Improving Few-Shot Relation Classification by Prototypical Representation Learning with Definition Text

Zhenzhen Li[1], **Yuyang Zhang**, **Jian-Yun Nie**[2], **Dongsheng Li**[1]
[1] National University of Defense Technology, Changsha, China
[2] DIRO, Université de Montréal, Montreal, Canada
{lizhenzhen14, dsli}@nudt.edu.cn, nie@iro.umontreal.ca

Abstract

Few-shot relation classification is difficult because the few instances available may not represent well the relation patterns. Some existing approaches explored extra information such as relation definition, in addition to the instances, to learn a better relation representation. However, the encoding of the extra information has been performed independently from the labeled instances. In this paper, we propose to learn a prototype encoder from relation definition text in a way that is useful for relation instance classification. To this end, we use a joint training approach to train both a prototype encoder from definition and an instance encoder. Extensive experiments on several datasets demonstrate the effectiveness and usefulness of our prototype encoder from definition text, enabling us to outperform state-of-the-art approaches.

1 Introduction

Relation classification (RC) aims to determine the relation expressed between two entities in a sentence. Typical approaches to RC train a classification model or a prototype from labeled sentences, which are intended to represent the typical relation patterns including various syntactic, semantic and contextual features. In practical situations, we may have only a few labeled examples, which are insufficient for training a good classification model. In some cases, the labeled examples may also be atypical. For example, if a relation is exemplified by sentences describing Obama's presidency in the US, the resulting prototype would be largely tuned toward a political context, which would be difficult to apply outside the political context. Even for a human annotator, with only a relation ID and some examples, it would be difficult to learn to classify a relation correctly. This is exactly the same problem faced by a few-shot RC system: the best guess based on the limited labeled examples may fail.

If, however, the human annotator is told that the relation is named "position held", then he/she would have a better understanding of the relation to better generalize the examples to other instances. In this case, the human annotator indeed exploits the prior knowledge about the relation (from its name). If more information such as a description of the relation is available, then the annotator would do an even better job. The definition of a relation, being it the name or the description, is important to help the human annotator understand the relation. It is the same for an automatic RC model.

Several previous studies have explored using such extra knowledge about the relation (Qu et al., 2020; Dong et al., 2020; Zhang et al., 2021). For example, relation names have been used for better initializing the relation prototype before being trained with labeled instances (Dong et al., 2020). However, Dong et al. only considered a limited number of relations in the training data, making the few-shot RC model prone to over-fitting. Inspired by these studies and human behaviors, in this paper, we leverage relation definitions to build high-level prototype representations of the relations.

A second problem we consider in this paper is the ability to construct prototypes for new relations, i.e. zero-shot learning. In this case, if the new relation has a description, then one would be able to apply the same prototype encoder learned on the other relations to the new relation. In other words, the mechanism of building relation prototypes from their definitions could be generalized and transferred to a new relation. Some previous work (Qu et al., 2020) has considered prototype transfer based on a knowledge graph which provides relationships between relations. It is assumed that a relation's prototype can be partly transferred to a related relation in the graph. While this could be a possible way to generate a more reasonable prototype for a relation, we believe that relation definitions provide a better basis for constructing a

Findings of the Association for Computational Linguistics: NAACL 2022, pages 454 - 464

Relation name	Description	Labeled data
language of work	language associated with creative work, such as books, shows, songs, or websites	**Nokta** ("dot" in Turkish) was a leading **Turkish** weekly political news magazine **All Things Must Pass** is a triple album by **English** musician George Harrison **It** was performed in **French** by French singer France Gall
position held	subject currently or formerly holds the object position or public office	**Goebbels** succeeded him as **Chancellor** of Germany It is named after **Justus**, **Archbishop** of Canterbury from 624-627 **He** represented Central Lancashire as a Member of the European Parliament **(MEP)**

Table 1: Example of relation definitions (relation name and description) and weakly labeled data for relation P407 and P39 in Wikidata, both of which are used for prototypical representation learning.

relation prototype.

We propose to learn a general mapping function from relation definitions to their prototypes. The advantages of the approach come from the fact that the definition text expresses the intrinsic semantics of relation, which is not explicitly covered by labeled instances. For example, as shown in Table 1, the name of the relation "language of work" provides a general idea of the relation. The description further specifies that it is the language used for "creative work". Such information is not explicitly expressed, but only hinted, in the corresponding labeled examples. The definition and labeled instances provide different but complementary information about the relation, therefore can be combined to improve RC.

Mapping a relation definition to a prototype vector could be done naively using a pre-trained language model such as BERT (Devlin et al., 2019), but the resulting prototype may not be the most useful for relation instance classification. We believe that a good prototype encoded from a definition should be the one that helps RC. Therefore, we propose to train a prototype encoder together with RC of some examples. To tackle the problem of limited labeled examples, inspired by the RC-oriented pre-training work (Baldini Soares et al., 2019), we use abundant distantly labeled data to help train a prototype representation. The pre-training will generate the relation prototype vector that can best classify the weakly labeled data.

To integrate the prototype representation learned from relation definitions with the given limited hand-labeled instances, we adapt the Bayesian meta-learning approach (Qu et al., 2020) to learn a posterior distribution of the prototype vectors of relations based on both the initial prototype representations and the labeled instances. This process helps adapt the prototype to the labeled instances.

We test our approach on two few-shot RC datasets. It outperforms previous competitive models that apply pre-trained instance encoder or relation definition text. We also show that the encoder can be easily generalized to new relations in zero-shot RC setting. Experimental results demonstrate the effectiveness and generalization ability of our pre-trained prototype encoder.

Our main contribution in this paper is twofold: 1. We propose a new relation prototype construction method from relation definition; 2. We experimentally show that the approach is effective in few-shot RC and can be generalized to new relations in zero-shot RC.

2 Problem Definition

RC aims to predict whether a sentence (instance) $\mathbf{x}$ expresses a pre-defined relation between two given entities (e_1, e_2). Neural RC models usually contain an instance encoder that encodes the relation expressed by an instance into a dense vector and a classification layer to classify the dense vector to the relation which has the most similar prototype vector. In this work, we leverage the definition text denoted as $\mathbf{y}$ to help learn the relation prototype.

The prototype encoder is trained to help the classification of relation instances. Some training instances are required. Inspired by the RC-oriented pre-training work (Baldini Soares et al., 2019; Peng et al., 2020), we leverage a large set of distantly labeled data. The distant labeling (Mintz et al., 2009) is done as follows: given a known relation containing a pair of entities, any sentence mentioning the same entity pair is labeled by the relation.

Let us denote the set of distantly labeled instances as $\mathbb{D} : \{(\mathbf{x}_i, e_{1i}, e_{2i}, r_i)\}_{i=1}^{z}$ and a set of definition texts for all relations in $\mathbb{D}$ denote as $\mathbb{T} : \{(\mathbf{y}_t, r_t)\}_{t=1}^{n}$. The instance encoder (instance relation encoder) RelEnc_ϕ, parameterized by ϕ, produces an instance embedding $\mathbf{s} \in \mathbb{R}^d$ by $\mathbf{s} = \mathrm{RelEnc}_\phi(\mathbf{x}, e_1, e_2)$, where d is the vector dimension of prototype representation. The prototype encoder $\mathrm{ProtoEnc}_\theta$, parameterized by θ, produces a set of definition representations $\{\mathbf{v}_t\}_{t=1}^{n}$, where $\mathbf{v}_t \in \mathbb{R}^d$ is produced by $\mathbf{v}_t = \mathrm{ProtoEnc}_\theta(\mathbf{y}_t)$. The

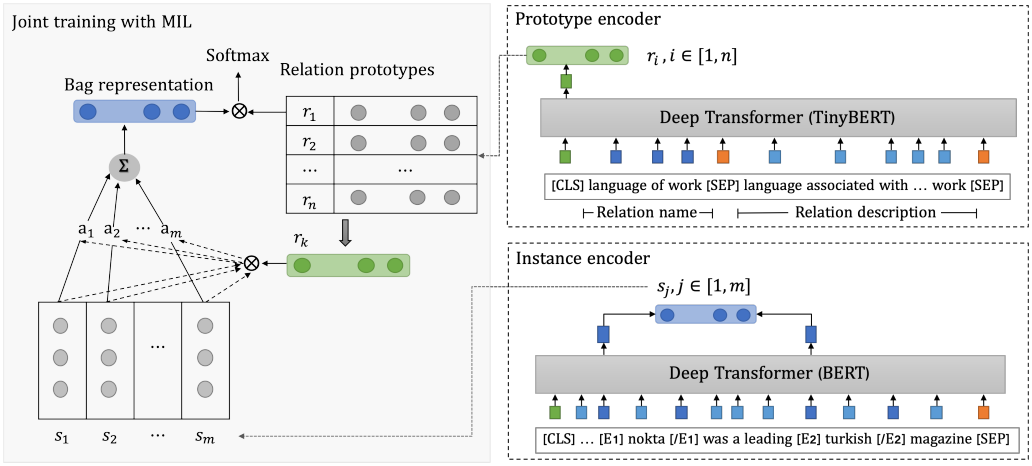

Figure 1: Overview of the prototype encoder pretraining framework. On the right side, two modules encode different text separately: the prototype encoder encodes the definition text as prototypes; the instance encoder encodes the possible relation expressed by an instance as the instance embedding. On the left side, a multi-instance learning method takes the representations from the two encoders for relation classification.

training goal is to map definition representations to relation prototypes useful for classifying instances.

3 Methodology

In this section, we first explain how a relation prototype is learned with the help of a set of distantly labeled instances. Then we describe how the learned prototype is further enhanced by the limited labeled instances in a Bayesian meta-learning framework.

3.1 Pre-training Framework

Pre-training a general prototype encoder involves two processes: generalizing contextual features of distantly labeled data to prototypes and mapping the relation definitions to their prototypes. As instances and the definition text describe relations in different forms, we adopt two pre-trained language models based on BERT (Devlin et al., 2019), as the backbones to encode them separately. The utilization of two distinct encoders is motivated by the different natures of the inputs: Relation definition provides a high-level, conceptual description of the relation, while a relation instance provides a concrete example about relation expression. From the former, the encoder would be asked to capture the general concepts, while from the latter, features relating to the syntactic/semantic pattern or the context can emerge. [1]. To cope with the large quantity of relations and the noisy labeling problem of distantly labeled data, we propose a simple yet effective design of instance encoder, prototype encoder and their joint training process, as illustrated in Figure 1. The joint training process learns the mapping function of prototype encoder by directly regarding definition representations as relation prototypes for classification. We elaborate the three components of the pre-training framework in the following parts.

3.1.1 Instance Encoder

We use the entity marker strategy to extract the instance relation following previous investigation about architectures of instance encoder (Baldini Soares et al., 2019). An example is shown in Figure 1. Four special entity markers - `[E1]`,`[/E1]`, `[E2]`, `[/E2]` are used to delimit entity1 and entity2. Then the hidden vectors of `[E1]` and `[E2]` at the last layer are concatenated to represent the relation expressed by the instance. Thus the hidden representation of an instance $\mathbf{x}_i$ is obtained as $H_i = \text{RelEnc}_\phi(\mathbf{x}_i, e_{1i}, e_{2i}) =< \mathbf{h}_{e_1} | \mathbf{h}_{e_2} >$, where $H_i \in \mathbb{R}^{2d}$ and d is the the size of the hidden representation space of the pre-trained language model. The final instance embedding is transformed into a d-dimensional representation by a linear layer as $\mathbf{s}_i = H_i^T W_t$, where parameters $W_t \in \mathbb{R}^{2d \times d}$.

3.1.2 Prototype Encoder

Each relation from KBs typically has a name and a description text. We use `[SEP]` to separate the relation name and description, and add the special token `[CLS]` at the beginning, another `[SEP]` at

[1] We also tested with the same encoder, but obtained poor results.

the end. Figure 1 shows an example. After tokenizing the input text, we feed the input tokens into TinyBERT and use the hidden vector of `[CLS]` (of dimension d) of the last layer as the definition representation. We adopt TinyBERT (Jiao et al., 2020) as the prototype encoder for its high efficiency. We also experimented the alternative with $\text{BERT}_{\text{BASE}}$ to update relation prototypes within a mini batch. However, this alternative converges much more slowly and at a higher final training loss than using TinyBERT.

3.1.3 Joint Training with MIL

Multi-instance learning (MIL) has been widely used to alleviate the noisy labeling problem (Ji et al., 2017; Alt et al., 2019) in distantly labeled data. It regards a set of instances containing the same entity pair as a bag and assign the bag with one relation label. Then relation classification is relaxed from sentence level to bag level. By selecting the most reliable instance or assigning different attention weights among the instances of the bag (Lin et al., 2016), the impact of wrong labels is reduced.

In this work, we use the relation definition to guide the instance attention learning among the bag. We assume that *instances that are semantically closer to the relation definition are more likely to express such a relation*. Let us denote a bag sample from the noisy dataset $\mathbb{D}$ as $(B_k, r_k, e_{1,k}, e_{2,k})$, where $B_k = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ is a set of instances containing the entity pair $(e_{1,k}, e_{2,k})$, which are encoded as $\{\mathbf{s}_i\}_{i=1}^m$ (also called 'instance embeddings') using RelEnc_ϕ. The bag representation $\mathbf{b}_k$ is computed by aggregating all instances according to the selective attention weights as follows:

$$\mathbf{b}_k = \sum_{i=1}^{m} a_i \mathbf{s}_i, \; a_i = \frac{\exp(\mathbf{s}_i^T \mathbf{v}_k)}{\sum_{j=1}^{m} \exp(\mathbf{s}_j^T \mathbf{v}_k)} \quad (1)$$

where the attention weight a_i represents the confidence score of instance $\mathbf{x}_i$ expressing relation r_k. a_i is calculated according to the similarity of instance embedding $\mathbf{s}_i$ and the definition representation $\mathbf{v}_k$.

We use dot product to compute the similarity of $\mathbf{b}_k$ and candidate relation prototypes $\{\mathbf{v}_i\}_{i=1}^n$. Then the bag-level prediction probability for relation r_k is computed as follows:

$$p(r_k|B_k, W_t, \phi, \theta) = \frac{\exp(\mathbf{b}_k^T \mathbf{v}_k)}{\sum_{j=1}^{n} \exp(\mathbf{b}_k^T \mathbf{v}_j)}. \quad (2)$$

Standard cross entropy is used to compute RC loss. We also add an auxiliary loss about language modeling over training instances to avoid catastrophic forgetting. We follow the same setting as previous work (Devlin et al., 2019; Baldini Soares et al., 2019) to compute the masked language modeling loss ($\mathcal{L}_{\text{MLM}}$). The final loss is defined as Eq. 3,

$$\mathcal{L} = \alpha * \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{RC}}, \quad (3)$$

where α controls the importance of the language modeling loss, set as 0.5 by default. We update the parameters of $\{\phi, \theta, W_t\}$ for both encoders at each training iteration to minimize the final loss.

3.2 Application to Downstream RC

After the pre-training phase, we use the limited labeled instances to further enhance the prototype representation.

We adopt the Bayesian meta-learning approach proposed in (Qu et al., 2020), which models the uncertainty of prototype vectors by regarding them as random variables and learning the probability distribution for each relation. It could effectively learn the posterior distribution of the relation prototypes by combining the prior knowledge and a few labeled instances (i.e. the support set). In the model of (Qu et al., 2020), the prior of prototype vectors is derived from the structural relationship of different relations in a knowledge graph. The assumption is that a relation (a node in the graph) can gain some information from its neighbors. Relation representations are derived by applying graph neural network. In our case, prior relation prototypes are obtained from relation definition texts. A big advantage is that we do not require a relation be included in the knowledge graph to be able to handle it. A new relation can be handled if we have a definition of it, which is often the case (or the definition can be easily created) in practice. We will show later that our prototype representation performs better than that of (Qu et al., 2020).

The Bayesian meta-learning framework works as follows. For each sample in few-shot learning, given the candidate relations C, we denote their labels and textual definitions as $\mathbf{r}_C$ and $\mathbf{y}_C$; given a few supporting instances S, we denote their sentences and relation labels as $\mathbf{X}_S$ and $\mathbf{r}_S$, where each relation $r_s \in \mathbf{r}_C$. In Bayesian statistics, we infer the posterior distribution of relation prototypes as follows. Considering context variables $\{\mathbf{X}_S, \theta, \mathbf{y}_C\}$, we formulate a Bayesian formula for $p(\mathbf{v}_C|\mathbf{r}_S)$ in Eq. 5. Eq. 6 could be obtained when we assume that $\mathbf{r}_S$ is uniformly distributed. Since

$\theta, \mathbf{y}_C$ are independent to $\mathbf{r}_S$ and $\mathbf{X}_S$ is independent to $\mathbf{v}_C$, we could finally get Eq. 7.

$$p(\mathbf{v}_C|\mathbf{r}_S; \mathbf{X}_S, \theta, \mathbf{y}_C) \tag{4}$$

$$= \frac{p(\mathbf{r}_S|\mathbf{v}_C; \mathbf{X}_S, \theta, \mathbf{y}_C)p(\mathbf{v}_C|; \mathbf{X}_S, \theta, \mathbf{y}_C)}{p(\mathbf{r}_S|; \mathbf{X}_S, \theta, \mathbf{y}_C)} \tag{5}$$

$$\propto p(\mathbf{r}_S|\mathbf{v}_C; \mathbf{X}_S, \theta, \mathbf{y}_C)p(\mathbf{v}_C|; \mathbf{X}_S, \theta, \mathbf{y}_C) \tag{6}$$

$$\propto p(\mathbf{r}_S|\mathbf{v}_C, \mathbf{X}_S)p(\mathbf{v}_C|\theta, \mathbf{y}_C) \tag{7}$$

Therefore, the posterior distribution of relation prototypes could be factorized as the likelihood of supporting instances and the prior knowledge of relation prototypes, which is derived from the definition text and pre-trained prototype encoder. $p(\mathbf{v}_C|\theta, \mathbf{y}_C)$ is the prior distribution for relation prototypes and each relation is assumed to follow a Gaussian distribution independently (i.e., $\mathcal{N}(\mathbf{v}_c|\text{ProtoEnc}_\theta(\mathbf{y}_c), I)$). $p(\mathbf{r}_S|\mathbf{X}_S, \mathbf{v}_C)$ is the likelihood of supporting instances computed by the softmax function, where dot product is used to compute the similarity of instance embeddings and the final relation prototypes.

Following the implementation of (Qu et al., 2020), we sample multiple prototypes for estimating the posterior distribution and each sampling is obtained via multiple stochastic updates. The optimization process is end-to-end and further details can be found in the paper (Qu et al., 2020).

Given a query instance $\mathbf{X}_q$ and a list of candidate relations C whose prototype vectors are denoted as $\mathbf{v}_C$, the relation distribution of the query instance over candidate relations can be computed by a softmax function as follows:

$$p(r_q|\mathbf{X}_q, \mathbf{v}_C) = \frac{\exp(\mathcal{E}(\mathbf{X}_q) \cdot \mathbf{v}_q)}{\sum_{c \in C} \exp(\mathcal{E}(\mathbf{X}_q) \cdot \mathbf{v}_c)}, \tag{8}$$

where $\mathcal{E}$ is the instance encoder (like BERT or RelEnc_ϕ).

4 Experiments

Pre-training details To run experiments with limited computation resources, we use $\text{BERT}_{\text{BASE}}$ [2] as the backbone of instance encoder and a four-layer TinyBERT [3] for the prototype encoder. We obtain a large-scale distantly-labeled dataset by processing the largest available alignments - T-REx (Elsahar et al., 2018), which align the documents from Wikipedia and triplets from Wikidata. We removed the relations that do not have the textual definition, i.e., all the relations in the distantly labeled data have the textual definition. After removing the repetitive instances, the final dataset contains 636 relations and $\sim$ 8M instances. We get the relation name and descriptions from Wikidata.

The hyper-parameters we used during the pre-training process are: batch size is 96; the number of training epochs is 3; optimizer is Adam with a learning rate of 1e-4, which decays by 0.8 per epoch. Our pre-training takes about 68 hours on four V100 GPUs.

4.1 Few-Shot Relation Learning

Dataset and evaluation metrics We adopt two few-shot RC datasets: FewRel 1.0 (Han et al., 2018) and FewRel 2.0 (Gao et al., 2019b). FewRel 1.0 contains 100 relations split into training, validation and test sets with respectively 64, 16 and 20 relations without overlapping. Each relation has 700 instance sentences from Wikipedia. FewRel 2.0 is constructed to evaluate models for domain adaption challenge and its validation data and test data are from biomedical domain.

The typical N-way K-shot setting means each evaluation episode will sample N relations, each of which has K labeled instances, and some query instances. The models are asked to classify query instances into the sampled N relations given the N$\times$K labeled data. Accuracy is used to evaluate the classification performance. Note that we exclude entity pairs in test set of FewRel 1.0 that appear in the pre-training dataset to avoid data leakage.

Baselines We choose the following representative and strong baselines for comparison. (1) few-shot learning methods relying only on given training instances: **ProtoNet** (Snell et al., 2017), **Pair** (Han et al., 2018) ; (2) methods that integrate extra information: **REGRAB** (Qu et al., 2020) uses structural relationship between different relations in KBs, **MIML** (Dong et al., 2020) uses class semantic information from relation names; (3) RC-oriented pre-training methods that provide new instance encoders: **MTB** (Baldini Soares et al., 2019) constructs sentence pairs as training samples based on entity linking techniques, **COL** (Ding et al., 2021) uses relation prototypes for regularization assuming they are uniformly dispersed in a unit ball. **CP** (Peng et al., 2020) conducts contrastive pre-training over distantly labeled data. CP achieves state-of-the-art performance. For fair com-

[2] https://huggingface.co/bert-base-uncased

[3] https://github.com/huawei-noah/Pretrained-Language-Model/tree/master/TinyBERT

Method	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
ProtoNet†	80.68	89.60	71.48	82.89
Pair†	88.32	93.22	80.63	87.02
REGRAB†	90.30	94.25	84.09	89.93
MIML†	92.55	96.03	87.47	93.22
MTB†	89.09	95.32	82.17	91.73
COL†	92.51	95.88	86.39	92.76
BERT-EM	88.12	95.55	83.44	91.19
CP-RI	93.03	96.10	88.69	93.09
REGRAB+Proto	90.72	94.87	84.44	90.43
REGRAB+Rel+Proto	93.20	96.50	87.32	92.80
Rel+Proto	**96.69**	**97.52**	**93.43**	**94.64**

Table 2: Classification accuracies (%) on FewRel 1.0 test set. Results with † are reported as published.

Method	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
FewRel 1.0				
MTB	91.10	95.40	84.30	91.80
CP	95.10	97.10	91.20	94.70
CP+Proto	**96.64**	**98.14**	**93.76**	**96.48**
FewRel 2.0 Domain Adaptation				
MTB	74.70	87.90	62.50	81.10
CP	79.70	84.90	68.10	79.80
CP+Proto	**83.11**	**90.80**	**73.02**	**83.08**

Table 3: Accuracy (%) on FewRel datsets. Our prototype learning method improves previous best pre-trained RC model (CP) and the reported baseline results are from their paper (Peng et al., 2020).

parison, we re-run CP model with our pre-training dataset and denote the model as **CP-RI**. We also implement a baseline **BERT-EM** that is optimized by the cross-entropy loss on every instance during pre-training and uses exemplar comparison for few-shot classification (Baldini Soares et al., 2019).

Our model and variants Based on the Bayesian meta-learning approach (Qu et al., 2020), we present two kinds of implementations:
(1) Models denoted as REGRAB+* construct a global relation graph with relation embeddings. Detailed ablation analysis is conducted on those variations. **REGRAB+Proto** only replaces the original relation embeddings with definition representations by our ProtoEnc. Based on this, **REGRAB+Rel+Proto** further replaces original instance encoder with our pre-trained RelEnc.
(2) Other models that discard the global graph construction as introduced in section 3.2. Those models are denoted as "instance encoder" +Proto, e.g., **Rel+Proto** uses both of our pre-trained encoders, **CP+Proto** applies the public instance encoder CP (Peng et al., 2020) to our approach.

Results and analysis Table 2 shows the performance of different models. Our model Rel+Proto achieves state-of-the-art performance over strong baselines. This demonstrates the advantages of leveraging both extra knowledge from the definition text and pre-trained instance encoder.

Table 3 shows that our approach can further improve other strong pre-trained instance encoder CP by a large margin on both FewRel 1.0 and FewRel 2.0, verifying the wide applicability of our approach. We empirically found that REGRAB cannot be easily applied to FewRel 2.0 and MIML often produced unstable results.

Ablation analysis of the pre-trained prototype encoder and instance encoder is conducted with REGRAB. Compared with REGRAB, REGRAB+Proto consistently improves the performance on four settings, indicating the semantic relationship by our prototype encoder is more effective than the structural relationship from KBs. Compared with REGRAB+Proto, REGRAB+Rel+Proto achieves obvious improvements over four settings, showing the importance of pre-trained instance encoder. Besides, removing the operation of explicitly constructing a global relation graph, Rel+Proto further improves REGRAB+Rel+Proto, verifying our assumption that relation definitions could imply semantic similarities of different relations, and they provide a better way to construct representations for relations than through the relationships between them.

From Table 2, we observe CP-RI, whose instance encoder is pre-trained with the open-sourced code on our pre-training dataset, performs worse than the published results of CP, indicating our pre-training dataset may be noisier. Compared with CP-RI, BERT-EM performs more poorly, implying that the model using the label of each instance may suffer from noisy labels. Thus our design to leverage MIL algorithm is necessary.

4.2 Generalizability in Zero-Shot Relation Learning

To test how generalizable the prototype encoder is, we apply it to unseen relations in zero-shot RC to construct the prototype representation from their definitions. We test how effective such a relation prototype representation is.

# relations	2	5	10	15	25
BERT-Def	48.98	19.83	9.34	7.98	3.98
ZS-BERT	**87.60**	57.84	38.54	30.73	24.05
BERT-Proto	84.01	**65.60**	**52.46**	**44.80**	**34.72**

Table 4: Zero-shot classification accuracies (%) on NYT-25 with increasing candidate relations.

# relations	2	5	8	10
BERT-Def	45.68	21.92	11.80	12.10
ZS-BERT	50.00	23.32	13.60	13.00
BERT-Proto	**63.78**	**32.82**	**21.40**	**18.00**

Table 5: Zero-shot classification accuracies (%) on PubMed-10 with increasing candidate relations.

Dataset and evaluation setting We use two datasets, NYT-25 and PubMed-10, whose relations are nonoverlapping with FewRel training data and are obtained from the FewRel website[4]. NYT-25 contains 25 relations from Wikidata and its sentences are from New York Times; PubMed-10 contains 10 relations and both its relations and instances are from the biomedical domain. Each relation in the two datasets has 100 manually labeled instances. Relations from Wikidata have names and descriptions as the definition text, while the relations in PubMed-10 have only relation names as the definition text.

For N-way zero-shot RC setting, the classification difficulty is increased with the increase of candidate relation number N. We vary N from 2 to the max number. In each setting, a candidate set is made of N-1 negative relations and a positive relation.

Compared models We denote our combined RC model by the pre-trained prototype encoder and instance encoder as **BERT-Proto**, whose pre-training data exclude relations in NYT-25. To verify if the pre-trained language models understand the definition text for RC, we present a baseline **BERT-Def** that has the same model structure as BERT-Proto but does not perform the joint training with the instance encoder and prototype encoder.

We also present a competitive baseline **ZS-BERT** (Chen and Li, 2021) that classifies sentences based on embedding similarity. Similar to BERT-Proto, ZS-BERT learns two functions to project instances and relation definitions into an embedding space. The difference is that it uses a fixed pre-trained model, sentenceBERT (Reimers and Gurevych, 2019), to encode the definition text into an attribute vector. The attribute vectors are used to regularize the instance encoder during training and to compare with instance embeddings for classification during testing. We train ZS-BERT with FewRel training data for adapting to RC tasks.

[4]https://github.com/thunlp/FewRel

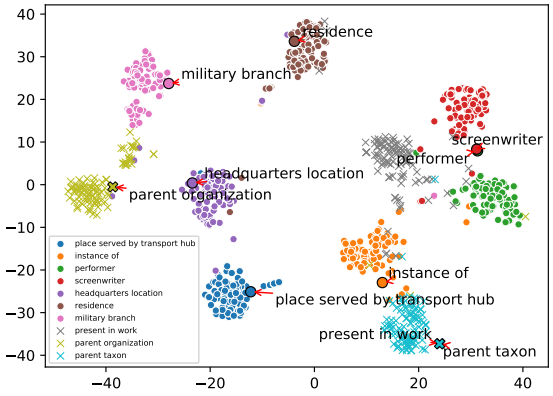

Figure 2: The t-SNE visualization of instance embeddings and definition representations (indicated by arrows) by our pre-trained BERT-Proto on selected relations. Each relation are represented by one color and the relations seen in pre-training are in circles and the unseen are crosses.

Zero-shot results and discussion From Table 4 and Table 5, we see BERT-Proto outperforms ZS-BERT, showing that our model is competitive for zero-shot RC even applied to a new domain and under the most difficult setting, while ZS-BERT performs slightly better than random on PubMed-10. We believe the advantages are mainly generated by the pre-training technique that helps learn a general mapping function, i.e. the prototype encoder. Note that BERT-Def presents almost random results, implying the ineffective encoding of relation definitions when no joint training is performed.

4.3 Visualization

To figure out how relation representations are distributed in the semantic space, we visualize the representations by our pre-trained instance encoder and prototype encoder with TSNE (Van der Maaten and Hinton, 2008) method. We also present relations of seen and unseen in the same space. The seen relations are selected from the FewRel training data and the unseen are from NYT-25.

From Figure 2, we observe that: (1) The definition representations of seen relations are close to instance embeddings of the same relation, show-

ing that the prototype encoder could effectively map the definition text to its prototype vector. For some unseen relations, their own definition representations could also be the closest to their instance embeddings, providing effective prior knowledge for RC. (2) The distance between different relations reflects their similarities, which are mainly determined by contextual features, such as entity types, sentence structures and context semantics. For example, the relations of "performer", "screenwriter" and "present in work" all express the relationship between a person and an artwork. Their instance embeddings are much closer to each other than with the relations about locations such as "headquarters location", even though "present in work" is the unseen relation. This shows that pre-trained instance encoder could extract effective contextual features for relation representations and build a meaningful semantic space to guide our prototype encoder learning. (3) For new-emerging relations, their prototype vectors are determined by both the definition text and labeled instances, so as to benefit from the possible connections with seen relations through pre-trained prototype encoder and instance encoder.

5 Related Work

Relation classification (RC) is pivotal for natural language understanding and has been studied for a long time (Chieu and Ng, 2002). Supervised machine learning approaches achieve remarkable progress on RC (Kambhatla, 2004; Hendrickx et al., 2009), but rely on high-quality labeled data. To relieve the heavy burden of manual annotation, researchers study RC under distant supervision (Mintz et al., 2009) or few-shot RC (Han et al., 2018; Gao et al., 2019b). The former focuses on robust classifier training with automatically labeled noisy data (Lin et al., 2016; Li et al., 2020). Our work belongs to the latter, which aims to learn general knowledge transferable to new relations.

Few-shot learning methods have been well studied for image classification (Ravi and Larochelle, 2017) and some classical approaches such as prototype network (Snell et al., 2017), model-agnostic meta-learning (Finn et al., 2017) have been applied for RC (Han et al., 2018). Two types of efforts have been devoted to improving few-shot RC. Firstly, some approaches (Ye and Ling, 2019; Gao et al., 2019a; Wang et al., 2020; Han et al., 2021; Ren et al., 2020; Ohashi et al., 2021) design specific model architectures such as using attention mechanism to model complex interactions between labeled instances. However, these approaches are still limited when the few labeled instances are atypical and does not reflect the general patterns of the relation. Secondly, researchers leverage extra information to complement the insufficient labeled data (Qu et al., 2020; Dong et al., 2020). Our method belongs to this line.

Some methods (Qu et al., 2020; Zhang et al., 2021) leverage extra knowledge from KBs but they cannot deal with relations not covered by the KBs, showing limited applicability. Similar to our work, some studies (Dong et al., 2020; Yang et al., 2020) use relation names or descriptions as extra information. They design specific modules to integrate the extra information, but the whole few-shot RC model suffers from over-fitting and unstable performance due to the limited number of training relations. In contrast, our method adopts pre-training techniques to learn a general mapping function, which is more applicable and proven effective for domain adaptation.

Some recent studies (Baldini Soares et al., 2019; Ding et al., 2021; Peng et al., 2020) conduct RC-oriented pre-training to learn a general-purpose instance encoder. Such instance encoders improve RC on both supervised and few-shot learning settings. We adopt the same idea of pre-training but focus on learning the general function of mapping relation definitions to the prototype space for relation classification.

6 Conclusion

This paper studies prototypical representation learning for few-shot relation classification, and the key idea is to encode the definition text as prior knowledge to help classify new relations. We proposed to train a general-purpose prototype encoder that could encode the definition text of any relation into the prototype space. An instance encoder and the prototype encoder are trained jointly with a multi-instance learning method on distantly labeled data. Applying our prior prototypes with a Bayesian meta-learning approach, our method outperforms previous state-of-the-art models by using pre-trained instance encoder on two datasets, verifying its wide applicability. Our prototype model also presents competitive performance on the zero-shot learning setting.

Acknowledgements

We would like to thank Meng Qu, Benyou Wang and other teammates of Yuyang Zhang for their insightful suggestions and valuable discussion. We also thank all the reviewers for the generous comments. This research work is partly supported by the National Natural Science Foundation of China under Grant No. 62025208.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

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*, pages 3470–3479, Online. Association for Computational Linguistics.

Hai Leong Chieu and Hwee Tou Ng. 2002. A maximum entropy approach to information extraction from semi-structured and free text. In *Eighteenth National Conference on Artificial Intelligence*, page 786–791, USA. American Association for Artificial Intelligence.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Xiaobin Wang, Yao Fu, Guangwei Xu, Rui Wang, Pengjun Xie, Ying Shen, Fei Huang, Hai-Tao Zheng, and Rui Zhang. 2021. Prototypical representation learning for relation extraction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Bowen Dong, Yuan Yao, Ruobing Xie, Tianyu Gao, Xu Han, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Meta-information guided meta-learning for few-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1594–1605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019a. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6407–6414. AAAI Press.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019b. FewRel 2.0: Towards more challenging few-shot relation classification. 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 6250–6255, Hong Kong, China. 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, Brussels, Belgium. Association for Computational Linguistics.

Yi Han, Linbo Qiao, Jianming Zheng, Zhigang Kan, Linhui Feng, Yifu Gao, Yu Tang, Qi Zhai, Dongsheng Li, and Xiangke Liao. 2021. Multi-view interaction learning for few-shot relation classification. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, page 649–658, New York, NY, USA. Association for Computing Machinery.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian

Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99, Boulder, Colorado. Association for Computational Linguistics.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3060–3066. AAAI Press.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *ACL*, pages 178–181.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8269–8276. AAAI Press.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Sora Ohashi, Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. 2021. Distinct label representations for few-shot text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 831–836, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Meng Qu, Tianyu Gao, Louis-Pascal A. C. Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7867–7876. PMLR.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Haopeng Ren, Yi Cai, Xiaofeng Chen, Guohua Wang, and Qing Li. 2020. A two-phase prototypical network model for incremental few-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1618–1629, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Yingyao Wang, Junwei Bao, Guangyi Liu, Youzheng Wu, Xiaodong He, Bowen Zhou, and Tiejun Zhao. 2020. Learning to decouple relations: Few-shot relation classification with entity-guided attention and confusion-aware training. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5799–5809, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kaijia Yang, Nantao Zheng, Xinyu Dai, Liang He, Shujian Huang, and Jiajun Chen. 2020. Enhance prototypical network with text descriptions for few-shot relation classification. In *CIKM '20: The 29th*

ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, pages 2273–2276. ACM.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2872–2881, Florence, Italy. Association for Computational Linguistics.

Jiawen Zhang, Jiaqi Zhu, Yi Yang, Wandong Shi, Congcong Zhang, and Hongan Wang. 2021. Knowledge-enhanced domain adaptation in few-shot relation classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '21, page 2183–2191, New York, NY, USA. Association for Computing Machinery.