MapRE: An Effective Semantic Mapping Approach for Low-resource Relation Extraction

Manqing Dong, Chunguang Pan, and Zhipeng Luo

DeepBlue Technology (Shanghai) Co., Ltd

`dongmanqing@gmail.com, luozp@deepblueai.com`

Abstract

Neural relation extraction models have shown promising results in recent years; however, the model performance drops dramatically given only a few training samples. Recent works try leveraging the advance in few-shot learning to solve the low resource problem, where they train label-agnostic models to directly compare the semantic similarities among context sentences in the embedding space. However, the label-aware information, i.e., the relation label that contains the semantic knowledge of the relation itself, is often neglected for prediction. In this work, we propose a framework considering both label-agnostic and label-aware semantic mapping information for low resource relation extraction. We show that incorporating the above two types of mapping information in both pretraining and fine-tuning can significantly improve the model performance on low-resource relation extraction tasks.

Support Instances	
(a) mountain range	It is approximately 9 km away from <u>Mount Korbu</u>, the tallest mountain of the **Titiwangsa Mountains**.
	The <u>San Ysidro Mountains</u> are part of the **Peninsular Ranges System**.
(b) head of government	One of the umpires was **Edmund Barton**, who became <u>Australia</u>'s first prime minister.
	Keith Burdette was the Secretary of Commerce for the state of <u>West Virginia</u> under the administration of Governor **Earl Ray Tomblin**.
Query Instance	
(a) or **(b)** ?	Senator Patrick Leahy and <u>Vermont</u> Governor **Phil Scott**.

Figure 1: Example for a 2-way 2-shot relation extraction task. The entities with <u>underlines</u> are head entities, and the entities in **bold** are tail entities. The target is to predict the relation between the head and the tail entities for a given query instance.

1 Introduction

Relation Extraction (RE), which aims at discovering the correct relation between two entities in a given sentence, is a fundamental task in NLP (Gao et al., 2019). The problem is generally regarded as a supervised classification problem by training on large-scale labelled data (Zhang et al., 2017). Neural models, e.g. RNN-based methods (Zhou et al., 2016), or more recently, BERT-based methods (Soares et al., 2019; Peng et al., 2020), have shown promising results on RE tasks, where they achieve state-of-the-art performance or even comparable with human performance on several public RE benchmarks.

Despite the promising performance of the existing neural relation classification frameworks, recent studies (Han et al., 2018) found that the model performance drops dramatically as the number of instances for a relation decreases, e.g., for long-tail relations. An extreme condition is few-shot relation extraction, where only few support examples are given for the unseen relations, see Figure 1 as an example.

A conventional way to solve the data deficiency problem of RE is distant supervision (Mintz et al., 2009; Hu et al., 2019), which assumes same entity-pairs have same relations in all sentences so that to augment training data for each relation from external corpus. However, such an approach can be rough and noisy since same entity-pairs may have different relations given different contexts (Ye and Ling, 2019; Peng et al., 2020). Besides, distant supervision may exacerbate the long-tail problem in RE for the relations with only a few instances.

Inspired by the advances in few-shot learning (Nichol et al., 2018; Mishra et al., 2018), recent attempts adopt metric-based meta-learning frameworks (Snell et al., 2017; Koch et al., 2015) to few-
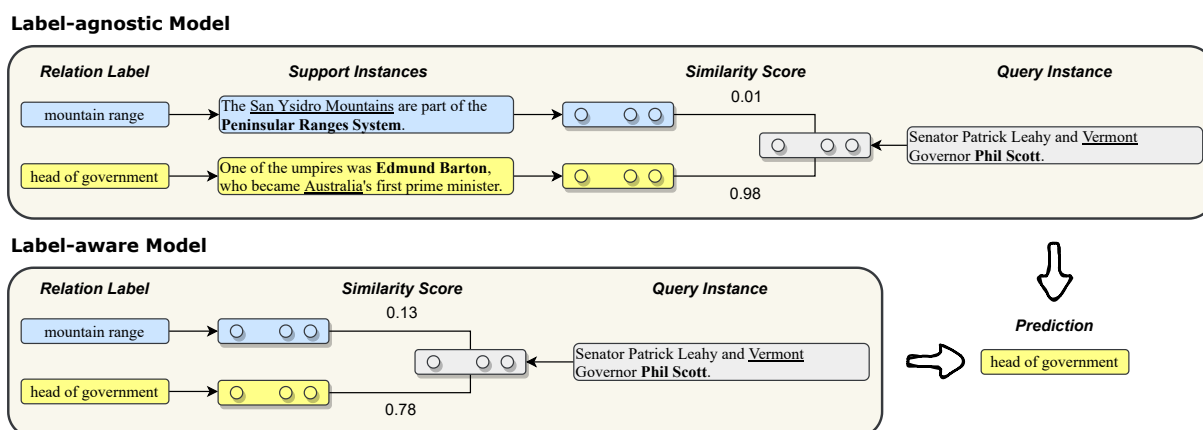
Label-agnostic Model

| Relation Label | Support Instances | Similarity Score | Query Instance |

mountain range → The San Ysidro Mountains are part of the **Peninsular Ranges System**. → 0.01

head of government → One of the umpires was **Edmund Barton**, who became Australia's first prime minister. → 0.98

Senator Patrick Leahy and Vermont Governor **Phil Scott**.

Label-aware Model

| Relation Label | Similarity Score | Query Instance |

mountain range → 0.13

head of government → 0.78

Senator Patrick Leahy and Vermont Governor **Phil Scott**.

Prediction

head of government

Figure 2: Examples for label-agnostic and label-aware models to relation extraction.

shot RE tasks (Gao et al., 2019; Ye and Ling, 2019). The key idea is to learn a label-agnostic model that compares the similarity between the query and support samples in the embedding space (see Figure 2 for an example). In this way, the target for RE changes from learning a general and accurate relation classifier to learning a projection network that maps the instances with the same relation into close regions in the embedding space.

Recent metric-based relation extraction frameworks (Peng et al., 2020; Soares et al., 2019) achieve the state-of-the-art on low-resource RE benchmarks. However, these approaches are not applicable when there is no support instance for the unseen relations, since they need at least one support example to provide the similarity score of a given query sentence. Besides, most of the existing few-shot RE frameworks neglect the relation label for prediction, whereas the relation label contains valuable information that implies the semantic knowledge between the two entities in a given sentence. In this work, we propose a semantic mapping framework, MapRE, which leverages both label-agnostic and label-aware knowledge. Specifically, we hope two types of matching information, i.e., the context sentences and their corresponding relation label (label-aware) as well as the context sentences denoting the same relations (label-agnostic), to be close in the embedding space. We show that leveraging the label-agnostic and label-aware knowledge in pretraining improves the model performance in low-resource RE tasks, and utilizing the two types of information in fine-tuning can further enhance the prediction results. With the contribution of the label-agnostic and label-aware information in both pretraining and fine-tuning, we

achieve the state-of-the-art in nearly all settings of the low-resource RE tasks (e.g., we improve the SOTA on two 10-way 1-shot datasets by 1.98% and 2.35%, respectively).

Section 2 summarizes the related work and briefly introduces the difference between our proposed method and the others. Section 3 illustrates the pretraining framework with considering both label-agnostic and label-aware information. We evaluate the proposed model on supervised RE in Section 4 and few & zero-shot RE in Section 5, and leave concluding remarks in Section 6.

2 Related Work

Meta-learning One branch of meta-learning is optimization-based frameworks (Nichol et al., 2018), e.g. model-agnostic meta-learning (MAML) (Finn et al., 2017), which learn a shared parameter initialization across training tasks to initialize the model parameters of testing tasks. However, a single shared parameter initialization cannot fit diverse task distribution (Hospedales et al., 2020); besides, the gradient updating strategies for the sharing parameters are complex and will take more computation resources. Metric-based meta-learning approaches (Snell et al., 2017; Koch et al., 2015) learn a projection network that maps the support and the query samples into the same semantic space to compare the similarities. The metric-based approaches are non-parametric, easier for implementation, and less computationally expensive; they have shown better performance than the optimization-based approaches on a series of few-shot learning tasks (Triantafillou et al., 2019), thus have been widely used in recent few-shot RE frameworks (Ye and Ling, 2019).

2695

Few-shot RE   Prototypical network (Snell et al., 2017) is probably the most widely used metric-based meta-learning framework for few-shot RE. It learns a prototype vector for each relation with a few examples, then compares the similarity between the query instance and the prototype vectors of the candidate relations for prediction (Han et al., 2018). For example, Gao et al. (2019) proposed hybrid attention-based prototypical networks to handle noisy training samples in few-shot learning. Ye and Ling (2019) further propose a multi-level matching and aggregation network for few-shot RE. Recent studies (Soares et al., 2019; Peng et al., 2020) also suggest the effectiveness of applying the metric-based approaches on pretrained models (Devlin et al., 2019), where optimizing the matching information between the support and query instances in embedding space obtained from the pretrained models can improve the model performance on the few-shot RE tasks. However, the metric-based approaches are not applicable for zero-shot learning scenarios, since they need at least one example for each support instance. To fill in this gap, we propose a semantic mapping framework that leverages both label-aware and label-agnostic information for relation extraction.

Zero-shot learning   An extreme condition of few-shot learning is zero-shot learning, where there is no instance provided for the candidate labels. A standard approach is to match the inputs with the predefined label vectors (Xian et al., 2017; Rios and Kavuluru, 2018; Xie et al., 2019), which assumes the label vectors take an equally crucial role as the representations of the support instances (Yin et al., 2019). The label vectors are often obtained by pretrained word embeddings such as GloVe embeddings (Pennington et al., 2014) and will be directly used for prediction (Rios and Kavuluru, 2018; Wang et al., 2018). For example, Xia et al. (2018) study the zero-shot intent detection problem: they use the sum of the word embeddings as the representation for each intent label, and the prediction is based on the similarity between the inputs and the intent representations. Zhang et al. (2019) enrich the label representation with external knowledge such as the label description and the label hierarchy. However, the label representations are fixed in most existing zero-shot learning approaches, which will lead the input-representation-learning model overfit to the label representations. Besides, the superiority of the label-aware models are somewhat

limited to zero-shot learning scenarios – according to our experimental results on FewRel dataset (Han et al., 2018) (refer to Table 3), the label-agnostic models perform better than the label-aware models once given support examples. To overcome the above issues, we propose a pretraining framework considering both label-aware and label-agnostic information for low-resource RE tasks, where the label representations are obtained via a learnable BERT-based (Devlin et al., 2019) model.

RE with external knowledge   Some works try leveraging external knowledge to address the low-resource RE tasks. For example, Cetoli (2020) formalize RE as a question-answering task: they fine-tune on a BERT-based model that pretrained on SQUAD (Rajpurkar et al., 2016) then use the BERT-based model to generate the prediction for the relation label. Qu et al. (2020) follows the key idea of zero-shot learning by introducing knowledge graphs to obtain the relation label representations. Both works show good performance on low-resource RE tasks while need extra knowledge to fine-tune the framework. However, the extra knowledge is not always available for all cases. In this work, we focus on enhancing the generalization ability of the model without referring to external knowledge, where we obtain SOTA performance on most low-resource RE benchmarks.

3   Pretraining with Semantic Mapping

3.1   Preliminary

Task definition   Each instance $x = (c, p_h, p_t)$ includes a triple of context sentence tokens $c = [c_0 \ldots c_m]$ and the head and tail entity positions, where $c_0 = $ `[CLS]` and $c_m = $ `[SEP]` are two special tokens denoting the start and the end of the sequence, $p_h = (p_h^s, p_h^e)$ and $p_t = (p_t^s, p_t^e)$ are the indices for head and tail entities with $0 \leq p_h^s \leq p_h^e \leq m$ and $0 \leq p_t^s \leq p_t^e \leq m$. For a **supervised learning** problem, given $\mathcal{N}$ relations $\mathcal{R} = \{r_1, \ldots, r_\mathcal{N}\}$ and the instances for each relation, our target is to predict the correct relations $r \in \mathcal{R}$ for the testing instances. For a N-**way** K-**shot learning** problem, given support instances $\mathcal{S} = \{x_r^j | r \in R, j = \{1, \ldots, K\}\}$ with N relations $R = \{r_1, \ldots, r_N\}$ and K examples for each relation, our target is to predict the correct relation $r \in R$ of the entities for a query instance x_q.
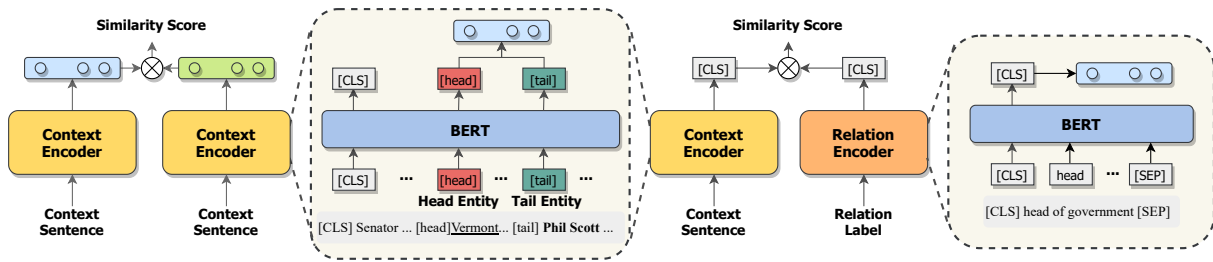
Figure 3: The pretraining framework for MapRE, where we consider both label-agnostic and label-aware semantic mapping information in training the whole framework.

Differences between supervised RE and few-shot RE There are several differences between supervised RE and few-shot RE. First, supervised RE tries to learn a $\mathcal{N}$-way relation classifier that could fit all training instances, while few-shot RE tries to learn a N-way classifier (normally $N \ll \mathcal{N}$) by learning from only a few samples. Second, the training and testing data for few-shot RE have no intersection in relation types, i.e. during the testing phase, the model is required to generalize to unseen labels with only a few samples.

Pretraining for low-resource RE Recent studies (Soares et al., 2019; Peng et al., 2020) find that pretrain the model with contrastive ranking loss (Sohn, 2016; Oord et al., 2018) can improve the generalization ability of the model in low-resource RE tasks. The key idea is reducing the semantic gap between the instances with the same relations in the embedding space. In other words, instances with same relations should have similar representations.

3.2 Matching Sample Formulation

Following the idea of Soares et al. (2019) and Peng et al. (2020), we construct mapping functions for relation extraction. Specially, we hope two types of matching samples to be close in the semantic space: **1)** the context sentences denoting same relations, and **2)** the context sentences and the corresponding relation labels.

Given a knowledge graph $\mathcal{G}$ containing extensive examples of relation triples $T = (h, r, t), T \in \mathcal{G}$, we will first randomly sample the relation triples; then, sentences containing the same head h and tail h entities and denoting the same relation r will be sampled from the corpus for this triple, i.e. $\{x = (c, p_h, p_t)|x \in T\}$. Specially, at each sampling step, N triples with N different relations $\{r_i|i = 1, \ldots, N\}$ are sampled from $\mathcal{G}$. For each triple $T = (h, r, t)$, a pair of sentences

$\{(x_A, x_B)|x_A, x_B \in T\}$ will be extracted from the corpus, so that we have $2N$ sentences in total. For each sentence, we take a similar strategy as in (Soares et al., 2019; Peng et al., 2020) that a probability of 0.7 is set to mask the entity mentions when fed into the sentence context encoder to avoid the model memorizes the entity mentions or shallow cues during pretraining.

Suppose the sentence context encoder is denoted as f_{CON}, and the relation encoder is denoted as f_{REL}, we hope the semantic gap between each pair of sentences that denote for same relation, i.e., $d(f_{\text{CON}}(x_A), f_{\text{CON}}(x_B)), x_A, x_B \in T$, and the semantic gap between the context sentences and their relation labels, i.e., $d(f_{\text{CON}}(x_A), f_{\text{REL}}(r))$ and $d(f_{\text{CON}}(x_B), f_{\text{REL}}(r))$, to be small in in the embedding space. Figure 3 shows an example of the matching samples, where both the context encoder f_{CON} and the relation encoder f_{REL} are a BERT$_{\text{BASE}}$ model (Devlin et al., 2019). According to Soares et al. (2019), the concatenation of the special tokens (i.e., [head] and [tail]) at the start of the head and the tail entities, provides best performance for downstream relation classification tasks, thus we take $f_{\text{CON}}(x)[[\text{head}], [\text{tail}]]$ to compare the label-agnostic similarities between sentences. We use the embedding of the special [CLS] token in the context encoder $f_{\text{CON}}(x)[\text{CLS}]$ to denote the label-aware information for the context sentence, and the [CLS] token in relation encoder $f_{\text{REL}}(r)[\text{CLS}]$ to denote the relation representation. This is to avoid the override of the memorization in the head and tail special tokens and to improve the generalization ability of the sentence context encoder. Another reason is the dimension of the concatenation [[head], [tail]] and the [CLS] token does not match, which needs extra parameter space to optimize. The extra parameter space can be easily over-fitted to training data and produce biased prediction performance

when distinct distribution between the training and testing sets exists.

3.3 Training Objectives

At each sampling step, we have $2N$ sentences with N pairs of sentences denoting N distinct relations. For each sentence x, we get its context embedding $\mathbf{u} = f_{\text{CON}}(x)[\text{[head]}, \text{[tail]}]$ and its label-aware embedding $\mathbf{w} = f_{\text{CON}}(x)[\text{CLS}]$. The corresponding relation representation is obtained by $\mathbf{v} = f_{\text{REL}}(r)[\text{CLS}]$. We use contrastive training (Oord et al., 2018; Chen et al., 2020) to train the MapRE, which pulls the 'neighbors' together and pushes 'non-neighbors' apart. Specifically, we consider three training objectives to optimize the whole framework.

Contrastive Context Representation Loss We follow the work by (Peng et al., 2020) to calculate the contrastive loss of the sentence context representations [1]. For example, for sentence x_A^i from the positive pair (x_A^i, x_B^i) (both represents relation r_i), any sentence in other pairs forms the negative pair with x_A^i, i.e., (x_A^i, x_B^j) and (x_A^i, x_A^j), for $1 \leq j \leq N, j \neq i$ (examples are shown in Figure 4). Then for x_A^i, we maximize $\frac{\exp(\mathbf{u}_A^{i\top}\mathbf{u}_B)}{\Sigma_j \exp(\mathbf{u}_A^{i\top}\mathbf{u}_B^j) + \Sigma_j \exp(\mathbf{u}_A^{i\top}\mathbf{u}_A^j)}$. Sum the log loss for each sentence, we get the contrastive context representation loss as $\mathcal{L}_{CCR}$.
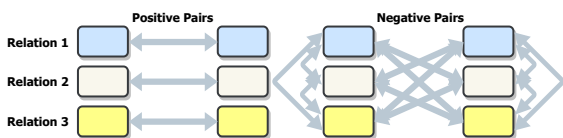

Figure 4: Examples for positive and negative sentence context representation pairs.

Contrastive Relation Representation Loss We also calculate the contrastive loss between the label-aware representation $\mathbf{w}$ and the relation representations $\mathbf{v}$. For the $2N$ sampled sentences of N relations, we hope to minimize the loss

$$\mathcal{L}_{CRR} = -\Sigma_{i=1}^{2N} \log \frac{\exp(\mathbf{w}_i^\top \mathbf{v}_i)}{\Sigma_{j=1}^N \exp(\mathbf{w}_i^\top \mathbf{v}_j)}. \quad (1)$$

Masked Language Modeling (MLM) We also consider the conventional Masked Language Mod-

[1] https://kevinmusgrave.github.io/pytorch-metric-learning/losses/#ntxentloss

eling objective (Devlin et al., 2019), which randomly masks tokens in the inputs and predicts them in the outputs to let the context encoder engaging more semantic and syntactic knowledge. Denoting the loss by $\mathcal{L}_{MLM}$, the overall training objective is

$$\mathcal{L} = \mathcal{L}_{CCR} + \mathcal{L}_{CRR} + \mathcal{L}_{MLM} \quad (2)$$

We pretrain the whole framework on Wikidata (Vrandečić and Krötzsch, 2014) with a similar strategy as in (Peng et al., 2020), where we exclude any overlapping data between Wikidata and the datasets for further experiments.

4 Supervised RE

4.1 Fine-tuning for supervised RE

We obtain a pretrained context encoder f_{CON} and a relation encoder f_{REL} after the pretraining process mentioned above. A conventional way for supervised RE is to append several fully connected layers to the context encoder f_{CON} for classification, which can also be regarded as computing the similarity between the output of the context encoder and the one-hot relation label vectors (see the left part of Figure 5 as an example). Instead of using one-hot representation for the relation labels, we use the relation representation obtained from the relation encoder to calculate the similarities. An example is shown in the right part of Figure 5. The prediction is made by

$$\hat{r} = \arg\max_r \frac{\exp(\sigma(f_{\text{CON}}(x))^\top f_{\text{REL}}(r))}{\Sigma_{r' \in \mathcal{R}} \exp(\sigma(f_{\text{CON}}(x))^\top f_{\text{REL}}(r'))} \quad (3)$$

where σ stands for fully connected layers, $f_{\text{REL}}(r)$ denotes the embedding of the special token $[\text{CLS}]$ in the relation encoder, and $f_{\text{CON}}(x)$ here outputs the concatenation of the special tokens of head and tail entities $[\text{[head]}, \text{[tail]}]$. We optimize the context encoder, relation encoder, and the fully connected layers with cross-entropy loss for supervised training.

4.2 Evaluation

Datasets We evaluate on two benchmark datasets, ChemProt (Kringelum et al., 2016) and Wiki80 (Han et al., 2019) for supervised RE tasks. The former includes 56,000 instances for 80 relations, and the latter includes 10,065 instances for 13 relations.
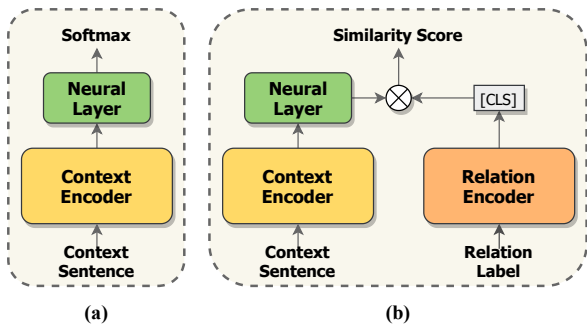
Figure 5: The frameworks for supervised learning. **Left**: uses fully connected layers to predict the probability distribution over all relations, used in BERT, MTB, CP, and MapRE-L. **Right**: compares the sentence context embedding with the relation representations, and regards the relation with highest similarity score as the prediction, used in MapRE-R.

Dataset	Method	1%	10%	100%
Wiki80	BERT	0.559	0.829	0.913
	MTB	0.585	0.859	0.916
	CP	0.827	0.893	0.922
	MapRE-L	**0.850**	**0.915**	**0.933**
	MapRE-R	**0.904**	**0.921**	**0.933**
ChemProt	BERT	0.362	0.634	0.792
	MTB	0.362	0.682	0.796
	CP	0.361	**0.708**	0.806
	MapRE-L	**0.424**	0.666	**0.813**
	MapRE-R	0.416	0.693	**0.814**

Table 1: Comparison results on supervised learning tasks in accuracy. 1%, 10%, and 100% denote the proportion of the training data used for fine-tuning.

Comparison Methods Numerous studies have been done for supervised RE tasks. Here we focus on low-resource RE and choose the following three representative models for comparison. 1) BERT (Devlin et al., 2019): the widely used pretrained model for NLP tasks. In this case, the model takes the embedding of the special tokens of the head and tail entities for prediction via several fully connected layers, similar to the conventional strategy shown in the left part of the Figure 5. 2) MTB (Soares et al., 2019): a pretrained framework for RE, which regards the sentences with the same head and tail entities as positive pairs. The fine-tuning strategy is same as in BERT. 3) CP (Peng et al., 2020): a pretrained framework that is analogous to MTB. The difference is that the model treats sentences with the same relations as positive pairs during the pretraining phase. The fine-tuning strategy is the same as BERT and MTB.

Comparison Results Table 1 shows the comparison results on the two datasets with training on

different proportions of the training sets. For our model, we consider the model performance with different fine-tuning strategies as shown in the left and right part in Figure 5. We denote the two variants as MapRE-L and MapRE-R. The detailed parameter settings can be found in the Appendix. We can observe that: 1) pretraining on the BERT with matching information (i.e., MTB, CP, and our MapRE) can improve the model performance on low-resource RE tasks; 2) comparing MapRE-L with CP and MTB, adding the label-aware information during pretraining can significantly improve the model performance, especially on extremely low-resource conditions, e.g., when only 1% of training sets are available for fine-tuning; and 3) comparing MapRE-R with MapRE-L, which also considers the label-aware information in fine-tuning, shows better and more stable performance in most conditions. Overall, the results suggest the importance of engaging the label-aware information in pretraining and fine-tuning to improve the model performance on low-resource supervised RE tasks.

5 Few & Zero-shot RE

5.1 Fine-tuning for few-shot RE

In the case of few-shot learning, the model is required to predict for new instances with only a few given samples. For a N-way K-shot problem, the support set $\mathcal{S}$ contains N relations that each is with K examples, and the query set contains Q samples that each belongs to one of the N relations. To fine-tune the model for few-shot RE, we construct the training set in a series of N-way K-shot learning tasks. For each task, the prediction for a query instance x_q is made by comparing the **label-agnostic mapping information**, i.e., the similarity between the query context sentence representation u_q and the support context sentence representation u_r, as well as the **label-aware mapping information**, i.e., the semantic gap between the query label-aware representation $w_q = f_{\text{CON}}(x_q)[\text{CLS}]$ and the relation label representation $v_r = f_{\text{REL}}(r)[\text{CLS}]$:

$$\hat{r} = \arg\max_r \frac{\exp(\alpha \cdot u_q^\top u_r + \beta \cdot w_q^\top v_r)}{\Sigma_{r' \in \mathcal{R}} \exp(\alpha \cdot u_q^\top u_{r'} + \beta \cdot w_q^\top v_{r'})} \quad (4)$$

$$u_r = \frac{1}{K}\Sigma_k u_r^k = \frac{1}{K}\Sigma_k f_{\text{CON}}(x_r^k)[[\text{head}],[\text{tail}]] \quad (5)$$
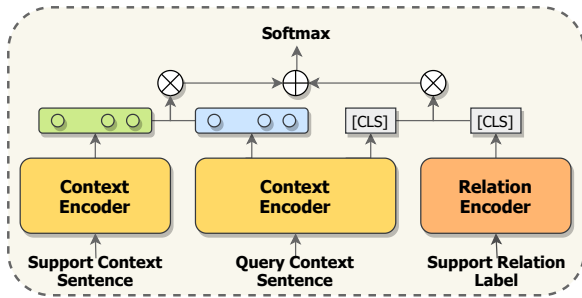
Figure 6: The framework for few-shot learning with MapRE. Both label-agnostic information, i.e., the matching information among the context sentence representations, and label-aware information, i.e., the semantic gap between the sentence label-aware representation and the relation label representation, are considered for fine-tuning.

where u_r is the prototype sentence representation for K support instances denoting relation r; α and β are two learnable coefficients controlling the contribution of the two types of semantic mapping information. An example of the few-shot learning framework is shown in Figure 6. We update both context encoder and relation encoder with cross-entropy loss on the generated N-way K-shot training tasks. We use dot product as the measurement of the similarities, which shows the best performance compared with other measurements. Details about the model settings can be found in the Appendix.

5.2 Evaluation

Datasets We evaluate the proposed method on two few-shot learning benchmarks: FewRel (Han et al., 2018) and NYT-25 (Gao et al., 2019). The FewRel dataset consists of 70,000 sentences for 100 relations (each with 700 sentences) derived from Wikipedia. There are 64 relations for training, 16 relations for validation, and 20 relations for testing. The testing dataset contains 10,000 query sentences that each is given N-way K-shot relation examples and has to be evaluated online (the labels for the testing set is not published). The NYT-25 dataset is a processed dataset by (Gao et al., 2019) for few-shot learning. We follow the preprocessing strategy by Qu et al. (2020) to randomly sample 10 relations for training, 5 for validating, and 10 for testing.

Comparison methods Many recent studies try employing the advances of meta-learning (Hospedales et al., 2020) to few-shot RE tasks. We consider the following representative

methods for comparison. 1) **Proto** (Han et al., 2018) is a work using Prototypical Networks (Snell et al., 2017) for few-shot RE. The model tries to find the prototypical vectors for each relation from supporting instances, and compares the distance between the query instance and each prototypical vector under certain distance metrics. Each instance is encoded by a BERT$_{\text{BASE}}$ model. 2) **BERT-pair** (Gao et al., 2019) is a BERT-based model that encodes a pair of sentences to a probability that the pair of sentences expressing the same relation. 3) **REGRAB** (Qu et al., 2020) is a label-aware approach that predicts the relations based on the similarity between the context sentence and the relation label. The relation label representation is initialized via an external knowledge graph, where a Bayesian meta-learning approach is further used to infer the posterior distribution of the relation representation. The representation of the context sentence is learned by a BERT$_{\text{BASE}}$ model. 4) **MTB** (Soares et al., 2019) is a pretraining framework with the assumption that the sentences with the same head and tail entities are positive pairs. During the testing phase, it ranks the similarity score between the query instance and the support instances and chooses the relation with the highest score as the prediction. 5) **CP** (Peng et al., 2020) is also a pretraining framework that regards the sentences with the same relations as positive pairs. The fine-tuning strategy of CP is much like the strategy in Proto; the difference is that they use the dot product instead of Euclidean distance to measure the similarities between instances. Our method differs from CP in that we also consider label-aware information in both pretraining and fine-tuning.

Comparison results We consider four types of few-shot learning tasks in our experiments, which are 5-way 1-shot, 5-way 5-shot, 10-way 1-shot, and 10-way 5-shot learning tasks. For the comparison methods, most results are collected from the published papers (Gao et al., 2019; Peng et al., 2020; Qu et al., 2020). While for MTB (Soares et al., 2019), which does not have publicly available code for reproduction, we present the results reproduced with a BERT$_{\text{BASE}}$ model trained with the MTB pretraining strategies (Soares et al., 2019; Peng et al., 2020). As for CP (Peng et al., 2020), which does not include the results for the NYT-25 dataset, we reproduce the results by fine-tuning the

Method	FewRel				NYT-25			
	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Proto	80.68	89.60	71.48	82.89	77.63	87.25	66.49	79.51
BERT-pair	88.32	93.22	80.63	87.02	80.78	88.13	72.65	79.68
REGRAB	90.30	94.25	84.09	89.93	89.76	95.66	84.11	92.48
MTB	91.10	95.40	84.30	91.80	88.90	95.53	83.08	92.23
CP	95.10	97.10	91.20	94.70	91.08	94.73	83.99	90.18
MapRE	**95.73**	**97.84**	**93.18**	**95.64**	**91.90**	**96.01**	**86.46**	**92.68**

Table 2: Comparison results on the test set of the FewRel and NYT-25 datasets in accuracy.

pretrained CP [2] on the NYT-25 datasets. For our model, we fine-tune on our pretrained MapRE with the approaches described in Section 5.1, which considers both label-agnostic and label-aware information in fine-tuning. More details about the parameter settings can be found in the Appendix. Table 2 presents the comparison results on two few-shot learning datasets in different task settings. We can observe that, pretraining the framework with matching information between the instances (i.e., MTB, CP, and ours) can significantly improve the model performance in few-shot scenarios. Comparing the label-aware methods (i.e., REGRAB and ours) with label-agnostic methods on the NYT-25 dataset, which lies in a different domain than Wikipedia, the label-aware methods can grasp more hints from the relation semantic knowledge for prediction. Such improvements become much significant with a larger number of relations N and fewer support instances K, which suggests that the label-aware information is valuable in extreme low-resource conditions. For all settings, the proposed MapRE, which considers both label-agnostic and label-aware information in pretraining and fine-tuning, provides steady performance and outperforms a series of baseline methods as well as the state-of-the-art. The results prove the effectiveness of the proposed framework, and suggest the importance of the semantic mapping information from both label-aware and label-agnostic knowledge.

Discussion We further consider two variants of MapRE, i.e., employing only the label-agnostic information or only the label-aware information, to discover how the two types of information contribute to the final performance. Table 3 shows the model performance on different options in fine-tuning the framework. Comparing the results of label-agnostic only MapRE with the model CP in Table 2, where the only difference is that we con-

Method	FewRel			
	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Label-agnostic	95.56	97.60	92.55	95.19
Label-aware	72.97	72.74	61.05	60.98
Both	95.73	97.84	93.18	95.64

Table 3: Accuracy on the test set of the FewRel dataset.

sider the label-aware information in pretraining the framework, we can see that the incorporating the relation label information does help the model to capture more semantic knowledge. However, if we only consider the label-aware information in fine-tuning, the performance drops since the model does not utilize any support instances, which is much like zero-shot learning. Note that there are fluctuates in 5-way 5-shot and 10-way 5-shot of the relation-aware only MapRE; this may be caused by the difference in the testing set of the FewRel for the four few-shot learning tasks provided online [3]. We will discuss more details about zero-shot RE in the following subsection. The results of the label-aware only MapRE suggest the importance of the label-agnostic knowledge in few-shot RE. Overall, both label-agnostic and label-aware knowledge are valuable for few-shot RE tasks, and using them in both pretraining and fine-tuning can significantly improve the results.

5.3 Zero-shot RE

We further consider an extreme condition of low-resource RE, i.e., zero-shot RE, where no support instance is provided for prediction. Under the condition of zero-shot RE, most of the above few-shot RE frameworks are not applicable since they need at least one example for each support relation for comparison. Previous studies for zero-shot learning lie in representing the label by vectors, then compare the input embedding with the label vec-

Method	FewRel		NYT-25	
	5-way 0-shot	10-way 0-shot	5-way 0-shot	10-way 0-shot
Qu et al. (2020)♣	52.50	37.50	40.50	24.50
Cetoli (2020)	86.00	76.20	-	-
MapRE	**90.65**	**81.46**	**72.14**	**59.94**

Table 4: The comparison results of the zero-shot RE on FewRel and NYT-25 datasets in accuracy. The results for the FewRel dataset and the NYT-25 dataset are evaluated on the validation set and test set, respectively.♣ The results for Qu et al. (2020) are observed from the figures in the paper with a standard deviation of 2%.

tors for comparison (Xian et al., 2017; Rios and Kavuluru, 2018; Xie et al., 2019). The work by Qu et al. (2020) extends the idea by inferring the posterior of the relation label vectors initialized by an external knowledge graph. Another direction is to formalize the zero-shot RE problem as a question-answering task, where Cetoli (2020) fine-tune on a BERT-based model pretrained on SQUAD (Rajpurkar et al., 2016), then use it to generate the relation prediction. Both work needs extra knowledge to tune the framework; however, the external knowledge is not always available for the given tasks. In our work, we fine-tune on the pretrained MapRE with only label-aware information for zero-shot learning, which can be regarded as a special case in Equation (4) when $\alpha = 0$ and $\beta = 1$. The results show that, compared to the two recent zero-shot RE methods, the proposed MapRE obtains outstanding performance on all zero-shot settings, which proves the effectiveness of our proposed framework.

6 Conclusion

In this work, we propose MapRE, a semantic mapping approach considering both label-agnostic and label-aware information for low-resource relation extraction (RE). Extensive experiments on low-resource supervised RE, few-shot RE, and zero-shot RE tasks present the outstanding performance of the proposed framework. The results suggest the importance of both label-agnostic and label-aware information in pretraining and fine-tuning the model for low-resource RE tasks. In this work, we did not investigate the potential effect caused by the domain shift problem, and we will leave the analysis on this to future works.

References

Alberto Cetoli. 2020. Exploring the zero-shot limit of fewrel. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1447–1451.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. 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 6251–6256.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. Opennre: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 169–174.

Xu Han, 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 *EMNLP*.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2020. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*.

Linmei Hu, Luhao Zhang, Chuan Shi, Liqiang Nie, Weili Guan, and Cheng Yang. 2019. Improving distantly-supervised relation extraction with joint label embedding. 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 3812–3820.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. 2016. Chemprot-3.0: a global chemical biology diseases mapping. *Database*, 2016.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mike Mintz, Steven Bills, Rion Snow, and Dan 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.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A simple neural attentive meta-learner. In *International Conference on Learning Representations (ICLR)*.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Meng Qu, Tianyu Gao, Louis-Pascal Xhonneux, and Jian Tang. 2020. Few-shot relation extraction via bayesian meta-learning on relation graphs. In *International Conference on Machine Learning (ICML)*, pages 7867–7876. PMLR.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4080–4090.

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 (ACL)*, pages 2895–2905.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1857–1865.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. 2019. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and S Yu Philip. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4582–4591.

Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. 2019. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9384–9393.

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 (ACL)*, pages 2872–2881.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3905–3914.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. 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 1031–1040.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

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

A Appendix

A.1 Pretraining Details

Data preparation We take similar strategies as in CP (Peng et al., 2020) for pretraining the models. The difference is we also consider the label-aware information to pretrain the model. The pretraining corpus is from Wikidata (Vrandečić and Krötzsch, 2014), where we exclude any overlapping data between Wikidata and the datasets we used for evaluation. The training instances are sampled from the Wikidata as we described in the section of matching sample formulation.

Implementation details We train on the $\text{BERT}_{\text{BASE}}$ model from the open-source transformer toolkits [4] and use AdamW (Loshchilov and Hutter, 2018) as the optimizer. The max length for the input is set as 60. The pretraining is implemented with eight Tesla V100 32G GPUs, which will take about 6 hours for about 11,000 training steps with the first 500 steps as the warmup steps. The batch size is set as 2040, the learning rate is 3×10^{-5}, the weight decay rate is

Parameter	Wiki80	ChemProt
Batchsize	64	64
Max training epochs	20	100
Learning rate	3×10^{-5}	3×10^{-5}
Weight decay rate	1×10^{-5}	1×10^{-5}
Warmup steps	500	500
Max sentence length	100	100

Table 5: Fine-tuning settings for supervised RE.

1×10^{-5}, and the max gradient norm for clipping is set as 1.0.

A.2 Fine-tuning Details

Supervised Relation Extraction The two supervised datasets, Wiki80 and ChemProt, can be found in the repository [5]. We follow the same strategy to split each dataset into training, validation, and testing samples, where we have accordingly 39,200, 5,600, and 11,200 samples for Wiki80 dataset, and 4,169, 2,427, and 3,469 for the ChemProt dataset. We also follow their settings to 1%, 10%, and 100% of the training sets to evaluate the model performance in low-resource scenarios. The parameter settings to fine-tune on the two datasets can be found in table 5.

Few & Zero-shot Relation Extraction The details about the two datasets can be found in (Han et al., 2018; Gao et al., 2019; Qu et al., 2020). The

Parameter	FewRel	NYT-25
Training task	5-way 1-shot	5-way 5-shot
# Training query instances	1	1
Max sentence length	60	200
Batch size	4	4
Training iteration	10,000	1,000
Learning rate	3×10^{-5}	3×10^{-5}
Weight decay rate	1×10^{-5}	1×10^{-5}

Table 6: Fine-tuning settings for few-shot RE.

general parameter settings for both few and zero-shot learning are shown in Table 6. The difference of the settings for few and zero settings lies in the settings of the coefficients α and β, which controls the contribution of the relation-agnostic and relation-aware information. For few-shot learning, we initialize the two coefficients as 0.95 and 1.05, where they will be optimized during fine-tuning. As for the zero-shot learning, which only uses the relation-aware information, we set α as 0 and β as 1.0.

---

[4] https://github.com/huggingface/transformers

[5] https://github.com/thunlp/RE-Context-or-Names