A Unified Label-Aware Contrastive Learning Framework for Few-Shot Named Entity Recognition

Haojie Zhang, **Yimeng Zhuang**

Samsung R&D Institute China-Beijing
tayee.chang@gmail.com, ym.zhuang@hotmail.com

Abstract

Few-shot Named Entity Recognition (NER) aims to extract named entities using only a limited number of labeled examples. Existing contrastive learning methods often suffer from insufficient distinguishability in context vector representation because they either solely rely on label semantics or completely disregard them. To tackle this issue, we propose a unified label-aware token-level contrastive learning framework. Our approach enriches the context by utilizing label semantics as suffix prompts. Additionally, it simultaneously optimizes context-context and context-label contrastive learning objectives to enhance generalized discriminative contextual representations. Extensive experiments on various traditional test domains (OntoNotes, CoNLL'03, WNUT'17, GUM, I2B2) and the large-scale few-shot NER dataset (FEW-NERD) demonstrate the effectiveness of our approach. It outperforms prior state-of-the-art models by a significant margin, achieving an average absolute gain of 7% in micro F1 scores across most scenarios. Further analysis reveals that our model benefits from its powerful transfer capability and improved contextual representations [1].

1 Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that involves extracting entity spans from input sentences and assigning them the correct entity types, such as location, person, and organization Sang and De Meulder [2003]. This information is crucial for downstream tasks like question answering, information retrieval, knowledge graph He *et al.* [2021], machine translation, etc. However, the process of labeling data is time-consuming and labor-intensive Huang *et al.* [2020]; Cui *et al.* [2021]. Oftentimes, only a few labeled examples are available, and these examples may come from a different domain, posing a significant challenge due to data sparsity Huang *et al.* [2020]; Ding *et al.* [2021]. Consequently, there is growing

attention on how to effectively adapt models to handle unseen data in such scenarios.

Recently, there have been two emerging trends in existing methods. One direction revolves around prompt technology. Prompt-based methods focus on bridging the gap between pre-training and fine-tuning by incorporating suitable prompts. These methods exhibit remarkable performance in the few-shot domain, positioning them as promising candidates for both few-shot and zero-shot learning scenarios Ding *et al.* [2021]; Chen *et al.* [2022]. However, prompt-based methods must carefully design templates Cui *et al.* [2021]; Ma *et al.* [2022a].

Metric learning is an innovative technology that involves calculating the distance metric between a query token and the reference, and subsequently assigning the corresponding entity type to the query token Hou *et al.* [2020]; Yang and Katiyar [2020]. Recent advancements in this field have leveraged contrastive learning techniques Das *et al.* [2021], leading to state-of-the-art performance. However, it is important to acknowledge that metric learning methods still have limitations. One notable limitation is their tendency to overlook the implicit semantic information underlying entity types, which leads to insufficient exploration of the relationships between different entity types. As a result, the contextual representations may not be sufficiently discriminative.

To tackle these challenges, we propose a label-aware contrastive learning approach that aims to enhance the discriminative power of contextual representations. To enhance the context, we convert entity types into natural language forms and use them as suffix prompts. We also employ both context-to-context and context-to-label contrastive learning, leveraging the connection between context and its corresponding label for improved learning. By expanding the context using natural language forms and integrating context-to-label contrastive learning, our approach achieves more robust and discriminative contextual representations. As a result, our model demonstrates significant improvements in handling few-shot NER scenarios and addressing data sparsity issues.

Our approach builds upon the contrastive learning method proposed by Das *et al.* [2021] to learn improved semantic space representations. However, we take three additional steps to address the aforementioned issues. In summary, our contributions can be outlined as follows:

- To thoroughly extract the implicit semantic information

---

[1] The source code is available at https://github.com/TayeeChang/Unified-Framework-for-FS-NER
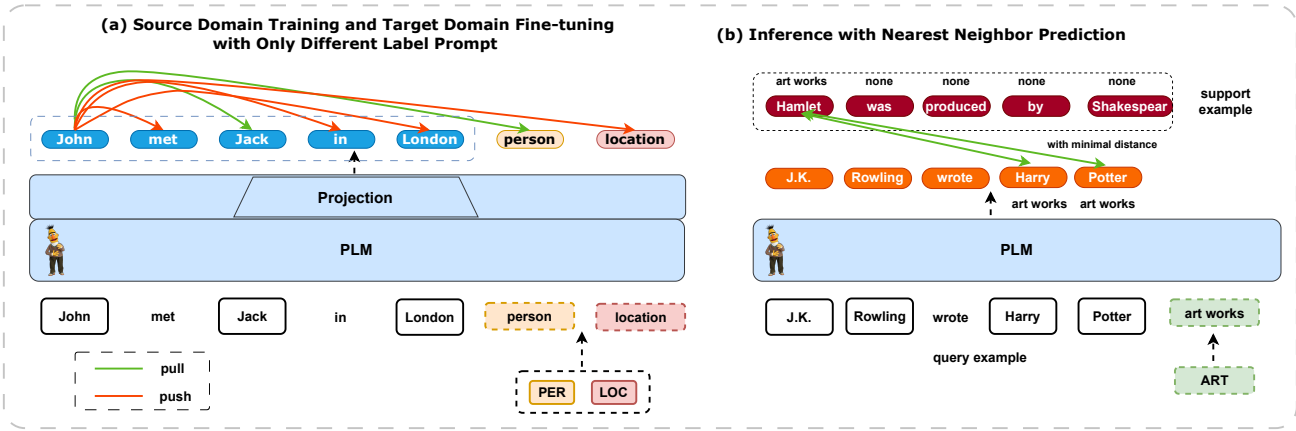
Figure 1: An overview of the architecture of our proposed model. (a) During the training and fine-tuning process in the source domain, the fine-tuning follows a similar approach as training, but with a different label prompt. Utilizing contrastive learning, tokens belonging to the same entity types are attracted toward each other, while tokens representing different entity types are pushed apart. This encourages the model to learn a more distinct and effective representation of entity-specific information. The contrastive learning includes two aspects: context-context and context-label. (b) Inference process with nearest neighbor prediction. Similarity scores between query tokens and support tokens will be calculated according to the distance metric.

associated with entity types, we employ label information as a suffix prompt. In particular, we transform the abbreviated form of the entity type into its corresponding natural language description and append it to the original input sentence. This approach allows us to effectively capture the complete meaning conveyed by the entity types.

- In addition to constructing context-to-context contrastive learning, we also incorporate context-to-label contrastive learning into our approach. This expansion enables us to leverage the relationship between the context and its associated label, further enhancing the learning process.

- We have enhanced the original contrastive loss function at a low cost, which can be seen as an advantageous benefit without any additional expense.

2 Related Work

Meta Learning Originating from computer vision, meta learning aims to learn how to learn across tasks Vanschoren [2018]. Matching Network Vinyals *et al.* [2016] employing ideas from the metric learning, was first proposed to adapt to the few-shot scenario. After that, many metric learning based methods are proposed. Prototypical Network Snell *et al.* [2017] was proposed to aim to learn the prototype representations by computing the average representations of unseen entity types given a support set. Yang and Katiyar [2020] proposed NNShot and StructShot, which employ the nearest neighbor prediction and Viterbi decoding to inference. Recently, metric learning based methods have become more and more popular in NLP tasks such as text classification Bao *et al.* [2019]; Geng *et al.* [2019], named entity recognition Ding *et al.* [2021]; Yang and Katiyar [2020]; Das *et al.* [2021]; Huang *et al.* [2022] and machine learning Gu *et al.* [2018].

Prompt Technology Recently, prompting has become a new paradigm in NLP Liu *et al.* [2023]; Brown *et al.* [2020]. Many prompt based methods are proposed for varied NLP tasks and show great potentiality in few-shot NER. TemplateNER Cui *et al.* [2021] utilizes prompt for the entity spans and scores these entity candidates. DemonstrationNER Lee *et al.* [2021] introduces task demonstrations for in-context learning. Similar to TemplateNER, EntLM Ma *et al.* [2021] converts sequence labeling task to pre-training LM task, which abandons the prompt but needs to introduce label words relying on a specific domain. Ma *et al.* [2022a] proposed that utilizing label semantics can improve the effect of few-shot NER models especially in the 1-shot scenarios. However, it completely depends on label information to learn the label representations. Huang *et al.* [2022] inherits the idea of Ma *et al.* [2022a] by aligning contextual token representations with label representations in the same encoder. But the alignment is limited as it doesn't consider that the contextual tokens include more alignment information.

Few-shot NER It aims to adapt the model to unseen entity types. Existing few-shot NER methods present two trends, metric learning based and prompt based. With the great success of prompt technology, many prompt based methods are proposed to apply in the few-shot scenario Cui *et al.* [2021]; Lee *et al.* [2021]; Ma *et al.* [2021]. However, these methods require an elaborate prompt design. Recently, label semantics have been proven to be effective for few-shot NER. Ma *et al.* [2022a]; Huang *et al.* [2022] introduce label semantics to learn the label representations by aligning the contextual token representations and label representations. However, both of them ignore those contextual token representations that include more alignment information. As a result, as shown in Ma *et al.* [2022a], with the number of shots increasing, such simple alignment brings marginal effects. Metric learning based methods are more popular and gradually become mainstream in recent years Snell *et al.* [2017]; Yang and Kati-

yar [2020]; Hou *et al.* [2020]. In order to better simulate the reality situation, Ding *et al.* [2021] proposed a new dataset FEW-NERD designed for the few-shot scenario. Inspired by the feature extractors and nearest neighbor inference Wang *et al.* [2019], learning representations and employing nearest neighbor inference seems to have become a standard configuration. As a result, the key point of metric learning based methods is to learn more generalized representations of context tokens. Das *et al.* [2021] utilizes contrastive learning to learn to distinguish between different tokens. Introducing Gaussian Embedding proves to be more effective in adapting to different domains and achieving current SOTA results in few-shot NER tasks.

3 Problem Formulation

NER is generally considered as a sequence labeling task Sang and De Meulder [2003]; Chiu and Nichols [2016]. For the input sentence $x = \{x_1, x_2, ..., x_n\}$, NER aims to assign a certain label y_i for each token x_i. Generally, $y_i \in \mathcal{C}$, $\mathcal{C}$ refers to a specific label set like {person, location, organization, other}. Here, other means that a token doesn't belong to any entity type.

3.1 Few-Shot NER

It's pretty easy to achieve state-of-the-art performance by employing the pre-training language model as the basic encoder when the data resource $\mathcal{D}_H$ is large. But when we face the case of low resource data $\mathcal{D}_L$, which means only a few examples can be obtained, it poses a new challenge.

Formally, for few-shot NER, the source domain label set is $\mathcal{C}_H$ and the target domain label set is $\mathcal{C}_L$, $\mathcal{C}_L \cap \mathcal{C}_H = \emptyset$. Since the target label set is unseen for the source label set, it is very challenging for the model to predict the right label sequence. Thus, it requires the model to have a great transfer capability. Different the high resource data set scenario, for the training and evaluation, we follow the N-way K-shot setting Ding *et al.* [2021], which means the label set size is N, and for each label, there are K examples.

3.2 Evaluation Protocols

Typically, there are two main evaluation protocols for few-shot NER tasks. One is the episode evaluation. For each episode or task, a support and query pair set is sampled from the high resource data set according to the N-way K-shot setting Ding *et al.* [2021]. The model is trained and evaluated on this downsampling data for T turns. The final micro-F1 is calculated over all the episodes. The other one is the low resource evaluation Yang and Katiyar [2020]. Instead of training and evaluation on each episode, the model is trained on the downsampling N-way K-shot dataset but evaluated on the whole standard hold-out test set. The operation can also be implemented for T turns and the average value of the micro-F1 is the final metric value. For the traditional datasets (e.g., OntoNotes, CoNLL'03, WNUT'17, GUM, I2B2), we follow the low resource evaluation protocol and for the FEW-NERD dataset, which is explicitly designed for the few-shot NER episode task, we adopt the episode evaluation protocol.

3.3 Tagging Scheme

For fair comparison, we adopt the IO tagging scheme for all our experiments, where I-type represents a token that is inside an entity and O-type represents a token that doesn't belong to any entity type.

4 Method

We propose a novel approach for few-shot Named Entity Recognition (NER) that integrates metric-based contrastive learning and prompt technology within a unified framework. Our method involves a two-stage training process, starting with training on the source domain and then fine-tuning on the target domain. The key aspect of our proposed method lies in its unified framework, which combines metric-based contrastive learning and prompt technology. Additionally, we have improved upon the original contrastive learning scheme by making slight adjustments to the contrastive loss. This effectively enhances model training and provides a significant boost in performance without any additional cost.

As demonstrated in Figure 1, we first train our model in source domain spaces. Next, we finetune our model in the target domain spaces. The fine-tuning process shares the same methods as the training process. Finally, the nearest neighbor prediction is employed to predict the label sequence of the query set in the target domain.

4.1 Source Domain Training

Figure 1 illustrates our unified framework, which combines the original context and label prompt using contrastive learning technology. This technology encompasses two key aspects: 1) context-context contrastive learning and 2) context-label contrastive learning. The goal of contrastive learning is to distinguish the differences between different token representations, including both context-context and context-label relationships. Label semantics have been proven effective for named entity recognition, particularly in few-shot scenarios Ma *et al.* [2022a]. To incorporate label semantic information, our model employs a suffix prompt. Firstly, we convert abbreviated-form labels to their natural language descriptions to obtain sufficient semantic information. For example, we manually map labels like ⟨PER⟩ to ⟨person⟩, ⟨LOC⟩ to ⟨location⟩, ⟨ORG⟩ to ⟨organization⟩, etc. The label prompt is constructed by arranging the natural language label sequences separated by a special token [CLS]. Similar to BERT Devlin *et al.* [2018], the [CLS] token represents the entire input sentence and can be used as the representation of the label itself. The context text and label prompt are then concatenated and fed into a pre-trained language model (PLM). A projection layer is applied to project the semantic representations into a different representative space. Finally, we utilize the context-context and context-label contrastive learning losses to train our model effectively. By combining contrastive learning with label prompts, our unified framework enhances the discriminative power of contextual representations for few-shot NER tasks.

Given a context sequence of n tokens $[x_{c_1}, x_{c_2}, ..., x_{c_n}]$, by label prompt technology, we can get the whole input tokens $[x_{c_1}, x_{c_2}, ..., x_{c_n}, x_{l_1}, x_{l_2}, ..., x_{l_k}]$, where k is the whole label

Algorithm 1 Fine-tuning in the target domain
___
Input: Support Data $\mathcal{X}_{sup}$, Encoder PLM, Projection head f_μ, f_Σ;
Output: PLM, μ, Σ;
 1: $\mathcal{L}_{prev} \in \mathbb{R}_+$ (infinity)
 2: $\mathcal{L} = \mathcal{L}_{prev} - 1$
 3: **repeat**
 4:    $\mathcal{L}_{prev} = \mathcal{L}$
 5:    **for** all $(x_i, y_i) \in \mathcal{X}_{sup}$ **do**
 6:      Calculate ℓ_i as in Eq. (1), Eq. (5), Eq. (7);
 7:    **end for**
 8:    Calculate $\mathcal{L}_{\text{context-context}}$ as in Eq. (6)
 9:    Calculate $\mathcal{L}_{\text{context-label}}$ as in Eq. (8)
10:    Calculate $\mathcal{L}$ as in Eq. (9)
11:    Update PLM, f_μ, f_Σ by gradient descent to reduce $\mathcal{L}$
12: **until** $\mathcal{L} > \mathcal{L}_{prev}$
___

number including entity and none-entity and for simplicity, x_{l_i} represents i-th label token. Here, for simplicity, we omit the word-piece tokenization and the special separative tokens [CLS] and [SEP].

After feeding to the PLM, we can get the last hidden layer outputs of PLM as the representations of input tokens.

$$[h_i, h_2, ...h_m] = \text{PLM}([x_1, x_2, ..., x_m]) \qquad (1)$$

We adopt a projection head for mapping the representations to a different space like Das *et al.* [2021].

$$\mu_i = f_\mu(h_i), \Sigma_i = f_\Sigma(h_i) \qquad (2)$$

For context-context contrastive learning, we adopt the Jensen–Shannon divergence between valid context tokens in the same batch. JS-divergence is as follows:

$$d(p, q) = \frac{1}{2}(\mathcal{D}_{\text{KL}}(\mathcal{N}_p || \mathcal{N}_q) + \mathcal{D}_{\text{KL}}(\mathcal{N}_q || \mathcal{N}_p)) \qquad (3)$$

where D_{KL} refers to the Kullback-Leibler divergence. The contrastive loss ContaiNER Das *et al.* [2021] employed is as follows:

$$\ell(p) = -\log \frac{\sum_{(x_q, y_q) \in \chi_p} \exp(-d(p, q))/ |\chi_p|}{\sum_{(x_q, y_q) \in \chi, p \neq q} \exp(-d(p, q))} \qquad (4)$$

Here, we propose a simple variant,

$$\ell_{\text{out}}(p) = -\frac{1}{|\chi_p|} \sum_{(x_q, y_q) \in \chi_p} \log \frac{\exp(-d(p,q))}{\sum_{(x_q, y_q) \in \chi, p \neq q} \exp(-d(p,q))} \qquad (5)$$

For p-th valid token (x_p, y_p), the positive sample set χ_p is composed of tokens with the same label as (x_p, y_p) in the batch set χ.

The context-context contrastive loss is as follows:

$$\mathcal{L}_{\text{context-context}} = \frac{1}{|\chi_{\text{context}}|} \sum_{p \in \chi_{\text{context}}} \ell_{out}(p) \qquad (6)$$

The context-label contrastive loss is as follows:

$$\ell_{\text{context-label}}(p) = -\log \frac{\exp(-d(p, q)/\tau)}{\sum_{k, k \in \mathcal{C}} \exp(-d(p, k)/\tau)} \qquad (7)$$

Dataset	Domain	# Sentences	# Entity
FEW-NERD	Wikipedia	188k	66
CoNLL'03	News	20.7k	4
GUM	Wiki	3.5k	11
WNUT	Social	5.6k	6
OntoNotes	Mixed	76k	18
I2B2'14	Medical	140K	23

Table 1: Summary Statistics of Original Datasets

where $y_p = q$, q is the prompt token and $\mathcal{C}$ is the prompt label set. For the context-label contrastive loss, there is only one positive example of the token x_p, which is its golden label token in the prompt.

$$\mathcal{L}_{\text{context-label}} = \frac{1}{|\chi_{\text{context}}|} \sum_{p \in \chi_{\text{context}}} \ell_{\text{context-label}}(p) \qquad (8)$$

As it's very easy to keep tabs on the saturation point to use the temperature coefficient when employing JS-divergence as the distance metric. Hence we set τ as 1.

The whole training loss will be:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{context-context}} + (1 - \alpha) \cdot \mathcal{L}_{\text{context-label}} \qquad (9)$$

4.2 Fine-tuning in the Target Domain

After completing the training phase in the source domain, the next step is to fine-tune the model for adaptation to a different target domain. Since the target domain typically has only a limited number of examples, we gather all the examples into a single batch for fine-tuning. The fine-tuning process closely resembles the training process. In the 5-shot setting, we utilize the same training setting as used during training. However, in the 1-shot setting, we find better results by fine-tuning the model using the Euclidean distance metric instead of the Jensen–Shannon divergence. Additionally, in the 1-shot setting, we omit the context-context contrastive loss as it does not contribute significantly to the training process. This suggests that the label semantic information is more crucial in the 1-shot setting. It is worth noting that due to the limited number of examples in the target domain, we implement an early stopping criterion based on contrastive loss to prevent overfitting. Algorithm 1 provides a detailed illustration of the entire process.

4.3 Inference Process

During the inference period, we adopt the nearest neighbor prediction to get the predicted label sequence of the query example Wang *et al.* [2019]. Similar to CONTaiNER Das *et al.* [2021], we only use the PLM hidden layer outputs as the token representations, so the projection heads f_μ, f_Σ are abandoned in this period.

More specifically, we first obtain the representation of the support token $h_{x'}$ from the support set and the representation of a query token h_x according to Eq. (1). Then, for each query token, we calculate its Euclidean distance between its

Model	1-shot					5-shot				
	I2B2	CoNLL	WNUT	GUM	Avg.	I2B2	CoNLL	WNUT	GUM	Avg.
ProtoBERT[†]	13.4 ± 3.0	49.9 ± 8.6	17.4 ± 4.9	17.8 ± 3.5	24.6	17.9 ± 1.8	61.3 ± 9.1	22.8 ± 4.5	19.5 ± 3.4	30.4
NNShot[†]	15.3 ± 1.6	61.2 ± 10.4	22.7 ± 7.4	10.5 ± 2.9	27.4	22.0 ± 1.5	74.1 ± 2.3	27.3 ± 5.4	15.9 ± 1.8	34.8
StructShot[†]	21.4 ± 3.8	62.4 ± 10.5	24.2 ± 8.0	7.8 ± 2.1	29.0	30.3 ± 2.1	74.8 ± 2.4	30.4 ± 6.5	13.3 ± 1.3	37.2
DecomposedModel Ma *et al.* [2022b]	-	46.1 ± 0.4	25.1 ± 0.2	17.5 ± 1.0	-	-	58.2 ± 0.9	31.0 ± 1.3	31.4 ± 0.9	-
CONTaiNER Das *et al.* [2021]	16.4 ± 1.7	57.8 ± 10.7	24.2 ± 2.9	17.9 ± 1.8	29.1	24.1 ± 1.9	72.8 ± 2.0	27.7 ± 2.2	25.2 ± 2.7	37.3
COPNER Huang *et al.* [2022]	34.6 ± 1.8	67.0 ± 3.8	33.8 ± 2.5	-	-	41.1 ± 1.6	74.9 ± 2.9	34.8 ± 3.1	-	-
Our model	$\textbf{38.4} \pm \textbf{5.0}$	$\textbf{72.8} \pm \textbf{4.3}$	$\textbf{35.7} \pm \textbf{4.5}$	$\textbf{23.9} \pm \textbf{7.2}$	**42.7**	$\textbf{46.1} \pm \textbf{2.1}$	$\textbf{77.6} \pm \textbf{1.8}$	$\textbf{35.8} \pm \textbf{2.1}$	$\textbf{33.7} \pm \textbf{4.7}$	**48.3**

Table 2: F1 scores with standard deviations for Domain Transfer Evaluation. [†] denotes the results reported by Das *et al.* [2021]. For some baselines, some values are not reported from the original paper and they are marked as -. For a fair comparison, we directly use sampled shots as in Yang and Katiyar [2020]. The best results are in **bold**.

PLM representation and support tokens. Finally, we will assign the query token to the label of its nearest support token.

$$y^{\star} = \arg\min_{y \in \mathcal{Y}} \min_{x' \in \mathcal{S}_y} d(h_x, h_{x'})$$
$$d(h_x, h_{x'}) = \|h_x - h_{x'}\|^2 \qquad (10)$$

where h_x denotes x-th context token in the query set and $h_{x'}$ denotes the context token in the support set.

5 Experiments Setups

5.1 Datasets

Our model is evaluated with different datasets. We utilize 6 different datasets: OntoNotes 5.0 Weischedel *et al.* [2013], WNUT-2017 Derczynski *et al.* [2017], CoNLL-2003 Sang and De Meulder [2003], I2B2-2014 Derczynski *et al.* [2017], GUM Zeldes [2017] and FEW-NERD Ding *et al.* [2021]. These datasets come from different domains. The first 5 datasets are Mixed, Social, News, Medical, and Wiki. As the entity distribution of these traditional datasets is usually considered uneven, it's not suitable for the few-shot scenario. Hence Ding *et al.* [2021] proposed a new dataset FEW-NERD. The FEW-NERD dataset is specially developed for the few-shot NER task, including INTRA and INTER two sub-tasks with a grand total of 66 fine-grained labels across 8 coarse-grained classes. A summary of statistics of these datasets is shown in Table 1.

5.2 Evaluation Settings

We employ two different evaluation settings to assess the performance of our model: the domain transfer setting and the FEW-NERD setting.

Domain Transfer Setting This setting evaluates the transfer capacity of the model from one domain to another different domain. In the few-shot scenario, the model transfer capacity is critically important. So in this setting, we evaluate our model on different datasets. We consider ONTONOTES 5.0, which is a mixed dataset, as the source domain. We first train our model on the ONTONOTES 5.0 and fine-tune the model on the CoNLL-2003, WNUT-2017, I2B2-2014, and GUM. For each experiment, 5 different support sets are evaluated by the micro-f1 score with mean and standard deviation reported. We show these results in Table 2. We see that our model outperforms prior SOTAs with a significant margin, even with existing great domain gaps between OntoNotes and

target domains (I2B2, WNUT, GUM). When the source domain and target domain (CoNLL) are closer, our model performs considerable performance like supervised learning.

FEW-NERD Setting We evaluate our model on the FEW-NERD dataset Ding *et al.* [2021]. This dataset is the first of its kind, specifically designed for the few-shot NER task, and it is also one of the largest manually annotated datasets available. It comprises 66 fine-grained and 8 coarse-grained entity types, making it more challenging compared to traditional datasets that are down-sampled in few-shot settings. Additionally, the dataset includes two subtasks. 1) **INTRA**, where fine-grained entity types in train, dev, and test sets belong to different coarse-grained entity types. To get it straight, there is no overlap between train, dev, and test sets in terms of coarse-grained entity types. 2) **INTER**, where fine-grained entity types in train, dev, and test sets share the same coarse-grained entity types. Obviously, the FEW-NERD (INTRA) task presents greater challenges compared to the FEW-NERD (INTER) task due to limitations in sharing coarse-grained entity types. Moreover, the FEW-NERD dataset adopts the episode learning approach within the meta-learning framework. Each episode or task is constructed based on the N-way K-shot setting. Hence, in the FEW-NERD scenario, we strictly follow this episode evaluation protocol. We assess our model's performance on all episodes of the two distinct sub-tasks and calculate the final micro-F1 score across all episodes. As shown in Table 3 and Table 4, our approach achieves new SOTA results in the leaderboard in both of the sub-tasks.

5.3 Implementation Details

We implement our model by `Tensorflow 2.4`. We leverage `bert-base-uncased` as our contextual encoder for all experiments. For the optimizer, we use AdamW Loshchilov and Hutter [2017] with a weight decay rate of 0.01, excluding the Norm layer and the bias terms. We train our mode with a learning rate of 5×10^{-5} for both training and fine-tuning processes, batch size of 32, and maximum sequence length of 128. For a fair comparison, we set the Gaussian Embedding dimension fixed to $l = 128$. For the hyperparameter α, we perform a grid search and select the best setting with the validation set. For the training process, we only run one epoch, as we find it will fall into the pattern collapse if continuing to train. For the label prompt, we manually convert the labels in abbreviation to the natural lan-

Model	5-way		10-way		Avg.
	1~2 shot	5~10 shot	1~2 shot	5~10 shot	
ProtoBERT†	23.45	41.93	19.76	34.61	29.94
NNShot†	31.01	35.74	21.88	27.67	29.08
StructShot†	35.92	38.83	25.38	26.39	31.63
DecomposedModel Ma *et al.* [2022b]	52.04	63.23	43.50	**56.84**	53.90
CONTaiNER Das *et al.* [2021]	40.43	53.70	33.84	47.49	43.87
COPNER Huang *et al.* [2022]	53.52	58.74	44.13	51.55	51.99
Our model	**56.51**	**65.66**	**46.42**	56.82	**56.35**

Table 3: F1 scores on FEW-NERD (INTRA). † denotes the results reported in Das *et al.* [2021]. The best results are in **bold**.

Model	5-way		10-way		Avg.
	1~2 shot	5~10 shot	1~2 shot	5~10 shot	
ProtoBERT†	44.44	58.80	39.09	53.97	49.08
NNShot†	54.29	50.56	46.98	50.00	50.46
StructShot†	57.33	57.16	49.46	49.39	53.34
DecomposedModel Ma *et al.* [2022b]	**68.77**	71.62	63.26	68.32	67.99
CONTaiNER Das *et al.* [2021]	55.95	61.83	48.35	57.12	55.81
COPNER Huang *et al.* [2022]	65.39	67.59	59.69	62.32	63.75
Our model	67.73	**71.81**	**63.93**	**68.65**	**68.03**

Table 4: F1 scores on FEW-NERD (INTER). † denotes the results reported in Das *et al.* [2021]. The best results are in **bold**.

guage descriptions in all the datasets. We run all experiments on an NVIDIA A6000 GPU.

6 Main Results

In this session, we present and analyze the comprehensive results of few-shot NER using two evaluation protocols. Additionally, we perform ablation experiments to demonstrate the effectiveness of our proposed method.

6.1 Baselines

We compare our model with various state-of-the-art few-shot NER models. These baseline models can be broadly categorized into two groups: prompt-based approaches and metric learning-based approaches.

- **ProtoBERT** is based on the Prototypical NetWork Snell *et al.* [2017] and employs BERT Devlin *et al.* [2018] as the encoder, which aims to learn the prototype of entity types. It belongs to the family of metric learning.

- **NNShot** is also based on metric learning, which aims to learn the token-level reference Yang and Katiyar [2020]. It leverages the nearest neighbor prediction for inference.

- **StructNN** leverages a Viterbi decoder using the source domain to build the abstract transition probability, compared to NNShot Yang and Katiyar [2020].

- **DecomposedModel** is a decomposed meta-learning approach that conducts entity span detection and entity typing respectively Ma *et al.* [2022b].

- **CONTaiNER** is based on contrastive learning, which makes them attractive for tokens of the same entity type and repulsive for tokens of the different entity types Das *et al.* [2021].

- **COPNER** leverages label information as the prompt and calculates the distance between the input token and the label token Huang *et al.* [2022]. It is one of the prompt-based families.

	INTRA	INTER
Ours w/ OCL + w/o context-label CL	57.53	63.12
Ours w/ ICL + w/o context-label CL	64.35	69.62
Ours w/ ICL + w/ context-label CL	**65.66**	**71.81**

Table 5: Ablation study: F1 scores on FEW-NERD 5-way 5~10-shot are reported.

6.2 Overall Results

Table 2-4 shows the overall results of our approach along with those reported by the previous SOTAs. Our approach demonstrates significant performance improvements over prior methods, indicating its effectiveness. Specifically, when compared to CONTaiNER, our approach achieves an impressive increase of up to +16.08 F1 scores on the FEW-NERD Evaluation (INTRA, 5-way 1~2-shot) and +22 F1 scores on the challenging Domain Transfer Evaluation (I2B2). The I2B2 dataset (Medical), with a different domain and its 23 entity types, is more challenging, further emphasizing the robustness of our model in handling challenging scenarios. Upon closer analysis of the FEW-NERD evaluation results, an interesting phenomenon emerges. In contrast, our approach exhibits a larger improvement in the 1~2 shot scenario compared to the 5~10 shot setting when compared to CONTaiNER. Similarly, when compared to COPNER, our method shows a greater margin of improvement in the 5~10 shot setting compared to the 1~2 shot setting. Contrastive learning-based methods show effectiveness in utilizing multiple few-shot examples. By optimizing the KL-divergence among diverse Gaussian embeddings, the model can improve contextual representations, particularly when there are fewer sparse few-shot examples, as in the 5~10 shot scenario. In contrast, COPNER solely relies on the label semantic information to align contextual token representations with label representations. But the alignment is limited as it doesn't consider that the contextual tokens include more alignment information. It shows improvement more prominently in the 1~2 shot scenarios, as is shown in Table 3 and Table 4. It reveals that when the few-shot examples are severely sparse, label information can be proper compensation for the sparsity.

Notably, FEW-NERD (INTER) consistently outperforms FEW-NERD (INTRA), aligning with their respective designs. FEW-NERD (INTER) utilizes coarse-grained entity types shared across train, dev, and test sets, leveraging joint information for enhanced performance. In summary, our method effectively leverages labels to address data sparsity and improves context token representations using more few-shot examples. This demonstrates the effectiveness of our unified framework, combining label-aware prompts and contrastive learning techniques.

6.3 Ablation Study and Analysis

To evaluate the effectiveness of key components in our model, we conducted an ablation study using the following settings: 1) Ours w/ ICL + w/ context-label CL: it's the final loss we use in our model. ICL refers to our improved contrastive loss. 2) Ours w/ ICL + w/o context-label CL: we use our improved contrastive loss (ICL) but remove the context-label

Our Model	CONTaiNER	Ground Truth
skyway committee_{organization-other} - deals with improvements to the st. paul skyway_{location-road/railway/highway/transit}	skyway committee_{organization-company} - deals with improvements to the st. paul skyway_{location-road/railway/highway/transit}	skyway committee_{organization-other} - deals with improvements to the st. paul skyway_{location-road/railway/highway/transit}
the cook islands football association_{organization-other} is the governing body of football in the cook islands_{location-island} .	the cook islands football association_{location-island} is the governing body of football_{organization-other} in the cook islands_{location-island} .	the cook islands football association_{organization-other} is the governing body of football in the cook islands_{location-island} .
the album is a conceptual work based on their past album stream of consciousness_{art-music} .	the album is a conceptual work based on their past album stream of consciousness .	the album is a conceptual work based on their past album stream of consciousness_{art-music} .

Table 6: Case study based on FEW-NERD dataset. We give three randomly sampled demonstrations where organization-other, organization-company, location-road/railway/highway/transit, location-island, art-music are entity types. Here gree color denotes right results while red color denotes mistakes.

contrastive loss. 3) Ours w/ OCL + w/o context-label CL: we use the origin contrastive loss (OCL) from CONTaiNER Das *et al.* [2021] and remove the context-label contrastive loss.

Effect of our improved contrastive loss (ICL) According to Table 5, our improved contrastive loss method outperforms the CONTaiNER method by a significant margin in both the FEW-NERD (INTRA) and FEW-NERD (INTER) datasets. With only minor adjustments to the original contrastive loss, we achieve substantial performance improvements, making it a nearly effortless enhancement.

Effect of our context-label contrastive loss As depicted in Table 5, the utilization of the context-label contrastive loss leads to further improvements in F1 scores. This demonstrates the effectiveness of our approach in leveraging context-label information for enhanced performance.

Case Study Table 6 shows the results of our case study. To assess entity prediction, we randomly sampled three demonstrations from the FEW-NERD dataset (INTRA and INTER). Comparing our model with CONTaiNER, which also utilizes contrastive learning, our method consistently achieves superior predictions for most cases.

We can see that our model performs well in the first case, accurately extracting all entities with correct spans and entity types. In comparison, CONTaiNER correctly predicts the entity span for "skyway committee" but assigns it the wrong entity type of `organization-company`. In the second case, both our model and CONTaiNER make errors in predicting the entity boundary (including an extra word "the"). However, CONTaiNER further assigns the wrong entity type of `location-island` to "the Cook Islands Football Association", while our model predicts the correct entity type. Additionally, CONTaiNER incorrectly extracts the non-existent entity "the governing body of football". Overall, our model's performance is notably superior. In the last case, CONTaiNER fails to make any predictions, whereas our model accurately predicts both the entity span and type.

t-SNE Visualization We used t-SNE Van der Maaten and Hinton [2008] to project BERT's token-level representations onto a 2-dimensional space. Figure 2 illustrates the visualization results on the FEW-NERD dataset. In CONTaiNER's visualization (left), there is considerable overlap among different entity classes. Our model (right) effectively preserves a clear distinction between different entity types, highlighting
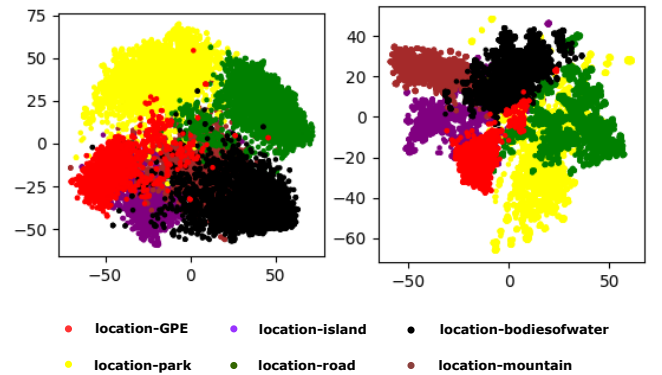

Figure 2: Two-dimensional t-SNE visualizations of the FEW-NERD test set. The token representations are from the sampled 6 fine-grained entity types of `location` category. The left is for CONTaiNER and the right is for ours.

its success in learning a superior metric space specific to each entity type compared to CONTaiNER.

7 Conclusion

We propose a unified label-aware contrastive learning framework tailored for few-shot NER tasks. Our model achieves state-of-the-art performance by leveraging label information and optimizing both context-context and context-label contrastive learning objectives. This approach enhances the discriminative power of contextual representations. Extensive experiments on benchmark datasets consistently demonstrate our model's superiority over previous state-of-the-art methods, even in challenging scenarios. In the future, we plan to extend our approach to other token-level classification tasks, such as POS tagging. Additionally, we plan to explore the potential of our model in addressing zero-shot challenges, where the model needs to generalize to unseen label classes.

References

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. Few-shot text classification with distributional signatures. *arXiv preprint arXiv:1908.06039*, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan,

Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*, pages 2778–2788, 2022.

Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370, 2016.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. Template-based named entity recognition using bart. *arXiv preprint arXiv:2106.01760*, 2021.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J Passonneau, and Rui Zhang. Container: Few-shot named entity recognition via contrastive learning. *arXiv preprint arXiv:2109.07589*, 2021.

Leon Derczynski, Eric Nichols, Marieke Van Erp, and Nut Limsopatham. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Hai-Tao Zheng, and Zhiyuan Liu. Few-nerd: A few-shot named entity recognition dataset. *arXiv preprint arXiv:2105.07464*, 2021.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. Induction networks for few-shot text classification. *arXiv preprint arXiv:1902.10482*, 2019.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*, 2018.

Kai He, Lixia Yao, JiaWei Zhang, Yufei Li, and Chen Li. Construction of genealogical knowledge graphs from obituaries: Multitask neural network extraction system. *Journal of Medical Internet Research*, 23(8):e25670, 2021.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. *arXiv preprint arXiv:2006.05702*, 2020.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. Few-shot named entity recognition: A comprehensive study. *arXiv preprint arXiv:2012.14978*, 2020.

Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. Copner: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International conference on computational linguistics*, pages 2515–2527, 2022.

Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. Good examples make a faster learner: Simple demonstration-based learning for low-resource ner. *arXiv preprint arXiv:2110.08454*, 2021.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuanjing Huang. Template-free prompt tuning for few-shot ner. *arXiv preprint arXiv:2109.13532*, 2021.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. Label semantics for few shot named entity recognition. *arXiv preprint arXiv:2203.08985*, 2022.

Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. Decomposed meta-learning for few-shot named entity recognition. *arXiv preprint arXiv:2204.05751*, 2022.

Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

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

Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens Van Der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23:170, 2013.

Yi Yang and Arzoo Katiyar. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. *arXiv preprint arXiv:2010.02405*, 2020.

Amir Zeldes. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017.

A Additional Implementation Details

Parameter Setting We use `BERT-base-uncased` [2] as our encoder following Ding et al. (2021). For optimization, we utilize AdamW (Loshchilov and Hutter, 2017) as our optimizer with a fixed learning rate of 5e-5 across all experiments. The embedding dimension is set to $l = 128$, and we maintain a batch size of 32 along with a maximum sequence length of 128. To mitigate overfitting, we apply a dropout rate of 0.1. During the source domain training, we train the model for only one epoch in all experiments. This measure is implemented to prevent mode collapse, ensuring that the representations of positive and negative examples do not collapse into a single point. Hyperparameter settings are determined through a grid search within the search space outlined in Table 7. The source domain training process requires approximately 40 minutes on an NVIDIA A6000 GPU.

Learning rate	5e-5
Batch size	32
Max sequence length	128
Embedding dimension	128
Training epoch	1
Optimizer	AdamW
Alpha	{0.8, 0.5, 0.3}

Table 7: Experiment setting and hyper-parameter search space in our paper.

B Fine-tuning Objectives

As depicted in Table 8, in the 1-shot scenario, a single example might not adequately capture the distribution information. Therefore, we opt for using the Euclidean distance as the metric instead of the KL-divergence distance. In our Domain Transfer experiments, we observe that optimizing the context-label contrastive loss alone, rather than both losses, yields improved performance in the 1-shot setting. However, in scenarios involving larger data examples like the 5-shot setting and source domain training, we continue to optimize both contrastive losses using the KL-divergence as the distance metric.

	KL-divergence	Euclidean
1-shot (WNUT)	32.83	35.72
1-shot (GUM)	18.32	23.86

Table 8: F1 scores comparison in Domain Transfer Evaluation of 1-shot (WNUT) and 1-shot (GUM) with different fine-tuning objectives.

[2]https://github.com/google-research/bert

C Label Names For Suffix Prompt

Table 9 shows the natural language forms of all original labels in all the datasets.

Dataset	Class	Natural Language
WNUT'17	person	person
	location	location
	group	group
	product	product
	creative-work	creative work
	corporation	corporation
CoNLL'03	PER	person
	LOC	location
	ORG	organization
	MISC	miscellaneous
OntoNotes	CARDINAL	cardinal
	DATE	date
	EVENT	event
	FAC	facility
	ORG	organization
	LAW	law
	QUANTITY	quantity
	ORDINAL	number
	...	...
I2B2	DATE	date
	PATIENT	patient
	DOCTOR	doctor
	HOSPITAL	hospital
	ZIP	zip
	STATE	state
	MEDICALRECORD	record
	HEALTHPLAN	plan
	...	...
GUM	person	person
	place	place
	organization	organization
	quantity	quantity
	time	time
	event	event
	abstract	abstract
	substance	substance
	object	object
	animal	animal
	plant	plant
FEW-NERD	person-actor	actor
	person-artist/author	artist
	location-island	island
	location-other	location
	organization-education	education
	organization-government/governmentagency	government
	building-airport	airport
	building-other	building
	art-broadcastprogram	broadcastprogram
	other-astronomything	astronomything
	...	...

Table 9: Natural language form of all label names of all data sets