Continual Contrastive Finetuning Improves Low-Resource Relation Extraction

Wenxuan Zhou[†], **Sheng Zhang**[‡], **Tristan Naumann**[‡], **Muhao Chen**[†] and **Hoifung Poon** [‡]

[†]University of Southern California, [‡]Microsoft Research

{zhouwenx,muhaoche}@usc.edu
{zhang.sheng,tristan,hoifung}@microsoft.com

Abstract

Relation extraction (RE), which has relied on structurally annotated corpora for model training, has been particularly challenging in low-resource scenarios and domains. Recent literature has tackled low-resource RE by self-supervised learning, where the solution involves pretraining the entity pair embedding by RE-based objective and finetuning on labeled data by classification-based objective. However, a critical challenge to this approach is the gap in objectives, which prevents the RE model from fully utilizing the knowledge in pretrained representations. In this paper, we aim at bridging the gap and propose to pretrain and finetune the RE model using consistent objectives of contrastive learning. Since in this kind of representation learning paradigm, one relation may easily form multiple clusters in the representation space, we further propose a multi-center contrastive loss that allows one relation to form multiple clusters to better align with pretraining. Experiments on two document-level RE datasets, BioRED and Re-DocRED, demonstrate the effectiveness of our method. Particularly, when using 1% end-task training data, our method outperforms PLM-based RE classifier by 10.5% and 6.1% on the two datasets, respectively.

1 Introduction

Relation extraction (RE) is a fundamental task in NLP. It aims to identify the relations among entities in a given text from a predefined set of relations. While much effort has been devoted to RE in supervised settings (Zhang et al., 2017, 2018; Nan et al., 2020), RE is extremely challenging in high-stakes domains such as biology and medicine, where annotated data are comparatively scarce due to overly high annotation costs. Therefore, there is a practical and urgent need for developing low-resource RE models without the reliance on large-scale end-task annotations.

To realize low-resource RE, previous work has focused on pretraining entity pair embedding on large corpora using RE-based pretraining objectives. Particularly, Baldini Soares et al. (2019) propose a self-supervised matching-the-blanks (MTB) objective that encourages embeddings of the same entity pairs in different sentences to be similar. Later work (Peng et al., 2020; Qin et al., 2021) extends this idea with distant supervision (Mintz et al., 2009) and improves representation learning using contrastive learning (Hadsell et al., 2006; Oord et al., 2018; Chen et al., 2020). To adapt to training on RE annotations, these works finetune pretrained entity pair embedding on labeled data using classification-based objectives. Although this paradigm produces better results compared to RE models initialized with pretrained language models (PLMs), it creates a significant divergence between pretraining and finetuning objectives, thus preventing the model from fully exploiting knowledge in pretraining.

In this paper, we aim to bridge this gap in RE pretraining and finetuning. Our key idea is to use similar objectives in pretraining and finetuning. First, we propose to continually finetune pretrained embedding by contrastive learning, which encourages the entity pair embeddings corresponding to the same relation to be similar. However, as pretraining and finetuning are conducted on different tasks, entity pairs of the same relation can form multiple different clusters in the pretrained embedding, where standard supervised contrastive loss (Khosla et al., 2020) may distort the representation because of its underlying one-cluster assumption (Graf et al., 2021). Therefore, we further propose a multi-center contrastive loss (MCCL), which encourages an entity pair to be similar to only a subset of entity pairs of the same relation, allowing one relation to form multiple clusters. Second, we propose to use class-wise k-nearest neighbors (kNN; Khandelwal et al.

2020, 2021) in inference, where predictions are made based on most similar instances.

We focus our work on document-level RE (Jia et al., 2019; Yao et al., 2019), which consists of both intra- and cross-sentence relations. To the best of our knowledge, this work represents the first effort to explore self-supervised pretraining for document-level RE. Unlike prior studies (Peng et al., 2020; Qin et al., 2021), we do not use distant supervision. Instead, we pretrain entity pair embedding with an improved MTB objective on unlabeled corpora, where we use contrastive learning to learn representations that suit downstream RE. We then finetune the pretrained model on labeled data with MCCL. Experiments on two datasets, BioRED (Luo et al., 2022) in the biomedical domain and Re-DocRED (Tan et al., 2022b) in the general domain, demonstrate that our pretraining and finetuning objectives significantly outperform baseline methods in low-resource settings. Particularly, in the low-resource setting of using 1% of labeled data, our method outperforms PLM-based classifiers by 10.5% and 6.1% on BioRED and Re-DocRED, respectively. Based on our pretrained representations, MCCL outperforms classification-based finetuning by 6.0% and 4.1%, respectively. We also find observe that as more data becomes available, the performance gap between MCCL and classification-based finetuning diminishes.

Our technical contributions are three-fold. First, we propose to pretrain the PLMs based on our improved MTB objective and show that it significantly improves PLM performance in low-resource document-level RE. Second, we present a technique that bridges the gap of learning objectives between RE pretraining and finetuning with continual contrastive finetuning and kNN-based inference, helping the RE model leverage pretraining knowledge. Third, we design a novel MCCL finetuning objective, allowing one relation to form multiple different clusters, thus further reducing the distributional gap between pretraining and finetuning.

2 Related Work

Document-level RE. Existing document-level RE models can be classified into graph-based and sequence-based models. Graph-based models construct document graphs spanning across sentence boundaries and use graph encoders such as the graph convolution network (GCN; Kipf and Welling 2017) to aggregate information. Particularly, Quirk and Poon (2017) build document graphs using words as nodes with inner- and inter-sentence dependencies (e.g., syntactic dependencies, coreference, etc.) as edges. Later work extends this idea by applying different network structures (Peng et al., 2017; Jia et al., 2019) or introducing other node types and edges (Christopoulou et al., 2019; Nan et al., 2020; Zeng et al., 2020). On the other hand, sequence-based methods (Zhou et al., 2021; Zhang et al., 2021; Tan et al., 2022a) use PLMs to learn cross-sentence dependencies without using graph structures. Particularly, Zhou et al. (2021) propose to enrich relation mention representation by localized context pooling. Zhang et al. (2021) propose to model the inter-dependencies between relation mentions by semantic segmentation (Ronneberger et al., 2015). In this work, we study a general method of self-supervised RE. Therefore, our method is independent of the model architecture and can be adapted to different RE models.

Low-resource RE. Labeled RE data may be scarce in real-world applications, especially in low-resource and high-stakes domains such as finance and biomedicine. Much effort has been devoted to training RE models in low-resource settings. Some work tackles low-resource RE by indirect supervision, which solves RE by other tasks such as machine reading comprehension (Levy et al., 2017), textual entailment (Sainz et al., 2021), and abstractive summarization (Lu et al., 2022). However, indirect supervision may not be practical in high-stake domains, where annotated data for other tasks are also scarce. Other efforts (Baldini Soares et al., 2019; Peng et al., 2020; Qin et al., 2021) improve low-resource RE by pretraining on large corpora with RE-based objectives. Specifically, Baldini Soares et al. (2019) propose an MTB objective that encourages embeddings of the same entity pairs in different sentences to be similar. Peng et al. (2020) propose to pretrain on distantly labeled corpora, where they make embeddings of entity pairs with the same distant label to be similar. They also introduce a contrastive learning based training objective to improve representation learning. Qin et al. (2021) further introduce an entity discrimination task and pretrain the RE model on distantly labeled document corpora. In this paper, we study self-

supervised pretraining for document-level RE. We study how to reduce the gap between pretraining and finetuning, which is critical to bridge the training signals obtained in these two stages but has been overlooked in prior work.

3 Method

In this work, we study a self-supervised approach for document-level RE. Given a document d and a set of entities $\{e_i\}_{i=1}^N$, where each entity e_i has one or multiple entity mentions in the document, document-level RE aims at predicting the relations of all entity pairs $(e_s, e_o)_{s,o \in \{1,...,N\}}$ from a predefined set of relationships $\mathcal{R}$ (including an NA class indicating no relation exists), where e_s and e_o are the subject and object entities, respectively. In the self-supervised RE setting, we have a large unlabeled document corpus for pretraining and a labeled RE dataset for finetuning. The document corpus has been annotated with entity mentions and the associated entity types but no relations. Our goal is to train a document-level RE classifier, especially in the low-resource setting.

Our training pipeline consists of two phases: pretraining and finetuning. In pretraining, we use the (unlabeled) document corpus to pretrain the entity pair embedding based on our improved *matching-the-blanks* training objective (MTB; Baldini Soares et al. 2019), where the LM learns to decide whether two entity pair embeddings correspond to the entity pairs or not, and the learning of representation is enhanced with contrastive learning. In finetuning, we continue to train the pretrained model on relation-labeled data using a multi-center contrastive loss (MCCL), which achieves better performance than the traditional classifier paradigm due to its better-aligned learning objective with pretraining. After training, we use classwise k-nearest neighbor (kNN) inference that suits well the contrastively finetuned model.

The rest of this section is organized as follows: we introduce the model architecture used in both pretraining and finetuning in Section 3.1, the pretraining process in Section 3.2, finetuning in Section 3.3, and inference in Section 3.4.

3.1 Model Architecture

Encoder. Given a document $d = [x_1, x_2, ..., x_l]$, we first mark the spans of the entity mentions by adding special entity markers [E] and [/E] to the start and the end of each mention. Then we encode the document with a PLM to get the contextual embedding of textual tokens:

$$\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_l] = \text{PLM}([x_1, x_2, ..., x_l]).$$

We take the contextual embedding of [E] at the last layer of the PLM as the embedding of entity mentions. We accumulate the embedding of mentions corresponding to the same entity by Log-SumExp pooling (Jia et al., 2019) to get the entity embedding $\boldsymbol{h}_{e_i}$.

Entity pair embedding. Given an entity pair $t = (e_s, e_o)$ in document d, where e_s and e_o are the subject and object entities, respectively, we calculate the entity pair embedding by:

$$\boldsymbol{z}^t = \boldsymbol{W}_{\text{linear}} \left[\boldsymbol{h}_{e_s}, \boldsymbol{h}_{e_o}, \boldsymbol{c}^{(e_s, e_o)} \right].$$

Here $\boldsymbol{h}_{e_s}, \boldsymbol{h}_{e_o} \in \mathbb{R}^d$ are embeddings of subject and object entities, $\boldsymbol{c}_{e_s, e_o} \in \mathbb{R}^d$ is the localized context encoding for (e_s, e_o), $\boldsymbol{W}_{\text{linear}} \in \mathbb{R}^{3d \times d}$ is a linear projector. The localized context encoding is introduced by Zhou et al. (2021) to derive the context embedding conditioned on an entity pair, which finds the context that both the subject and object entities attend to. Specifically, denote the multi-head attention in the last layer of PLM as $\boldsymbol{A} \in \mathbb{R}^{m \times l \times l}$, where m is the number of attention heads, l is the input length, we first take the attention scores from [E] as the attention from each entity mention, then accumulate the attention of this entity mention by mean pooling to get the entity-level attention $\boldsymbol{A}^{(e_i)} \in \mathbb{R}^{m \times l}$. Finally, we compute $\boldsymbol{c}^{(e_s, e_o)}$ by:

$$\boldsymbol{A}^{(e_s, e_o)} = \boldsymbol{A}^{(e_s)} \odot \boldsymbol{A}^{(e_o)},$$
$$\boldsymbol{q}^{(e_s, e_o)} = \sum_{i=1}^{m} \boldsymbol{A}_i^{(e_s, e_o)},$$
$$\boldsymbol{a}^{(e_s, e_o)} = \boldsymbol{q}^{(e_s, e_o)} / \boldsymbol{1}^\intercal \boldsymbol{q}^{(e_s, e_o)},$$
$$\boldsymbol{c}^{(e_s, e_o)} = \boldsymbol{H}^\intercal \boldsymbol{a}^{(e_s, e_o)}.$$

We introduce in the rest of the section how to pretrain and finetune the RE model based on the entity pair embedding $\boldsymbol{z}^{(e_s, e_o)}$.

3.2 Pretraining

We pretrain the LM on the document corpus using the MTB objective. MTB is based on a simple assumption that, in contrast to different entity pairs, it is more frequent for the same entity pair to be connected with the same relation. The

MTB objective transforms the similarity learning problem into a pairwise binary classification problem: given two relation-describing utterances where entity mentions are masked, the model classifies whether the entity pairs are the same or not. This pretraining objective has shown effectiveness in several sentence-level RE datasets(Zhang et al., 2017; Hendrickx et al., 2010; Han et al., 2018).

However, when it comes to document-level RE, Qin et al. (2021) have observed no improvement led by the vanilla MTB pretraining. Therefore, we replace the pairwise binary classification with contrastive learning, which is adopted in later RE pretraining works (Peng et al., 2020; Qin et al., 2021) and can effectively learn from more positive and negative examples. Details of training objectives are elaborated in the rest of the section. We introduce the details of data preprocessing of the pretraining corpus in Appendix A.

Training objective. The overall goal of pretraining is to make the embedding of the same entity pair from different documents more similar than different entity pairs. For clarity, we call two same entity pairs from different documents as a positive pair, and two different entity pairs as a negative pair. We use the InfoNCE loss (Oord et al., 2018) to model this objective. Given the documents in batch, $\mathcal{P}$ as the set of all positive pairs, and $\mathcal{N}_t$ denote the set of entity pairs different to t, the contrastive MTB loss is[1]:

$$\mathcal{L}_{\text{rel}} = -\frac{1}{|\mathcal{P}|} \sum_{t_i, t_i \in \mathcal{P}} \log \frac{e^{\text{sim}(\boldsymbol{z}^{t_i}, \boldsymbol{z}^{t_j})/\tau}}{\mathcal{Z}_{t_i}}, \qquad (1)$$

$$\mathcal{Z}_{t_i} = e^{\text{sim}(\boldsymbol{z}^{t_i}, \boldsymbol{z}^{t_j})/\tau} + \sum_{t_k \in \mathcal{N}_{t_i}} e^{\text{sim}(\boldsymbol{z}^{t_i}, \boldsymbol{z}^{t_k})/\tau},$$

where $\text{sim}(\boldsymbol{z}^{t_i}, \boldsymbol{z}^{t_j})$ denotes the similarity between the embeddings of t_i and t_j, and τ is a temperature hyperprameter. Following Chen et al. (2020), we use cosine similarity as the similarity metric. Similar to SimCSE (Gao et al., 2021), we further add a self-supervised contrastive loss that requires the same entity pair embedding augmented by different dropout masks to be similar, thus encouraging the model to learn more instance-discriminative features that lead to less collapsed representations. Specifically, denote the two entity pair embeddings of t derived by different dropout masks as

---

[1]Similar to Baldini Soares et al. (2019), we randomly mask the entities in documents with a probability of 0.7 to avoid shortcut learning.

$\boldsymbol{z}^t$ and $\hat{\boldsymbol{z}}^t$, respectively, the set of all entity pairs in the batch as $\mathcal{T}$, and the set of entity pairs in positive pairs as $\mathcal{T}_P$, the self-supervised loss is:

$$\mathcal{L}_{\text{self}} = -\frac{1}{|\mathcal{T}_P|} \sum_{t_i \in \mathcal{T}_P} \log \frac{e^{\text{sim}(\boldsymbol{z}^{t_i}, \hat{\boldsymbol{z}}^{t_i})/\tau}}{\mathcal{Z}_{t_i}},$$

$$\mathcal{Z}_{t_i} = e^{\text{sim}(\boldsymbol{z}^{t_i}, \hat{\boldsymbol{z}}^{t_i})/\tau} + \sum_{t_k \in \mathcal{T}\setminus\{t_i\}} e^{\text{sim}(\boldsymbol{z}^{t_i}, \hat{\boldsymbol{z}}^{t_k})/\tau}.$$

Finally, we use a masked language model loss $\mathcal{L}_{\text{mlm}}$ to adapt the LM to the document corpus. The overall pretraining objective is:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{rel}} + \mathcal{L}_{\text{self}} + \mathcal{L}_{\text{mlm}}.$$

For faster convergence, we initialize our model with a PLM that is pretrained on a larger corpus, and continually pretrain the PLM on the document corpus with our new pretraining objectives. We use BERT (Devlin et al., 2019) for the general domain and PubmedBERT (Gu et al., 2021) for the biomedical domain.

3.3 Finetuning

After pretraining, we finetune the LM on labeled document-level RE datasets. In previous studies (Baldini Soares et al., 2019; Peng et al., 2020; Qin et al., 2021), pretraining and finetuning are conducted in processes with different learning objectives. Specifically, after using the pretrained weights to initialize a RE classifier, the model is finetuned with a classification-based training objective. Based on our model architecture, a straightforward finetuning method is to add a softmax classifier upon the entity pair embedding, for which a cross-entropy loss for a batch of entity pairs $\mathcal{T}$ is formulated as:

$$P_r^{t_i} = \text{softmax}(\boldsymbol{W}_r \boldsymbol{z}^{t_i} + b_r),$$

$$\mathcal{L}_{\text{ce}} = -\frac{1}{|\mathcal{T}|} \sum_{t_i \in \mathcal{T}} \log(P_{y_{t_i}}^{t_i}),$$

where y_t is the ground-truth label for entity pair t, $\boldsymbol{W}_r, b_r$ are the weight and bias of the classifier. Though this approach has shown improvements, it may produce sub-optimal outcomes from MTB pretraining since it implicitly assumes that entity pairs corresponding to the same relation are in the same cluster, while MTB pretraining may learn multiple clusters for a relation. For example, the entity pairs *(Honda Corp., Japan)* and *(Mount Fuji, Japan)*, although likely to be expressed with

Classifier	BioRED	Re-DocRED
One-cluster		
Softmax	28.6	39.3
Nearest centroid	12.5	4.1
Multi-cluster		
classwise kNN	**36.7**	**54.1**

Table 1: Probing results (in F_1) on the test set of BioRED and Re-DocRED.

the same relation *country* in documents, are likely to be in different clusters since MTB views them as negative pairs due to different subject entities. Therefore, we propose an MCCL objective that can bridge these gaps. Next, we will discuss the distributional assumption of the softmax classifier as well as supervised contrastive loss, then present our MCCL objective.

Distributional assumption. We conduct a probing analysis on the distribution of pretrained representations to further justify the multi-cluster assumption. Specifically, we fix the weights of the pretrained MTB model and fit different classifiers on top of it, including a softmax classifier, a nearest centroid classifier (both assuming one cluster for a relation), and a classwise kNN classifier (assuming multiple clusters for a relation). We evaluate these classifiers on the test set. Results are shown in Table 1. We find that classwise kNN greatly outperforms others, showing that MTB pretraining learns multiple clusters for a relation.

Therefore, to accommodate this multi-cluster assumption, we need to finetune the representations with a training objective that suits multiple clusters for each relation. Beside using the softmax classifier with cross-entropy loss, we also consider supervised contrastive loss (SupCon; Khosla et al. 2020; Gunel et al. 2021). SupCon has a similar loss form to InfoNCE in Eq. (1), except that it uses instances of the same/different relations as positive/negative pairs. However, previous work (Graf et al., 2021) has shown that both softmax and SupCon are minimized when the representations of each class collapse to the vertex of a regular simplex. In our case, this means the entity pair embeddings corresponding to the same relation in pretraining collapses to a single point, which creates a distributional gap between pretraining and finetuning.

Training objective. We thereby propose the MCCL objective. Given entity pairs $\mathcal{T}$ and sets of entity pairs grouped by their relations $\{\mathcal{T}_r\}_{r \in \mathcal{R}}$, our loss is formulated as:

$$w_r^{(t_i, t_j)} = \frac{e^{\mathrm{sim}(\boldsymbol{z}^{t_i}, \boldsymbol{z}^{t_j})/\tau_1}}{\sum_{t_k \in \mathcal{T}_r \setminus \{t_i\}} e^{\mathrm{sim}(\boldsymbol{z}^{t_i}, \boldsymbol{z}^{t_k})/\tau_1}},$$

$$s_r^{t_i} = \sum_{t_j \in \mathcal{T}_r \setminus \{t_i\}} w_r^{(t_i, t_j)} \mathrm{sim}(\boldsymbol{z}^{t_i}, \boldsymbol{z}^{t_j}),$$

$$P_r^{t_i} = \mathrm{softmax}((s_r^{t_i} + b_r)/\tau_2),$$

$$\mathcal{L}_{\mathrm{mccl}} = -\frac{1}{|\mathcal{T}|} \sum_{t_i \in \mathcal{T}} \log(P_{y_{t_i}}^{t_i}),$$

where τ_1 and τ_2 are temperature hyperparameters, $b_r \in \mathbb{R}$ is the classwise bias. The loss calculation can be split into two steps. First, we calculate the similarity between t_i and relation r, which is a weighted average of the similarity between t_i and $t_j \in \mathcal{T}_r$ such that a more similar t_j has a larger weight. Next, we use the cross-entropy loss to make the similarity of ground-truth relation larger than others. In this way, MCCL only optimizes t_i to be similar to a few closest entity pairs of the ground-truth relation, and thus encourages multiple clusters in entity pair embedding. Note that MCCL can be easily extended to support multi-label classification scenarios, for which details are given in Appendix B.

Proxies. We use batched training for finetuning, where entity pairs in the current batch are used to calculate MCCL. However, it is possible that a subset of relations in $\mathcal{R}$, especially the long-tail relations, are rare or missing in the current batch. When $\mathcal{T}_r \setminus \{t_i\}$ is empty, $s_r^{t_i}$ and MCCL become undefined. To tackle this problem, we propose the use of proxies (Movshovitz-Attias et al., 2017; Zhu et al., 2022). We add one proxy vector p_r for each relation r, which is a trainable parameter and associated with an embedding $\boldsymbol{z}_r^p$. We incorporate the proxies into MCCL by changing $\mathcal{T}_r$ to $\mathcal{T}_r' = \mathcal{T}_r \cup \{p_r\}$, ensuring that $\mathcal{T}_r' \setminus \{t_i\}$ is never empty in training and preventing MCCL from becoming undefined. The proxies are randomly initialized and updated during training by backward propagation.

3.4 Inference

We use the classwise kNN (Christobel and Sivaprakasam, 2013) for inference, which predicts relations based on similarly represented instances and thus aligns with our contrastive finetuning objective. Given a new entity pair to predict, we first

find k most similar instances[2] in the training data of each relation (including NA), then calculate the average cosine similarity of each relation s_r^{avg}. Finally, the model returns the relation with the maximum $s_r^{\text{avg}} + b_r$ for single-label prediction, and all relations with higher $s_r^{\text{avg}} + b_r$ than NA for multi-label prediction. We use classwise kNN because it is more suitable for RE datasets, where the label distribution is usually long-tailed (Zhang et al., 2019).

4 Experiments

We evaluate our proposed method with a focus on low-resource RE (Sections 4.1-4.3), and present detailed analyses (Section 4.4) and visualization (Section 4.5) to justify method design choices.

4.1 Datasets

We conduct experiments with two document-level RE datasets. The **BioRED** dataset (Luo et al., 2022) is a manually labeled single-label RE dataset in the biomedical domain. The entity pairs are classified into 9 types (including an NA type indicating no relation). It has a training set consisting of 400 documents, which we use in finetuning. For pretraining, we use the PubTator Central corpus (Wei et al., 2019), which annotates the PubMed corpus with entity mentions and their named entity types. The **Re-DocRED** dataset (Tan et al., 2022b) is a multi-label large-scale dataset of the general domain. It is a relabeled version of the DocRED dataset (Yao et al., 2019). Re-DocRED addresses the incomplete annotation issue of DocRED, where a large percentage of entity pairs are mislabeled as NA. The entity pairs in Re-DocRED are classified into 97 types (incl. NA). It has a training set consisting of 3,053 documents, which we use in finetuning. For pretraining, we use the distantly labeled training set provided by DocRED, which consists of 101,873 documents. We remove the relation labels and use our improved MTB to pretrain the model.

4.2 Experimental Setup

Model configurations. We implement our models using Hugging Face Transformers (Wolf et al., 2020). We use AdamW (Loshchilov and Hutter, 2018) in optimization with a weight decay of 0.01. During pretraining, we use a batch size of 16, a learning rate of 5e-6, a temperature of 0.05, and epochs of 3 and 10 for BioRED and DocRED, respectively. During finetuning, we use a batch size of 32, a learning rate of 5e-5, and epochs of 100 and 30 for BioRED and DocRED, respectively. The temperatures in MCCL are set to $\tau_1 = \tau_2 = 0.2$ for BioRED and $\tau_1 = 0.01, \tau_2 = 0.03$ for DocRED. We search k from $\{1, 3, 5, 10, 20\}$ for classwise kNN using the development set[3]. We run experiments with Nvidia V100 GPUs.

Evaluation settings. In this work, in addition to the standard full-shot training, we consider low-resource settings. To create each of the settings, we randomly sample a fixed proportion $p\%$ of the entity pairs from the training set as our training data, and use the original test set for evaluation. We use the same evaluation metrics as the original papers. We use micro-F_1 for BioRED, and micro-F_1 and micro-F_1-Ign for Re-DocRED. The micro-F_1-Ign removes the relational facts in the test set that have appeared in training.

Compared methods. We experiment with the following finetuning objectives: (1) **Lazy learning**, which directly uses the pretrained embedding and training data to perform kNN without finetuning; (2) **Cross-entropy loss** (CE), which adds a softmax classifier on top of PLM and uses cross-entropy loss to finetune the model; (3) **Supervised contrastive loss** (SupCon); and (4) **Multi-center contrastive loss** (MCCL). In inference, classwise kNN is used for all methods except for CE. Note that as SupCon does not apply to multi-label scenarios, we only evaluate it on BioRED. For each objective, we also evaluate the PLM before and after MTB pretraining. We use different PLMs as the backbone of the model, namely PubmedBERT$_{\text{BASE}}$ for BioRED and BERT$_{\text{BASE}}$ for Re-DocRED, which are pretrained on the biomedical and general domains, respectively.

4.3 Main Results

The results on the test sets of Re-DocRED and BioRED are shown in Table 2 and Table 3, respectively. All results are averaged for five runs of training using different random seeds. Overall, the combination of MTB and MCCL achieves the best performance in low-resource settings where 1%, 5%, and 10% of relation-labeled data are used. Further, when using the same MTB-based

---

[2]Measured by cosine similarity. If a relation has fewer than k entity pairs in training data, we use all of them.

[3]For low-resource setting with $p\%$ of training data, we sample $p\%$ of development data as the development set.

Encoder	Objective	1%		5%		10%		100%	
		F_1	F_1-Ign	F_1	F_1-Ign	F_1	F_1-Ign	F_1	F_1-Ign
PLM	Lazy	15.6	14.9	20.1	19.4	21.6	19.2	28.7	28.0
	CE	40.3	38.9	54.1	52.6	61.3	60.3	70.9	69.4
	MCCL	44.7	43.1	59.1	57.5	63.2	61.8	68.2	66.7
MTB	Lazy	35.2	34.4	44.7	43.4	47.3	46.2	54.1	52.9
	CE	42.3	40.7	57.9	56.4	62.9	61.4	**71.2**	**69.9**
	MCCL	**46.4**	**44.5**	**59.7**	**58.2**	**63.8**	**62.1**	69.3	67.9

Table 2: Results on the test set of Re-DocRED.

Encoder	Objective	1%	5%	10%	100%
PLM	Lazy	14.5	17.6	18.8	28.3
	CE	24.1	35.4	42.5	57.7
	SupCon	20.0	30.9	38.0	52.2
	MCCL	20.8	41.3	45.5	55.1
MTB	Lazy	24.3	28.4	34.4	36.7
	CE	28.6	41.2	49.8	**61.5**
	SupCon	24.4	29.1	31.4	43.1
	MCCL	**34.6**	**48.5**	**54.2**	60.8

Table 3: F_1 on the test set of BioRED.

Pretraining Objective	1%	10%	100%
PLM	20.8	45.5	55.1
vanilla MTB	22.9	45.0	56.0
our MTB	34.6	54.2	60.8
w/o $\mathcal{L}_{rel}$	21.0	47.1	56.7
w/o $\mathcal{L}_{self}$	24.1	49.3	58.6
w/o $\mathcal{L}_{mlm}$	32.9	50.2	58.2

Table 4: F_1 on the test set of BioRED with different pretraining objectives. We use MCCL in finetuning.

4.4 Ablation Study

Pretraining objectives. We analyze the effectiveness of our proposed pretraining losses in Section 3.2. To do so, we pretrain the model with one loss removed at a time while keeping the finetuning setup on BioRED fixed with the MCCL. The results are shown in Table 4. Overall, we observe that all losses are effective. If we remove all proposed techniques and use the vanilla MTB pretraining objective of binary pairwise classification, the results are only slightly better or even worse. Among the techniques, removing $\mathcal{L}_{rel}$ leads to the largest performance drop, showing that MTB-based pretraining is critical to improve low-resource RE. Removing $\mathcal{L}_{self}$ also leads to a large performance drop. It is because $\mathcal{L}_{self}$ encourages the model to learn more discriminative features that lead to less collapsed representations. Our finding aligns with recent studies in computer vision (Islam et al., 2021; Chen et al., 2022), showing that reducing collapsed representations with self-supervised contrastive learning improves the transferability to downstream tasks.

Performance w.r.t. different temperatures. We discuss the impact of two temperatures in MCCL. In MCCL, τ_1 controls the weighting of instances. With a very small τ_1, each instance will only form a cluster with its nearest neighbor in the batch,
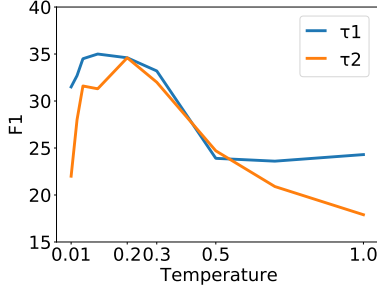
representations, MCCL shows better results than CE in low-resource settings. It shows that in low-resource settings, MCCL can better leverage the pretraining knowledge with a well-aligned finetuning objective. However, this improvement diminishes when abundant labeled data are available, as MCCL underperforms CE on both datasets with full training data on both datasets. In addition, we observe that MTB pretraining consistently improves MCCL and CE on both datasets. These results demonstrate the effectiveness of MTB pretraining for more precise document-level RE with less needed end-task supervision.

Considering other training objectives, we observe that lazy learning produces meaningful results. On both datasets, the results of lazy learning based on MTB with 10% of data are comparable to finetuning with 1% of data. This shows that the entity pair embedding pretrained on unlabeled corpora contains knowledge that can be transferred to unseen relations. We also observe that SupCon using kNN-based inference underperforms both CE and MCCL on BioRED, showing that its one-cluster assumption hurts the knowledge transfer.

Figure 1: F_1 using different temperatures on 1% of BioRED.

Figure 2: F_1 under different percentages of BioRED training data.

while with very large τ_1, instances of the same relation will collapse to the same cluster. τ_2 controls the importance of hard instances, which is also used in other contrastive losses (e.g., τ in Eq. (1)). Wang and Liu (2021) observe that small τ_2 makes the model focus more on hard instances, while Khosla et al. (2020) observe that too small τ_2 leads to numerical instability. We show the results of using different temperatures in Figure 1, where we keep one temperature fixed and change the other. For τ_1, we find that using large temperature harms the performance, showing that our multi-cluster assumption improves low-resource RE. For τ_2, we observe that both small and large values impair the performance, which is aligned with prior observations.

Performance w.r.t. different amount of data. The main results show that MCCL outperforms CE in the low-resource setting, while slightly underperforming CE when full training data is used. We further evaluate MCCL and CE using different amounts of end-task data. We experiment on BioRED and use the entity pair embedding pretrained with MTB. Results are shown in Figure 2. We observe that MCCL consistently outperforms CE by a large margin when less than 20% of training data is used, while it performs similarly or worse than CE after that. It again demonstrates the effectiveness of MCCL in low-resource RE. However, as the pretraining and finetuning are based on different tasks, fully adapting the model to downstream data by CE results in similar or better performance in data-sufficient scenarios.

4.5 Visualization

Figure 3 shows the t-SNE (Van der Maaten and Hinton, 2008) projection of entity pair embedding finetuned with different objectives on BioRED. For clarity, we visualize the embedding of the four most frequent relations in BioRED with differ-
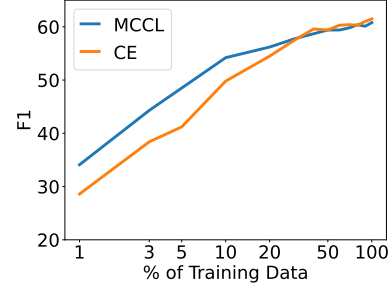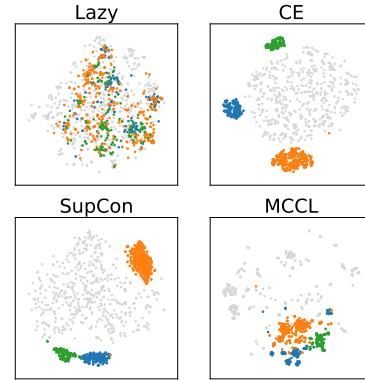

Figure 3: Visualization of entity pair embedding finetuned with different objectives on BioRED. NA instances are shown in grey.

ent colors, including the NA class shown in grey. The visualization shows that both CE and SupCon learn one cluster for each relation, while lazy learning and MCCL, as expected, generate multiple small clusters for a relation. This observation indicates that MCCL can better align with the pretraining objective, further explaining its better performance in low-resource settings.

5 Conclusion

In this paper, we study self-supervised learning for document-level RE. Our method conducts an improved MTB pretraining objective that acquires cheap supervision signals from large corpora without relation labels. To bridge the gap between pretraining and end-task finetuning, we propose a continual contrastive finetuning objective, in contrast to prior studies that typically use classification-based finetuning, and use kNN-based inference. As pretrained representation may form multi-cluster representation, we further propose a multi-center contrastive loss that aligns well with the nature of the pretrained representation. Extensive experiments on two document-level RE datasets demonstrate the effectiveness of

these key techniques in our method. Future work is adapting our method to other tasks in information extraction, such as n-ary relation extraction, named entity recognition, typing, and linking.

Limitations

The main limitation of MCCL is the requirement of a sufficiently large batch size in training (32 documents in our experiments), leading to a need for large GPU memory. This is because MCCL uses in-batch entity pairs for contrastive learning, and a small batch size does not provide enough instances to form multiple clusters. In addition, we need to store the entity pair embedding of the whole training set for kNN-based inference, which is less memory-efficient than CE.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. 2022. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pages 3090–3122. PMLR.

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.

Y Angeline Christobel and P Sivaprakasam. 2013. A new classwise k nearest neighbor (cknn) method for the classification of diabetes dataset. *International Journal of Engineering and Advanced Technology*, 2(3):396–200.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. 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 4925–4936, Hong Kong, China. Association for Computational Linguistics.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. 2021. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

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

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. 2021. A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8845–8855.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multi-scale representation learning. 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 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

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

Keming Lu, I-Hung Hsu, Mingyu Derek Ma, Wenxuan Zhou, and Muhao Chen. 2022. Summarization as indirect supervision for relation extraction. In *Findings of ACL: EMNLP*.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

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

Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. 2017. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. 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 3350–3363, Online. Association for Computational Linguistics.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain. Association for Computational Linguistics.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2022b. Revisiting docred - addressing the overlooked false negative problem in relation extraction. *arXiv preprint arXiv:2205.12696*.

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

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, 47(W1):W587–W593.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. 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 *IJCAI*.

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, Volume 1 (Long and Short Papers)*, pages 3016–3025, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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, Copenhagen, Denmark. 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*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *AACL-IJCNLP*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. 2022. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917.

Appendices

A Data Preparation

We acquire positive and negative pairs from the document corpus. We regard two entity pairs $(e_{s_1}, e_{o_1}), (e_{s_2}, e_{o_2})$ in different documents as a positive pair if they share the same subject and object entities, respectively (i.e., $e_{s_1} = e_{s_2}, e_{o_1} = e_{o_2}$), and otherwise negative.

However, for a large corpus, the number of such positive pairs is enormous. For instance, in biomedical RE pretraining, we extract 37 billion positive pairs in total. Using all these pairs in pretraining is computationally infeasible. Therefore, we select positive pairs as follows. Denote the number of documents mentioning an entity e or an entity pair (e_s, e_o) as $N(e)$ and $N(e_s, e_o)$, respectively, we use two metrics, frequency $= N(e_s, e_o)$

and $\text{PMI} = \frac{N(e_s, e_o)}{N(e_s) \times N(e_o)}$, to measure the popularity of entity pairs The frequency measures how often e_s and e_o co-occur. The PMI measures whether e_s and e_o have a strong association. In pretraining, we first discard the entity pairs with frequency $< N_{\text{threshold}}$, and then use the positive pairs constituted by the top K entity pairs measured by their PMIs. We set the frequency threshold to be 16 and 3 for BioRED and DocRED, respectively, and use the top 5,000 entity pairs in pretraining.

Besides, as MTB is fully self-supervised, the information of whether two relations mentions correspond to the same relation type is not available, but it is assumed that at least entity pairs with different subject or object types are likely to be of different relation types and can therefore be used as negative pairs. Such use of entity types to filter the pairs has indeed been shown a strong feature for RE (Zhong and Chen, 2021; Zhou and Chen, 2022). We only use two entity pairs with different subject or object entity types as negatives. While the entity type based filtering may also discard some hard negatives, our experiment (see Section C) shows improved results, meaning that its benefits outweigh the disadvantages.

B Adaptation to Multi-label RE

It is noteworthy that in some RE tasks, such as DocRED, one entity pair may have multiple relation labels, in which case the cross-entropy loss does not apply. Therefore, for multi-label scenarios, we substitute cross-entropy loss (also the softmax in MCCL) with the adaptive thresholding loss proposed by Zhou et al. (2021). Specifically, denote the logits as l (the input to softmax in cross-entropy loss), the set of positive relations as $\mathcal{P}$ (except NA), and the set of the remaining relations except for NA as $\mathcal{N}$, the adaptive thresholding loss is formulated as:

$$\mathcal{L}_1 = -\sum_{r \in \mathcal{P}} \log \left(\frac{e^{l_r}}{\sum_{r' \in \mathcal{P} \cup \{\text{NA}\}} e^{l_{r'}}} \right),$$

$$\mathcal{L}_2 = -\log \left(\frac{e^{l_{\text{NA}}}}{\sum_{r' \in \mathcal{N} \cup \{\text{NA}\}} e^{l_{r'}}} \right),$$

$$\mathcal{L}_{\text{at}} = \mathcal{L}_1 + \mathcal{L}_2.$$

This loss encourages the logits of positive relations to be higher than NA, and the logits of other relations to be lower than NA. In prediction, the model returns the relations with higher logits than
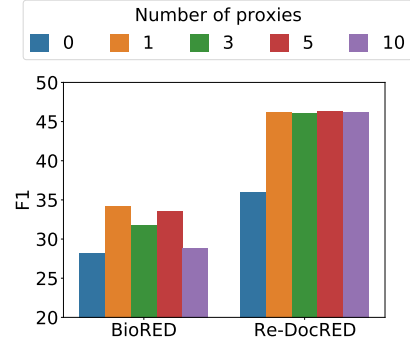

Figure 4: F_1 achieved with different numbers of proxies on 1% BioRED.

NA as predictions, or return NA if none of such relations exist.

C More Experiments

Performance w.r.t. number of proxies. We evaluate MCCL with different number of proxies. When no proxy is used, we ignore the relations that do not appear in the current batch. The F_1 on both BioRED and Re-DocRED in the 1% low-resource setting is shown in Figure 4, indicating that adding proxies improves F_1 significantly on both datasets. Using one proxy for each relation achieves an increase of 6.0% in F_1 on BioRED, and a larger increase of 10.2% in F_1 on Re-DocRED. Such a difference of increment is due to the fact that Re-DocRED is more long-tailed, where 97% of instances are NA compared to 80% in BioRED. We also observe that adding more proxies achieves similar or even worse results. These results make sense as the proxies are mainly in the place of long-tail relations that do not appear in the batch, and these relations contain too few instances to form multiple clusters.

Coarse-to-fine evaluation. To give another illustration of showing that MCCL learns multiple clusters, we experiment with it on 1% of BioRED in a coarse-to-fine setting. Specifically, we merge all relations except NA into one relation in finetuning, and apply kNN inference using the original labels. We find that MCCL achieves an F1 of 30.3%, which is even better than CE with all relations provided. However, if we remove the instance weights in MCCL to degrade it to one-cluster, the F_1 constantly degrades in finetuning. It shows that multi-cluster assumption helps preserve the fine-grained relation information in pretrained representation.

Pretraining Objective	1%	10%	100%
PLM	20.8	45.5	55.1
vanilla MTB	22.9	45.0	56.0
our MTB	34.6	54.2	60.8
w/o entity type filtering	25.1	48.6	58.1
Replace $\mathcal{L}_{rel}$ by $\mathcal{L}_{mccl}$	34.7	52.5	58.8

Table 5: F_1 on the test set of BioRED.

Other ablation studies. We analyze the effectiveness of entity type filtering in Section A. Results are shown in Table 5. Removing entity type filtering degrades performance significantly. It shows that entity type filtering can remove a lot of false negatives in pretraining and greatly improves the pretrained model.

Besides, as the main results have demonstrated the effectiveness of MCCL in finetuning, we wonder whether MCCL can also lead to improved pretraining. To do so, we replace the InfoNCE loss in Eq. (1) by MCCL and regard different entity pairs as different classes. The results are comparable or slightly worse in contrast to using $\mathcal{L}_{rel}$, showing that the multi-cluster assumption of MCCL does not necessarily help pretraining.

A For every submission:

☑ A1. Did you describe the limitations of your work?
At the end of the paper.

☒ A2. Did you discuss any potential risks of your work?
Our work studies general information extraction techniques and does not have ethical issues.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

In section 4

☑ B1. Did you cite the creators of artifacts you used?
In section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In section 4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In section 4

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In section 4

C ☑ Did you run computational experiments?

In section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In section 4

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We report the mean results In section 4.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D   ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.