models. The optimizer is Adam (Kingma and Ba, 2015), in which the learning rate is $1e-4$. The batch size is 100, and the dimension of output features is 768. We set the maximum training epoch number as 100 and test the model performance on the test set every epoch. The training will stop if there is no growth for 15 consecutive epochs. We set the pre-training epoch number μ as 5. All experiments are conducted using a TeslaA100 with 80 GB of memory. In most of our experiments, we chose Δ_1 and Δ_2 as 7 and 2 for best results, respectively. Table 6 shows our hyper-parameter settings.

B Supplementary Experiment Results

B.1 Percentage of Source Data

"Source percentage" indicates the percentage of pre-defined relational instances in the source domain used for pre-training knowledge transfer. This means we only randomly select the training data in the source domain by percentage, and do not change other data. The details of the datasets under different "Source percentage" are shown in Table 7. Specifically, we randomly select 6, 30, 60 instances per category on average under 1%, 5%, and 10% settings on FewRel, respectively. Similarity, we randomly select 3, 16, 31 instances per category on average under 1%, 5%, and 10% settings on TACRED, respectively.

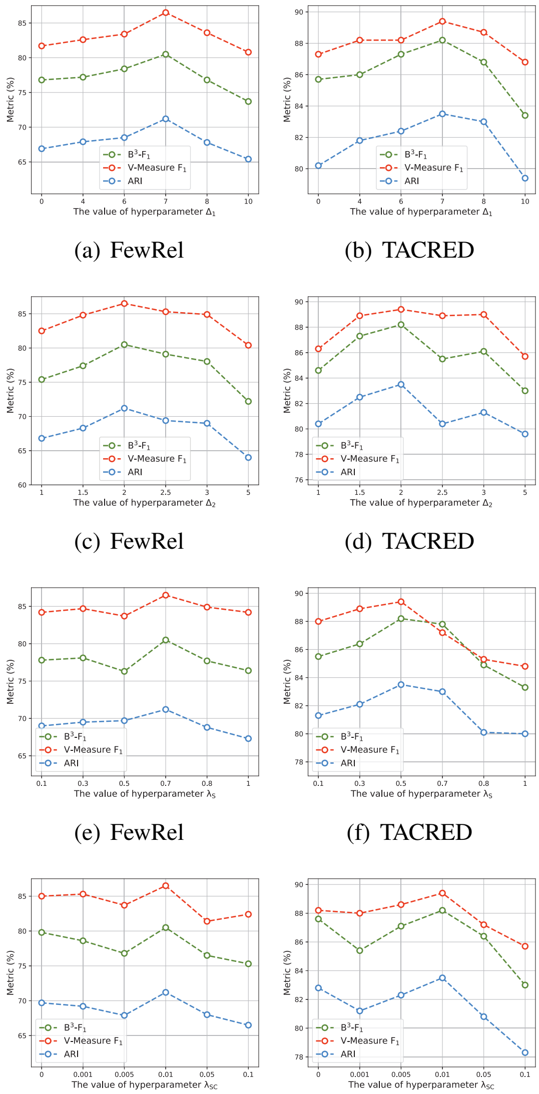
B.2 Hyper-parameter Analysis

In this section, we present the analysis of hyper-parameters in our model, including $\mathcal{L}_s$, $\mathcal{L}_{SC}$, Δ_1 and Δ_2. For any one, we fix the remaining three parameters. The results are shown in Fig. 4. The search scope of

Source Percentage	FewRel		TACRED	
	Source Doamin	Open Domain	Source Domain	Open Domain
	#Rel(train_n/test_n)	#Rel(train_n/test_n)	#Rel(train_n/test_n)	#Rel(train_n/test_n)
1%	64(384/6400)	16(9600/1600)	31(98/1644)	10(1038/174)
5%	64(1920/6400)	16(9600/1600)	31(493/1644)	10(1038/174)
10%	64(3840/6400)	16(9600/1600)	31(986/1644)	10(1038/174)
100%	64(38400/6400)	16(9600/1600)	31(9865/1644)	10(1038/174)

Table 7: The details of two datasets under different “Source percentage”. “Rel” refers to relation types, “train_n” refers to number of training instances and “test_n” refers to number of test instances.

$\mathcal{L}_s$ is within$\{0, 0.1, 0.3, 0.5, 0.7, 0.8, 1\}$, and the optimal value for FewRel and TACRED are 0.7 and 0.5, respectively. The search scope of $\mathcal{L}_{SC}$ is within$\{0, 0.001, 0.005, 0.01, 0.05, 0.1\}$, and the optimal value for FewRel and TACRED are both 0.01. The search scope of Δ_1 is within$\{0, 4, 6, 7, 8, 10\}$, and the optimal value for FewRel and TACRED are both 7. The search scope of Δ_2 is within$\{1, 1.5, 2, 2.5, 3, 5\}$, and the optimal value for FewRel and TACRED are both 2.



(a) FewRel (b) TACRED

(c) FewRel (d) TACRED

(e) FewRel (f) TACRED

(g) FewRel (h) TACRED

Figure 4: Hyperparameter Analysis.