Towards Realistic Low-resource Relation Extraction: A Benchmark with Empirical Baseline Study

Xin Xu[1,2]*, **Xiang Chen**[1,2]*, **Ningyu Zhang**[1,2]†, **Xin Xie**[1,2], **Xi Chen**[3], **Huajun Chen**[1,2]
[1]Zhejiang University & AZFT Joint Lab for Knowledge Engine
[2]Hangzhou Innovation Center, Zhejiang University, [3]Tencent
{xxucs@zju.edu.cn, xiang_chen, xx2020, huajunsir, zhangningyu}@zju.edu.cn,
jasonxchen@tencent.com
https://zjunlp.github.io/project/LREBench

Abstract

This paper presents an empirical study to build relation extraction systems in low-resource settings. Based upon recent pre-trained language models, we comprehensively investigate three schemes to evaluate the performance in low-resource settings: (i) different types of prompt-based methods with few-shot labeled data; (ii) diverse balancing methods to address the long-tailed distribution issue; (iii) data augmentation technologies and self-training to generate more labeled in-domain data. We create a benchmark with 8 relation extraction (RE) datasets covering different languages, domains and contexts and perform extensive comparisons over the proposed schemes with combinations. Our experiments illustrate: (i) Though prompt-based tuning is beneficial in low-resource RE, there is still much potential for improvement, especially in extracting relations from cross-sentence contexts with multiple relational triples; (ii) Balancing methods are not always helpful for RE with long-tailed distribution; (iii) Data augmentation complements existing baselines and can bring much performance gain, while self-training may not consistently achieve advancement to low-resource RE[1].

1 Introduction

Relation Extraction (RE) aims to extract relational facts from the text and plays an essential role in information extraction (Zhang et al., 2022b). The success of neural networks for RE has been witnessed in recent years; however, open issues remain as they still depend on the number of labeled data in practice. For example, Han et al. (2018) found that the model performance drops dramatically as the number of instances for one relation decreases, e.g., for long-tail. An extreme scenario is few-shot RE, where only a few support examples are given. This motivates a **Low-resource RE** (**LRE**) task where annotations are scarce (Brody et al., 2021).

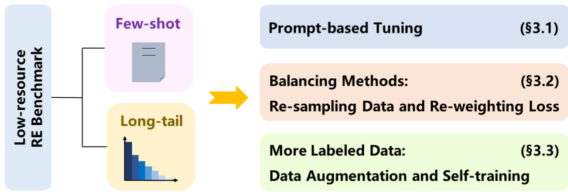


Figure 1: An overview of methods studied in our paper.

Many efforts are devoted to improving the generalization ability beyond learning directly from limited labeled data. Early, Mintz et al. (2009) proposes distant supervision for RE, which leverages facts in KG as weak supervision to obtain annotated instances. Rosenberg et al. (2005); Liu et al. (2021a); Hu et al. (2021) try to assign pseudo labels to unlabeled data and leverage both pseudo-labeled data and gold-labeled data to improve the generalization capability of models iteratively. Some studies apply meta-learning strategies to endow a new model with the ability to optimize rapidly or leverage transfer learning to alleviate the data-hungry issue (Gao et al., 2019; Yu et al., 2020b; Li et al., 2020a; Deng et al., 2021). Other studies (Zhang et al., 2019) focus on the long-tailed class distribution, especially in tail classes that only allow learning with a few instances. With the prosperity of the pre-trained language models (PLMs), the pre-train – fine-tune paradigm has become standard for natural language processing (NLP), leading to a tremendous increase in LRE performance. More recently, a new methodology named prompt learning has made waves in the community by demonstrating astounding few-shot capabilities on LRE (Han et al., 2021; Chen et al., 2022d).

In this work, we benchmark more realistic scenarios on diverse datasets for low-resource RE, in which models have to handle **both extreme few-**

* Equal contribution and shared co-first authorship.
† Corresponding author.
[1]Code and datasets are in https://github.com/zjunlp/LREBench.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 413 - 427
December 7-11, 2022

shot instances and long-tailed distribution, and can also make use of data augmentation or unlabeled in-domain data without cross-validation (Perez et al., 2021). These settings are appealing as: (i) Such models mirror deployment in applied settings; (ii) Few-shot settings are realistic with long-tailed distribution; (iii) Diverse datasets cover different languages (Chinese and English), domains (general, scientific), and contexts (one or more sentences with single or multiple relational triples).

Specifically, we focus on improving the generalization ability from three directions shown in Figure 1. Instead of using limited few-shot data, we create different types of prompts for RE and empirically analyze low-resource performance. We further implement many popular balancing methods for long-tailed distribution, which can mitigate performance decay in instance-scarce (tail) classes. We also leverage more generated training instances by data augmentation and self-training in conjunction with the limited labeled data.

Our contributions include: (i) We present the **first** systematic study for low-resource RE, an important problem in information extraction, by investigating three distinctive schemes with combinations. (ii) We conduct extensive comparisons with in-depth analysis on 8 RE datasets and report empirical results with insightful findings. (iii) We release both the data and the source code of these baselines as an open-sourced testbed for future research purposes.

To shed light on future research on low-resource RE, our empirical analysis suggests that: (i) Previous state-of-the-art methods in the low-resource setting still struggle to obtain better performance than that in the fully-supervised setting (Cross-sentence LRE is extremely challenging), which indicates that there is still much room for low-resource RE. (ii) Balancing methods may not always benefit low-resource RE. The long-tailed issue can not be ignored, and more studies should be focused on model development. (iii) With some simple data augmentation methods, better performance can be achieved, highlighting opportunities for future improvements on low-resource RE.

2 Background on Low-resource RE

2.1 Low-resource RE

RE is a classification task that aims to assign relation labels to entity pairs in given contexts. Formally, in a RE dataset denoted as $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$, $\mathbf{X}$ is the set of texts and $\mathbf{Y}$ is the set of relation labels. Given a text $x = \{w_1, w_2, \ldots, w_s, \ldots, w_o, \ldots, w_{|x|}\}$, where $x \in \mathbf{X}$, RE aims to predict the semantic relation $y_x \in \mathbf{Y}$ holding between the subject entity w_s and the object entity w_o. Conventional RE systems are trained in the standard supervised learning regime, where large amounts of labeled examples are required. Nevertheless, owing to various languages, domains, and the cost of human annotation, there is commonly a very small number of labeled examples in real-world applications. Thus, traditional supervised learning with few-shot labeled data struggle to achieve satisfactory performance (Schick and Schütze, 2021). Consequently, a challenging task, *low-resource RE*, has emerged.

2.2 Fine-tuning PLMs for RE

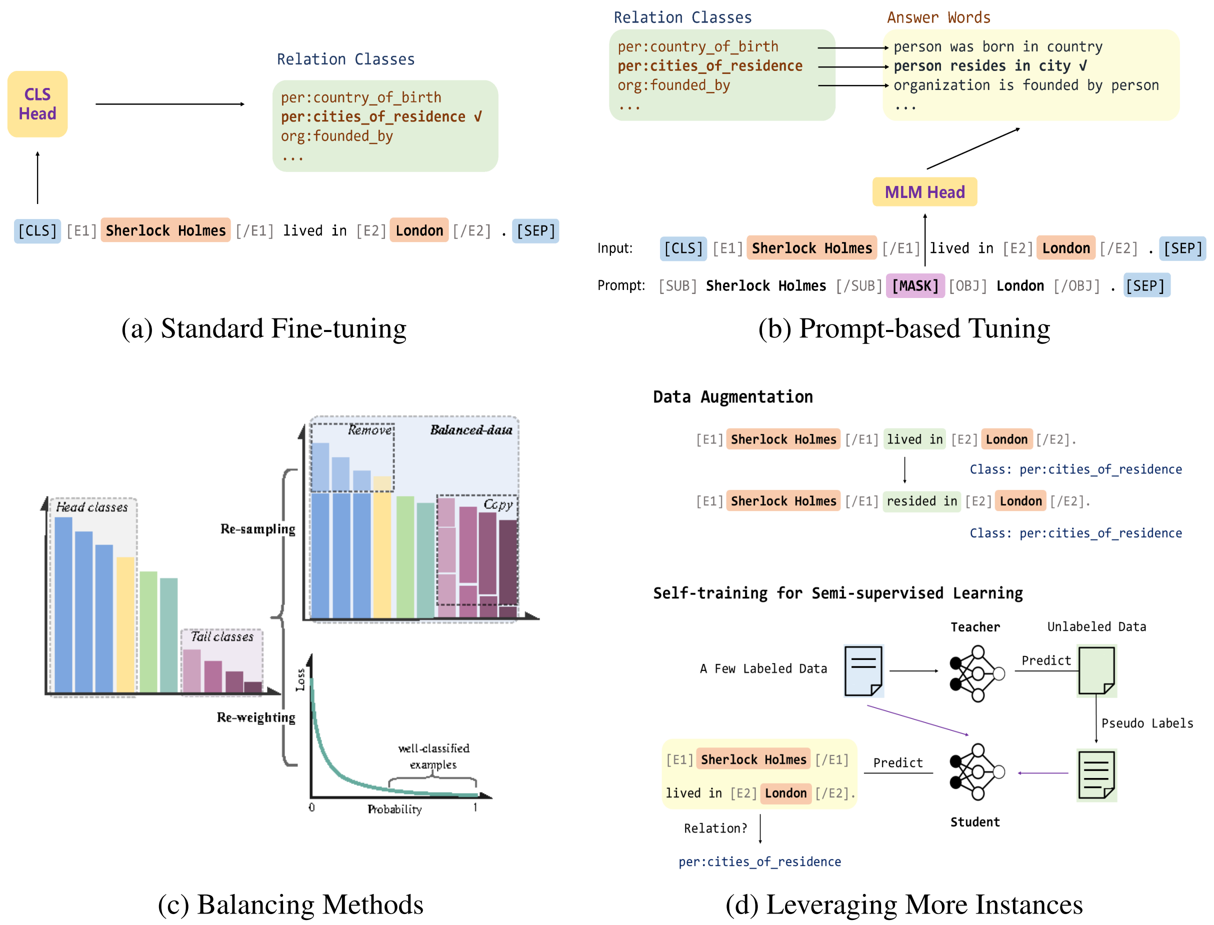
A typical baseline method for RE is to fine-tune a PLM $\mathcal{M}$ as shown in Figure 2(a). Firstly, the tokenizer of $\mathcal{M}$ converts the text x into the input tokens of $\mathcal{M}$, such as $\texttt{[CLS]}x_{\text{token}}\texttt{[SEP]}$, and then encodes tokens into the corresponding hidden vectors, such as $\mathbf{h} = \{\mathbf{h}_{\texttt{[CLS]}}, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_s, \ldots, \mathbf{h}_o, \ldots, \mathbf{h}_{\texttt{[SEP]}}\}$. Then, a [CLS] head is used to compute the probability distribution over the class set $\mathbf{Y}$ with the softmax $p(\cdot|x) = \texttt{Softmax}(\mathbf{W}\mathbf{h}_{\texttt{[CLS]}} + \mathbf{b})$, where $\mathbf{W}$ is a set of learnable weight parameters randomly initialized at the start of fine-tuning, $\mathbf{h}_{\texttt{[CLS]}}$ is the hidden vector of [CLS] and $\mathbf{b}$ is the learnable bias. All learnable parameters are fine-tuned by minimizing the cross-entropy loss over $p(y_x|x)$ on $\mathcal{D}$. Nevertheless, conventional supervised fine-tuning may over-fit a few training examples and perform poor generalization ability over test sets when encountering the low-resource RE task.

3 Methods for Low-resource RE

In this paper, we conduct a comprehensive empirical study with three distinctive schemes against difficulty in low-resource RE: PLMs-based prompt-based tuning, balancing long-tailed data and leveraging more instances, as shown in Figure 2.

3.1 Prompting for Few-shot Instances

To address the low-resource issue of data sparsity for RE, we firstly analyze **prompting methods**. Unlike standard fine-tuning, prompt-based tuning reformulates classification tasks as cloze-style language modeling problems and predicts an-



(a) Standard Fine-tuning

(b) Prompt-based Tuning

(c) Balancing Methods

(d) Leveraging More Instances

Figure 2: Illustrations of different methods used in our low-resource RE benchmark. (a) A standard RE pipeline of fine-tuning a PLM such as BERT and RoBERTa (§2.2). (b) Prompt-based tuning, which concatenates the original input with the prompt template to predict [MASK] by an MLM head and then injects the predicted answer words to the corresponding class sets (§3.1). (c) Two balancing methods, re-sampling data and re-weighting losses, to address the long-tailed issue (§3.2). (d) Levering more instances with data augmentation and self-training (§3.3).

swer words, denoted as $\mathbf{V}$, through the masked language model (MLM) head. Specifically, $\mathcal{T}_{\text{prompt}}$ converts every instance x into a prompt input $x_{\text{prompt}} = \mathcal{T}_{\text{prompt}}(x)$, in which there is at least one [MASK] for $\mathcal{M}$ to fill with right answer words $v \in \mathbf{V}$. Meanwhile, a verbalizer connects relation labels with answer words via an injective mapping $\gamma : \mathbf{Y} \rightarrow \mathbf{V}$. With the aforementioned functions, we can formalize the probability distribution over $\mathbf{Y}$ with the probability distribution over $\mathbf{V}$ at the masked position (Ma et al., 2021):

$$\begin{aligned} P(y_x|x) &= P(\texttt{[MASK]} = \gamma(y_x)|x_{prompt}) \\ &= Softmax(\mathbf{W}_{lm} \cdot \mathbf{h}_{\texttt{[MASK]}}) \end{aligned} \tag{1}$$

where $\mathbf{W}_{lm}$ is a set of parameters of the PLM head.

Note that the main difference between various prompt-based tuning methods lies in the design of the prompt template and verbalizer. Thus, we benchmark different kinds of prompting methods in low-resource RE to empirically investigate their performance. For the prompt template, given the input x, the first choice is manually designing the template. We utilize the natural language or task schema to formulate different prompt templates. Formally, we have:

Template Prompt:
[CLS] x. [SEP] The relation between [sub] and [obj] is [MASK]. [SEP]

Schema Prompt:
[CLS] x. [SEP] [[sub] | [obj]] relation: [MASK]. [SEP]

where <sub> is the head entity mention and <obj> is the tail entity mention. Since there exists rich semantic knowledge within relation labels and structural knowledge implications among relational triples, we also benchmark previous studies such as *PTR* (Han et al., 2021) and *KnowPrompt* (Chen et al., 2022d) which incorporates relational knowledge into prompt-based tuning as shown in Figure 2(b).

3.2 Balancing for Long-tailed Distribution

Learning with long-tailed data, where the number of instances in each class highly varies, is a common challenge in low-resource RE because instance-rich (head) classes predominate the training procedure. Note that the learnable parameters of the trained model prefer to perform better in these head classes and worse in less frequent (tail) classes (Kang et al., 2020a). To address this issue, we explore two balancing methods: re-sampling data and re-weighting losses for low-resource RE.

Re-sampling Data We re-sample RE datasets to balance the data distribution. For example, the tail classes can be over-sampled by adding copies of data, and the head classes can be under-sampled by removing data, as shown in Figure 2(c). Specifically, we use a toolkit[2], which can estimate the sampling weights automatically when sampling from imbalanced data to obtain datasets with the nearly balanced distribution.

Re-weighting Loss We utilize various re-weighting losses, assigning different weights to different training instances for each class. For instance, *DSC Loss* (Li et al., 2020b) attaches similar importance to false positives and false negatives. *Focal Loss* (Lin et al., 2020a) balances the sample-wise classification loss for model training by down-weighing easy samples. *GHM Loss* (Li et al., 2019a) applies a gradient harmonizing mechanism, making the model ignore outliers to conquer the disharmony in classification. *LDAM Loss* (Cao et al., 2019) expands the decision boundaries of few-shot classes.

3.3 Leveraging More Instances via Data Augmentation and Self-training

It is also beneficial to leverage more instances to address the low-resource issue. We conduct data augmentation and also leverage unlabeled in-domain data via self-training, as shown in Figure 2(d).

Data augmentation (DA) automatically generates more labeled instances based on only a few labeled instances. For example, we utilize token-level augmentation, which changes or inserts words and phrases in a sentence to generate augmented text remaining with the same labels as the original text. In this work, we apply three DA methods for English RE datasets to substitute words in training sets based on **WordNet's synonyms**, **TF-IDF similarity** and the **contextual word embedding** implemented by *nlpaug*[3]. And we replace words with their synonyms via *nlpcda*[4] to augment Chinese RE samples. We further analyze different types of augmentation objects in RE regarding contexts, entities, and both of them.

Since substantial easily-collected unlabeled data are also leveraged in this work for low-resource RE, we conduct self-training, a classical, intuitive and straightforward semi-supervised learning method. Specifically, we train a model with labeled data and then expand the labeled set according to the most confident predictions (a.k.a. pseudo labels) on unlabeled data. We combine the data with gold and pseudo labels to obtain the final RE model. The details of the whole self-training pipeline are described in Appendix A.5.

4 Benchmark Design

In this paper, we provide a comprehensive empirical study for low-resource RE and design the **LREBench** (**L**ow-resource **R**elation **E**xtraction **Bench**mark) to evaluate various methods. In the following section, we will detail the datasets chosen for experiments and the reproducibility of all baselines mentioned above.

4.1 Datasets Selection

As shown in Table 1, we select 8 RE datasets to evaluate baselines in low-resource settings, covering various domains: SemEval 2010 Task 8[5] (Hendrickx et al., 2009), TACREV[6] (Alt et al., 2020), DialogRE[7] (Yu et al., 2020a) and DuIE2.0[8] (Li et al., 2019b) on the general domain, Wiki80[9] (Han et al., 2019) on the encyclopedic domain, ChemProt[10] (Peng et al., 2019) on the biochemical domain, SciERC[11] (Luan et al., 2018) on the scientific domain, and CMeIE[12] (Zhang et al., 2022a) on the medical domain. Except for frequently-used English datasets, we select Chinese datasets,

[2] https://github.com/ufoym/imbalanced-dataset-sampler

[3] https://github.com/makcedward/nlpaug

[4] https://github.com/425776024/nlpcda

[5] https://github.com/zjunlp/KnowPrompt/tree/master/dataset/semeval

[6] https://github.com/DFKI-NLP/tacrev

[7] https://dataset.org/dialogre/

[8] https://www.luge.ai/#/luge/dataDetail?id=5

[9] https://github.com/thunlp/OpenNRE/blob/master/benchmark/download_wiki80.sh

[10] https://github.com/ncbi-nlp/BLUE_Benchmark

[11] http://nlp.cs.washington.edu/sciIE/

[12] https://tianchi.aliyun.com/dataset/dataDetail?dataId=95414

Datasets	**SemEval**	**TACREV**	**Wiki80***	**SciERC**	**ChemProt**	**DialogRE**	**DuIE2.0** (cn)	**CMeIE** (cn)
Domain	General	General	Encyclopedic	Scientific	Biochemical	Dialogue	General	Medical
# Train	6.5k	68.1k	12.0k	3.2k	19.5k	6.0k	153k	34k
# Test	2.7k	15.5k	5.6k	974	16.9k	1.9k	18k	8.7k
# Relation Class	19	42	80	7	14	37	48	44
MS / MT	× / ×	× / ×	× / ×	✓ / ✓	✓ / ✓	✓ / ✓	× / ✓	✓ / ✓

Table 1: Statistics on the 8 public RE datasets selected for evaluation in **LREBench**. **MS** indicates if datasets contain instances with multiple sentences in one text, and **MT** indicates if one text in these datasets can be related to multiple relational triples. "*" means that we **re-sample** and convert Wiki80 into long-tailed distribution through an exponential function since its original distribution is exactly balanced. "cn" represents datasets with Chinese.

such as DuIE2.0 and CMeIE. Besides, the SciERC, ChemProt, DialogRE, and CMeIE datasets contain the situation where multiple sentences are in one instance, which is for cross-sentence RE and more challenging than single-sentence RE in SemEval, TACREV and Wiki80.

For simplicity, we provide a unified input-output format for all datasets in the low-resource setting[13]. Specifically, each instance in LREBench consists of one text and one relational triple (one head entity and one tail entity in the text and the corresponding relation between them). For those datasets with instances having one text related to multiple relational triples, such as ChemProt, SciERC, DialogRE, DuIE2.0 and CMeIE, we follow Zhong and Chen (2021) to place such a text to multiple instances with only one relational triple. In this way, we can utilize a unified input-output format for widespread models.

We conduct experiments in three settings with different proportions of training data to simulate different resource levels: 8-shot, 10% and 100%. For the 8-shot setting, we sample 8 instances for each relation category in the training and test sets[14]. For the 10% and 100% settings, we sample 10 percent of the training set and use the whole training set, respectively. Since fine-tuning on small datasets can suffer from instability and results may change dramatically given a new split of data (Gao et al., 2021), **we sample all training datasets 5 times randomly in 8-shot and 10% settings and measure their average performance in experiments**. Also, we follow the same sampling strategy in the re-sampling long-tailed data method and data augmentation methods to obtain a fair comparison.

[13] We utilize a unified json format for evaluation, and it is straightforward to adapt to other datasets.

[14] If there are less than 8 instances in one relation class, we delete all instances of this class.

4.2 Reproducibility

Methods Throughout our experiments, we employ $\mathcal{M}$ = *RoBERTa-large* (Liu et al., 2019) for SemEval, TACREV, Wiki80 and DialogRE, Chinese *RoBERTa-large* (Cui et al., 2020) for DuIE2.0 and CMeIE, and *BioBERT-large* (Lee et al., 2020) for ChemProt and SciERC from *HuggingFace*[15] as the backbone network (detailed in Appendix A.1). For each method, we investigate the following three schemes in different settings for the comparative empirical study, as shown in Table 2: (i) **Normal** is the general scheme with the PLM for low-resource relation extraction, in which we evaluate with 8-shot, 10% and 100% settings. (ii) **Balance** refers to balancing methods in §3.2 for long-tailed data distribution with 10% and 100% settings. We list the best performance among all balancing methods for each dataset in Table 2 and detailed results in Table 3. (iii) **Data augmentation (DA)** methods are applied to 10% training sets. We list the best performance among all DA methods in Table 2 and all performance in Table 4. We also conduct **self-training (ST)** that firstly trains a teacher $\mathcal{M}$ on 10% training data and then tags the rest 90% training data with pseudo labels by $\mathcal{M}$. Both gold-labeled and pseudo-labeled data are used to obtain a final student RE model as introduced in §3.3.

Training and Evaluation We only train models on training sets **without validation on development sets** to ensure true few-shot learning with limited labeled data. For all training data sizes, we set the training epoch = 10 following Huang et al. (2021). Except for re-weighting losses for addressing the long-tailed problem, the cross-entropy loss is used in all training processes. Since the performance of head and tail classes varies a lot, we use both Macro F1 and Micro F1 together as the evaluation metrics. Implementation details can be found in Appendix A.

[15] https://huggingface.co/

Dataset	Metric	Fine-Tune							Prompt						
		Normal			Balance		DA	ST	Normal			Balance		DA	ST
		8-shot	10%	100%	10%	100%	10%	10%	8-shot	10%	100%	10%	100%	10%	10%
SemEval	MaF1	2.69	34.63	81.88	41.84	82.44	69.84	60.10	48.54	44.71	83.40	54.54	83.20	71.73	63.55
	MiF1	9.70	54.61	89.10	58.26	89.44	78.98	74.12	54.55	69.90	90.01	76.53	92.31	83.54	76.81
TACREV	MaF1	1.02	47.32	63.41	48.64	63.38	50.68	48.84	29.46	61.40	67.08	63.09	69.63	62.20	7.32
	MiF1	1.76	65.43	71.68	67.19	73.86	65.99	66.89	30.88	77.00	78.30	76.25	81.41	76.90	32.93
Wiki80	MaF1	37.89	37.82	71.31	44.37	73.36	49.40	37.47	75.11	60.67	82.79	63.99	83.72	63.40	60.86
	MiF1	44.85	46.50	72.82	49.74	74.20	55.00	45.91	76.34	64.86	82.96	67.86	83.86	66.96	65.04
SciERC	MaF1	10.41	10.31	83.41	10.11	81.17	30.09	31.48	23.26	51.71	83.27	60.55	84.83	65.98	56.94
	MiF1	39.12	54.66	89.12	54.72	87.78	61.79	64.07	22.07	74.00	89.01	76.90	90.04	79.92	76.32
ChemProt	MaF1	2.18	27.96	47.35	33.38	47.35	36.31	30.67	6.17	36.43	47.16	38.99	47.07	37.44	33.62
	MiF1	8.93	49.20	68.81	54.98	68.77	56.58	54.17	8.65	56.96	69.14	57.28	69.12	58.26	53.55
DialogRE	MaF1	1.13	2.17	25.31	5.84	27.28	9.74	0.00	44.96	45.51	64.49	46.22	71.73	49.47	34.70
	MiF1	3.92	23.37	41.52	24.53	41.24	27.40	0.00	45.70	54.16	73.66	55.65	73.52	57.53	46.54
DuIE2.0	MaF1	36.62	90.46	95.01	92.91	96.00	91.47	89.27	80.31	93.48	95.73	93.70	96.01	93.66	90.49
	MiF1	39.00	94.42	96.22	94.46	96.13	94.46	93.81	82.14	95.09	96.43	95.23	96.44	95.11	93.35
CMeIE	MaF1	13.68	62.30	84.37	67.22	86.31	63.82	58.46	36.54	67.59	86.42	67.84	86.68	69.95	65.79
	MiF1	17.05	79.82	90.48	80.43	90.56	80.14	78.92	38.02	83.38	92.08	83.40	92.14	83.71	81.26

Table 2: F1 Scores (%) on 8 datasets with various sizes of training data in different methods for the low-resource scenario. *MaF1* and *MiF1* mean Macro F1 Score (%) and Micro F1 Score (%) respectively. ***Normal*** means the standard PLM fine-tuning method and ***Prompt*** means prompt-based tuning implemented by *KnowPrompt*. ***Balance*** represents balancing methods for long-tailed data. ***DA*** is data augmentation. ***ST*** refers to self-training with unlabeled in-domain data. Results colored with red means prompt-based tuning works worse than fine-tuning between two *Normal* columns. blue, orange, and purple results indicates the performance of balancing methods, data augmentation and self-training is poorer than the *Normal* method in the same setting.

5 Results and Discussions

5.1 Main Results

We leverage the basic PLM fine-tuning code from *OpenNRE*[16] (Han et al., 2019) and the state-of-the-art prompt-based RE method *KnowPrompt* (Chen et al., 2022d) to conduct extensive experiments across 8 datasets in various methods and settings. The results of the main experiments are shown in Table 2, which illustrates the following findings:

Finding 1: Prompt-based tuning largely outperforms standard fine-tuning for RE, especially more effective in the low-resource scenario. The comparison between the results of standard fine-tuning and prompt-based tuning indicates that prompts can provide task-specific information and bridge the pre-train – fine-tune gap, thus, empowering PLMs in low-resource RE.

Finding 2: Though balancing methods obtain advancement with long-tailed distribution, they may still fail on challenging RE datasets, such as ChemProt, DialogRE and DuIE2.0. By comparing Macro F1 Scores of the *Balance* columns and *Normal* columns, blue (bad) results illustrate that balancing methods are affected by complexity of long contexts with multiple sentences and relational triples.

Finding 3: Data augmentation achieves much gain on RE and sometimes even better performance than prompt-based tuning, such as on SemEval, according to the difference between two pairs of *DA* columns and *Normal* columns in the 10% setting. More data generated through DA methods are complementary with other baselines, boosting the performance.

Finding 4: RE systems struggle against difficulty in obtaining correct relations from cross-sentence contexts and among multiple triples. The extremely low F1 scores for 8-shot ChemProt, and DialogRE datasets in standard fine-tuning demonstrate this finding. One text in ChemProt is related to too many relational triples (there are 347 texts related to 3 triples and 699 texts related to 2 triples in the training set). At the same time, in DialogRE, the input text is extremely long (one text can contain 10 sentences). Even with the powerful prompt-based tuning method, it is non-trivial to address the low-resource issue according to the unexpected drop in F1 scores of ChemProt and SciERC.

Finding 5: Self-training with unlabeled in-domain data may not always show an advantage for low-resource RE. There is much noise in those

[16] https://github.com/thunlp/OpenNRE

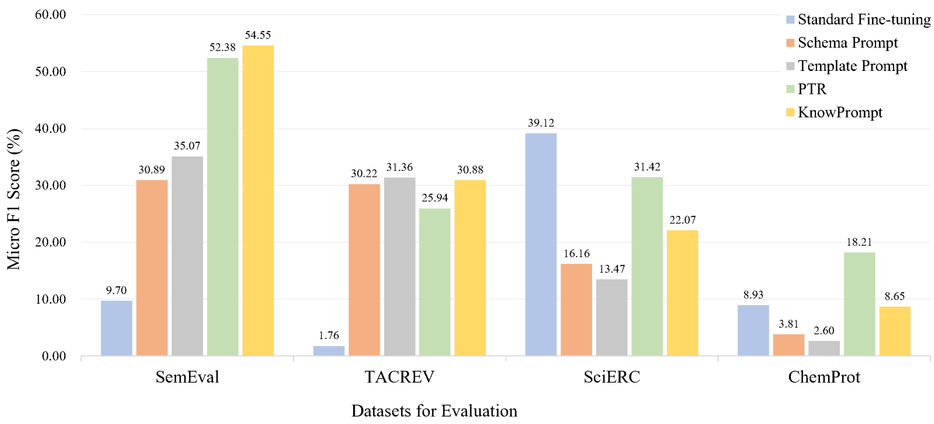


Figure 3: Micro F1 Scores (%) of different prompts on 8-shot datasets. *RoBERTA-large* is used on SemEval and TACREV and *BioBERT-large* is used on SciERC and ChemProt as backbone networks.

generated pseudo labels. Furthermore, for assigning labels in RE, both semantics and positions of entities in a text need to be considered, which is exceedingly challenging. Therefore, the model with self-training cannot always obtain better performance in low-resource settings.

5.2 Comprehensive Empirical Analysis

Different Prompting Methods To investigate the effects of different prompts, we conduct an empirical analysis on SemEval, TACREV, SciERC and ChemProt as shown in Figure 3. We observe the following insights: (i) **Prompt-based tuning is more beneficial in general domains than specific domains for low-resource RE.** Prompt-based tuning achieves the most gain, 44.85% Micro F1 Score, by comparing fine-tuning and *KnowPrompt* on 8-shot SemEval, while obtaining the worst drop, 25.65% Micro F1 Score, by comparing fine-tuning and the template prompt on 8-shot SciERC even with the domain-specific PLM. Except for the difficulty of these two datasets, general manual prompts have little domain knowledge related to vertical domains, hindering performance. (ii) **Entity type information in prompts is helpful for low-resource RE.** The head and tail entity types in prompts provide strong constraints between relations and their related entities. Prompting methods with entity type information in *KnowPrompt* and *PTR* perform better than the template and schema-based prompt in most datasets, which illustrates that prompts with entity-type information are more appropriate for low-resource RE. The reason for the abnormal phenomenon that KnowPrompt and PTR obtain worse results than the template and schema-based prompts in TACREV is that annotation errors in the training set of TACREV (Stoica et al., 2021) can lead to overestimation of the performance of models depending on the side information of entities such as entity names, spans and types (Zhou and Chen, 2021), and the templates of *KnowPrompt* and *PTR* are natural language sentences consisting of the head, and tail entities and their relations, which require high-quality annotated entity mentions, positions, types and relational words, while they are relatively trivial to the template and schema-based prompts.

Different Balancing Methods We also conduct experiments to validate the effectiveness of different balancing methods on two long-tailed datasets. We categorize the classes into three splits based on the number of training instances per class, including *Few*, *Medium*, and *Many*, and also report the results on the whole dataset with the *Overall* setting in Table 3 (split schemes are in Appendix B). We notice that with re-balancing methods (e.g., *Focal Loss* and *LDAM Loss*), the tail relations (Few) can yield better performance on both general and domain-specific datasets. However, some technologies, such as *GHM-C*, fail to contribute to performance gains. Overall, our empirical analysis illustrates that the RE performance can be improved with balancing methods, which indicates that long-tailed RE is a challenging classification task, and it should be paid more attention to developing suitable methodologies.

Different Data Augmentation To evaluate the low-resource RE performance with more instances, we generate 30% and 100% augmented instances

Method	SemEval								SciERC							
	Few		Medium		Many		Overall		Few		Medium		Many		Overall	
	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1
Normal	50.42	74.58	**89.53**	89.02	90.17	90.59	83.40	90.01	69.98	67.78	88.05	87.52	92.98	91.93	83.27	89.01
Re-sample	38.17	56.18	70.13	70.56	71.22	71.54	65.37	71.31	71.79	69.64	88.49	87.83	92.96	92.25	82.61	87.58
DSC	49.80	73.87	87.84	88.00	88.97	89.52	82.19	89.00	71.57	69.90	89.94	89.51	93.51	92.88	83.09	88.91
Focal	53.31	77.69	89.50	**89.57**	**90.71**	**91.06**	**84.21**	**90.55**	**73.47**	**72.38**	**91.88**	**91.54**	**94.83**	**94.08**	**84.83**	**90.04**
GHM-C	00.00	00.00	3.39	6.27	70.42	75.81	43.79	70.99	71.34	69.28	89.42	88.82	93.90	93.33	82.95	88.81
LDAM	**53.53**	**79.66**	88.71	88.98	90.32	90.60	83.83	90.15	72.32	70.55	88.48	87.73	94.61	93.98	83.31	89.22

Table 3: F1 Scores (%) on SemEval and SciERC datasets of diverse balancing methods via *KnowPrompt*. *MaF1* and *MiF1* mean Macro F1 Score (%) and Micro F1 Score (%) respectively. *Normal* means conducting the experiment without any balancing methods.

Method	SemEval						TACRED-Revisit					
	30%			100%			30%			100%		
	Context	Entity	All	Context	Entity	All	Context	Entity	All	Context	Entity	All
WordNet's Synonym	**75.49**	75.50	**76.47**	**83.54**	**83.50**	**82.56**	76.54	**76.87**	76.63	**76.12**	**76.59**	**76.37**
TF-IDF Similarity	73.93	**76.23**	74.30	82.92	82.61	82.33	**76.63**	76.05	**76.90**	75.44	75.80	75.15
Contextual Word Embedding (RoBERTa)	73.84	-	74.41	81.63	-	81.31	75.86	76.76	76.35	75.98	76.12	75.92
KnowPrompt (RoBERTa)	69.90						77.00					
	SciERC						**ChemProt**					
WordNet's Synonym	77.70	76.98	**77.54**	**79.36**	**79.40**	**79.92**	**57.37**	57.56	**57.03**	**53.36**	**57.11**	**54.27**
TF-IDF Similarity	**78.50**	**77.33**	73.92	78.30	79.38	79.38	41.22	**58.26**	47.95	43.06	54.60	43.63
Contextual Word Embedding (BioBERT)	76.24	73.55	74.62	75.50	77.35	76.59	56.01	53.48	56.28	45.95	53.26	46.68
KnowPrompt (BioBERT)	74.00						56.96					

Table 4: Micro F1 Scores (%) on four datasets generated by different data augmentation methods from 10% training sets. Three DA methods are conducted to substitute words at three positions: only in contexts, only in entities and in both of them. "-" represents non-repeated data generated based on contextual word embedding is not available.

from 10% training sets by substituting tokens based on three methods. From Table 4, we notice that DA with WordNet can obtain the best performance improvement in most cases. Further, we observe that DA methods can rise by 13.6% and 5.92% Micro F1 Scores mostly on SemEval and SciERC compared to origin prompt-based tuning, demonstrating that DA contributes a lot in the low-resource scenario. Besides, we observe that the performance improvement is much smaller in specific domains, such as SciERC and ChemProt, than in the general domain. We think that because there are many specific terms in vertical domains, it is challenging to obtain qualified augmented instances, which causes to yield lower performance improvement.

6 Related Work

General and Low-resource RE Relation extraction is essential in information extraction. Learning algorithms for RE models involve feature-based methods (Kambhatla, 2004), semi-supervised (Chen et al., 2006; Rosenfeld and Feldman, 2007; Sun et al., 2011), graph-based methods (Zhang et al., 2018; Guo et al., 2019, 2020) and applies PLMs as the backbone (Lin et al., 2020b; Zhang et al., 2021; Zheng et al., 2021; Wu et al., 2022; Chen et al., 2022c,b). Since labeled instances may be limited in practice, low-resource RE has appealed to researchers (Sabo et al., 2021).

Prompting Methods for RE Though fine-tuning PLMs has waved the NLP community, there is still a big gap between pre-training and fine-tuning objectives, hindering the few-shot performance. Hence, prompt-based tuning is proposed in GPT-3 (Brown et al., 2020) and drawn much attention. A series of researches have illustrate the decent performance of prompt-based tuning (Shin et al., 2020; Lester et al., 2021; Li and Liang, 2021), especially in few-shot classification tasks (Schick and Schütze, 2021; Liu et al., 2021b; Chen et al., 2022a). Typically, PTR (Han et al., 2021) encodes prior knowledge using logic rules in prompt-based tuning with several sub-prompts for text classification. KnowPrompt (Chen et al., 2022d) incorporates knowledge among relation labels into prompt tuning for RE with synergistic optimization for better performance.

Methods for Long-tailed Distribution Data Many re-balancing methods are proposed to tackle the long-tailed problem (Kang et al., 2020b; Nan et al., 2021). Data distribution re-balancing methods re-sample the dataset into a more balanced data distribution (Han et al., 2005; Mahajan et al., 2018). Various re-weighing losses (Cui et al., 2019; Li et al., 2019a, 2020b; Lin et al., 2020a; Cao et al., 2019) assign balanced weights to training samples from each class. For RE, Nan et al. (2021) introduces causal inference to mitigate the spurious correlation issues for information extraction.

Data Augmentation for NLP An effective method for NLP in low-resource domains is data augmentation. Token-level DA approaches include replacing tokens with their synonyms (Kolomiyets et al., 2011; Wang and Yang, 2015), deleting tokens (Iyyer et al., 2015), inserting random tokens (Wei and Zou, 2019; Miao et al., 2020) or replacing meaningless tokens with random tokens (Xie et al., 2020; Niu and Bansal, 2018).

7 Conclusion

We provide an empirical study on low-resource RE. Specifically, we analyze the prompt-based tuning for few-shot RE, balancing methods for long-tailed RE datasets, and use data augmentation or unlabeled in-domain data. We systematically evaluate baselines on 8 benchmark datasets in low-resource settings (e.g., 8-shot, 10%) and provide insightful findings. We hope this study can help inspire future research for low-resource RE with more robust models and promote transitioning the technology to real-world industrial scenarios.

8 Limitations

With the fast development of low-resource RE, we cannot compare and evaluate all previous studies due to the settings and non-available open-sourced code. Our motivation is to develop a universal, GLUE-like, and open platform on low-resource RE for the community. We will continue to maintain the benchmark by adding new datasets.

Acknowledgment

We would like to express gratitude to the anonymous reviewers for their kind comments. This work was supported by the National Natural Science Foundation of China (No.62206246, 91846204 and U19B2027), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Ningbo Natural Science Foundation (2021J190), and Yongjiang Talent Introduction Programme (2021A-156-G).

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Sam Brody, Sichao Wu, and Adrian Benton. 2021. Towards realistic few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5338–5345. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1565–1576.

Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2006. Relation extraction using label propagation based semi-supervised learning. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 129–136, Sydney, Australia. Association for Computational Linguistics.

Xiang Chen, Lei Li, Ningyu Zhang, Xiaozhuan Liang, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022a. Decoupling knowledge from memorization: Retrieval-augmented prompt learning. In *Proceedings of NeurIPS 2022*.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022b. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 904–915. ACM.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022c. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618, Seattle, United States. Association for Computational Linguistics.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022d. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2778–2788. ACM.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *CVPR*.

Shumin Deng, Ningyu Zhang, Luoqiu Li, Chen Hui, Huaixiao Tou, Mosha Chen, Fei Huang, and Huajun Chen. 2021. Ontoed: Low-resource event detection with ontology embedding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2828–2839. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. 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.

Zhijiang Guo, Guoshun Nan, Wei LU, and Shay B. Cohen. 2020. Learning latent forests for medical relation extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3651–3657. International Joint Conferences on Artificial Intelligence Organization. Main track.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy. Association for Computational Linguistics.

Hui Han, Wenyuan Wang, and Binghuan Mao. 2005. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing, International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I*, volume 3644 of *Lecture Notes in Computer Science*, pages 878–887. Springer.

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, Hong Kong, China. Association for Computational Linguistics.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2236–2245. Association for Computational Linguistics.

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

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.

Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. 2021. Gradient imitation reinforcement learning for low resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2737–2746. Association for Computational Linguistics.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin

Peng, Jianfeng Gao, and Jiawei Han. 2021. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 178–181, Barcelona, Spain. Association for Computational Linguistics.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020a. Decoupling representation and classifier for long-tailed recognition. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020b. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 271–276, Portland, Oregon, USA. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Buyu Li, Yu Liu, and Xiaogang Wang. 2019a. Gradient harmonized single-stage detector. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Juan Li, Ruoxu Wang, Ningyu Zhang, Wen Zhang, Fan Yang, and Huajun Chen. 2020a. Logic-guided semantic representation learning for zero-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2967–2978. International Committee on Computational Linguistics.

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye Jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. 2019b. Duie: A large-scale chinese dataset for information extraction. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*, page 791–800, Berlin, Heidelberg. Springer-Verlag.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020b. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 465–476. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2020a. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020b. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Kun Liu, Yao Fu, Chuanqi Tan, Mosha Chen, Ningyu Zhang, Songfang Huang, and Sheng Gao. 2021a. Noisy-labeled NER with confidence estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3437–3445. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv:2103.10385*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Qi Zhang, and Xuanjing Huang. 2021. Template-free prompt tuning for few-shot NER. *CoRR*, abs/2109.13532.

Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pages 185–201. Springer.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 617–628. ACM / IW3C2.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1003–1011. The Association for Computer Linguistics.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Adversarial over-sensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 486–496. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 58–65. Association for Computational Linguistics.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11054–11070.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models. In *7th IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing (WACV/MOTION 2005), 5-7 January 2005, Breckenridge, CO, USA*, pages 29–36. IEEE Computer Society.

Benjamin Rosenfeld and Ronen Feldman. 2007. Using corpus statistics on entities to improve semi-supervised relation extraction from the web. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 600–607, Prague, Czech Republic. Association for Computational Linguistics.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification: Evaluation data and classification schemes. *Trans. Assoc. Comput. Linguistics*, 9:691–706.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing (EMNLP)*.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the TACRED dataset. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13843–13850. AAAI Press.

Ang Sun, Ralph Grishman, and Satoshi Sekine. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of the 49th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 521–529, Portland, Oregon, USA. Association for Computational Linguistics.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Tongtong Wu, Guitao Wang, Jinming Zhao, Zhaoran Liu, Guilin Qi, Yuan-Fang Li, and Gholamreza Haffari. 2022. Towards relation extraction from speech. *CoRR*.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020a. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.

Haiyang Yu, Ningyu Zhang, Shumin Deng, Hongbin Ye, Wei Zhang, and Huajun Chen. 2020b. Bridging text and knowledge with multi-prototype embedding for few-shot relational triple extraction. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6399–6410. International Committee on Computational Linguistics.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022a. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3999–4006. ijcai.org.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3016–3025. Association for Computational Linguistics.

Ningyu Zhang, Xin Xu, Liankuan Tao, Haiyang Yu, Hongbin Ye, Shuofei Qiao, Xin Xie, Xiang Chen, Zhoubo Li, Lei Li, et al. 2022b. Deepke: A deep learning based knowledge extraction toolkit for knowledge base population. *arXiv preprint arXiv:2201.03335*.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. PRGC: Potential relation and global correspondence based joint relational triple extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6225–6235, Online. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 50–61. Association for Computational Linguistics.

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

A Implementation Details

A.1 Settings

We detail the training procedures and hyperparameters for each of the datasets. We utilize PyTorch to conduct experiments with one NVIDIA RTX 3090 GPU. All optimization is performed with the

AdamW optimizer (Loshchilov and Hutter, 2019). The training is always continuous for 10 epochs without validation. All pre-trained language models used in this work are downloaded from HuggingFace. The names of PLMs are "hfl/chinese-roberta-wwm-ext-large" for DuIE2.0 and CMeIE, "dmis-lab/biobert-large-cased-v1.1" for SciERC and ChemProtm, and "roberta-large" for other benchmark datasets.

A.2 Prompting Methods

In the prompt-based tuning experiments with *KnowPrompt* (PyTorch-Lightning), the early stop in the original code is dropped. The learning rate is set as $4e-5$ for all datasets. Instead of using the original code for multi-labeled DialogRE with BCEloss, we implement experiments with DialogRE the same as the other seven datasets to unify our benchmark.

A.3 Balancing Methods

For re-sampling methods, we firstly use the sampler on all 10% and 100% imbalanced training sets to get nearly balanced training sets and then use them in all methods the same way as imbalanced datasets. We leverage the official code of various re-weighting losses and provide the alternative parsing argument named "useloss" for developers to choose them.

A.4 Data Augmentation

Different DA methods mentioned in §3.3 are utilized on English and Chinese datasets via *nlpaug* and *nlpcda*. After generating augmented data, we merge them with original data in order to delete repeated instances that make no sense. Then both original and augmented data are combined and fed into models to evaluate their performance.

A.5 Self-training

Given unlabeled data $\mathcal{D}^{\mathrm{U}}$ and a few labeled data $\mathcal{D}^{\mathrm{L}}$, we conduct self-training for semi-supervised learning. The scheme is executed as the following steps (Huang et al., 2021):

1. Train a teacher model Θ^{T} with gold-labeled data $\mathcal{D}^{\mathrm{L}}$ via cross-entropy.

2. Use the trained teacher model Θ^{T} to generate soft labels on unlabeled data $\mathcal{D}^{\mathrm{U}}$:

$$\tilde{y}_i = f_{\Theta^{\mathrm{T}}}(\tilde{x}_i), \quad \tilde{x}_i \in \mathcal{D}^{\mathrm{U}} \tag{2}$$

Relation	Number	Level
Other	1145	-
Entity-Destination (e1,e2)	686	Many
Cause-Effect (e2,e1)	536	
Member-Collection (e2,e1)	498	
Entity-Origin (e1,e2)	462	
Message-Topic (e1,e2)	399	
Component-Whole (e2,e1)	383	
Component-Whole (e1,e2)	382	
Instrument-Agency (e2,e1)	331	
Product-Producer (e2,e1)	321	
Content-Container (e1,e2)	304	
Cause-Effect (e1,e2)	280	
Product-Producer (e1,e2)	263	Medium
Content-Container (e2,e1)	135	
Entity-Origin (e2,e1)	121	
Message-Topic (e2,e1)	117	
Instrument-Agency (e1,e2)	79	Few
Member-Collection (e1,e2)	64	
Entity-Destination (e2,e1)	1	

Table 5: Relation splits on SemEval.

Relation	Number	Level
USED-FOR	1690	Many
CONJUNCTION	400	
EVALUATE-FOR	313	Medium
HYPONYM-OF	298	
PART-OF	179	Few
FEATURE-OF	173	
COMPARE	166	

Table 6: Relation splits on SciERC.

3. Train a student model Θ^{S} via cross-entropy $\mathcal{L}$ on both gold-labeled data $\mathcal{D}^{\mathrm{L}}$ and soft-labeled data $\mathcal{D}^{\mathrm{SU}}$. The loss function of Θ^{S} is:

$$\begin{aligned}\mathcal{L}_{\mathrm{STU}} =& \frac{1}{|\mathcal{D}^{\mathrm{L}}|} \sum_{x_i \in \mathcal{D}^L} \mathcal{L}(f_{\Theta^{\mathrm{S}}}(x_i), y_i) \\ &+ \frac{\lambda_{\mathrm{U}}}{|\mathcal{D}^{\mathrm{U}}|} \sum_{\tilde{x}_i \in \mathcal{D}^U} \mathcal{L}(f_{\Theta^{\mathrm{S}}}(\tilde{x}_i), \tilde{y}_i)\end{aligned} \tag{3}$$

where λ_{U} is the weighting hyper-parameter, and we set it 0.2 in this work. It is an alternative to iterate from Step 1 to Step 3 multiple times by initializing Θ^{T} in Step 1 with newly learned Θ^{S} in Step 3. We only perform self-training once in our experiments for simplicity because the result is not good, and it is not sensitive to continue the next iteration.

B Class Splits in Balancing Methods Evaluation

The few-level, medium-level and many-level relation splits based on the number of each relation class on SemEVal and SciERC are shown in Table 5 and Tabel 6 for comparative experiments on different re-weighting losses in §5.2.