Structure and Label Constrained Data Augmentation for Cross-domain Few-shot NER

Jingyi Zhang[1] , Ying Zhang[1], Yufeng Chen[1]*, Jinan Xu[1]
[1]Beijing Key Lab of Traffic Data Analysis and Mining,
Beijing Jiaotong University, Beijing, China
{jyizhang,zhying,chenyf,jaxu}@bjtu.edu.cn

Abstract

Cross-domain few-shot named entity recognition (NER) is a challenging task that aims to recognize entities in target domains with limited labeled data by leveraging relevant knowledge from source domains. However, domain gaps limit the effect of knowledge transfer and harm the performance of NER models. In this paper, we analyze those domain gaps from two new perspectives, i.e., *entity annotations* and *entity structures* and leverage *word-to-tag* and *word-to-word* relations to model them, respectively. Moreover, we propose a novel method called **S**tructure and **L**abel **C**onstrained **D**ata **A**ugmentation (SLC-DA) for Cross-domain Few-shot NER, which novelly design a label constrained pre-train task and a structure constrained optimization objectives in the data augmentation process to generate domain-specific augmented data to help NER models smoothly transition from source to target domains. We evaluate our approach on several standard datasets and achieve state-of-the-art or competitive results, demonstrating the effectiveness of our method in cross-domain few-shot NER.

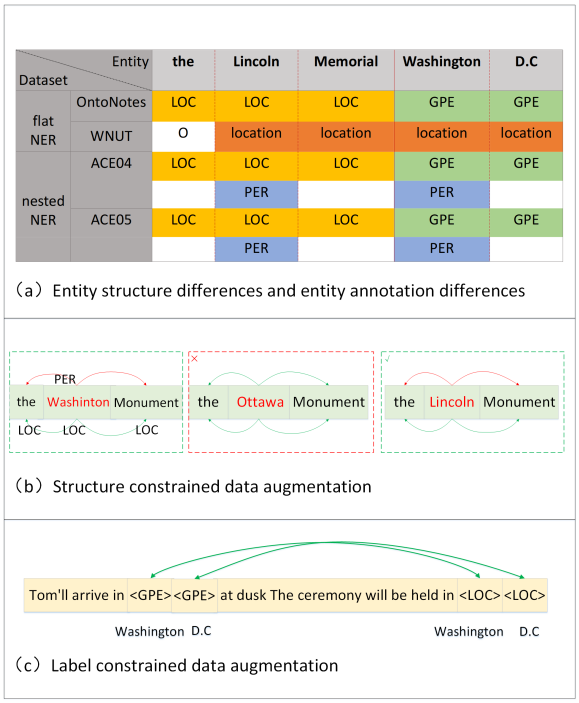

Figure 1: Examples for entities have different structures and labels between the source and target domains. Different colors and combinations of squares represent different entity types and structures, respectively.

1 Introduction

Named entity recognition (NER) is a fundamental Natural Language Processing (NLP) task to detect entity mentions and classify them into predefined labels (Grishman and Sundheim, 1996). Benefiting from powerful feature representations, deep learning based NER approaches (Lample et al., 2016; Devlin et al., 2019; Li et al., 2020) have achieved promising performances. However, their success depends heavily on the large-scale dataset with accurate annotations that are labor-intensive and time-consuming. Although some general domains (e.g., news) provide rich annotations, it is unaffordable to manually annotate NER labels in some new environments (e.g., bio-medicine). Therefore, few-shot NER (Fritzler et al., 2019; Hou et al., 2020) has attracted increasing attentions, aiming to build an NER model with only a small number of supporting samples in specific domains.

Mainstream researches on cross-domain few-shot NER aim to transfer relevant knowledge from source domains. Most of them focus on optimizing model architectures based on metric-learning, such as Prototypical Network based ProtoBERT (Snell et al., 2017; Fritzler et al., 2019; Hou et al., 2020), a nearest neighbor based network NNShot (Yang and Katiyar, 2020), a viterbi decoding variant nearest neighbor based network StructShot (Yang and Katiyar, 2020), and Container (Das et al., 2022), a contrastive model. There are also studies based on data augmentation methods, such as MELM Zhou

*Yufeng Chen is the corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 518–530
December 6-10, 2023

et al. (2022), which uses cross-lingual pre-trained models for data augmentation.

Cross-domain few-shot NER is full of challenges because of the domain gaps, especially for NER tasks. However, existing work lacks research on this problem. As a structure prediction task, NER requires synthetic entity with highly matching of label dependencies. However, there are distinctions among texts from various domains. The same entity mention in various domains is labeled with different entity types or distinct span boundaries. Unfortunately, existing studies have not adequately explored the influence of domain gaps on NER tasks. Consequently, these gaps significantly impact the performance of existing approaches.

In view of these challenges, we divide domain gaps into two categories as shown in Figure 1(a). **Category-I: The structure of entities differs across domains.** For example, 'the Lincoln Memorial' is represented as a contiguous location entity in flat NER datasets, while it is additionally labeled with a person entity "Lincoln" in the nested NER datasets. Additionally, entity boundaries may vary from domain to domain. For example, in different datasets, the inclusion of "the" in the phrase "the Lincoln Memorial" can vary. **Category-II: The annotations of entities differs across domains.** Difference domains usually have different pre-defined entity types. For example, in OntoNotes dataset, 'the Lincoln Memorial' and 'Washington D.C' are annotated as 'LOC' and 'GPE' types, respectively, where as in WNUT dataset, both are classified as 'location'. Our proposed method aims to alleviate the negative impact of structure and annotations differences on cross-domain few-shot NER. By doing so, we aim to enhance the performance of NER models in the target domain.

Based on the above analyses, we introduce two types of relationships to sufficiently model two kinds of domain gaps in cross domain few-shot NER method. **Word-to-word relation**: 'the Washington Monument' is annotated as one entity in flat NER datasets while annotated as two entities in nested ones. When the source domain is a flat entity dataset and the target domain is a nested entity dataset, it is likely to generate non-nested entity data, leading to the NER model not being able to learn the knowledge in the target domain. **Word-to-tag relation**: the corresponding entity types of 'Washington D.C.' are 'GPE' and 'location' in 'OntoNotes' and 'WNUT', respectively, which may cause label conflict if directly learned.

In this paper, we propose a novel method called **S**tructure and **L**abel **C**onstrained **D**ata **A**ugmentation (SLC-DA) for Cross-domain Few-shot NER. SLC-DA novelly design a label constrained pre-train task, which allows the model to learn the mapping relationships between entities across diverse domains. Furthermore, SLC-DA incorporate structure constrained optimization objectives in the data augmentation process to generate domain-aware augmented data to help NER models smoothly transition from source to target domains.

Concretely, for structure-constrained data augmentation as shown in figure 1(b), we calculate the word-to-word relation to model the entity structure between entity word tokens and other tokens, and generate structure-enhanced NER data in the target domain for training. For label-constrained data augmentation as shown in figure 1(c), we replace the same entity mentions with their corresponding different categories for each instance and utilize the language model to learn these word-to-tag relationships in different domains to avoid confusion.

To evaluate the effectiveness of our proposed approach, we compare our approach with previous works on flat NER and report results surpassing the current state-of-the-art. Additionally, we report competitive results on nested NER. Our findings demonstrate that our proposed method is simple yet highly effective. Finally, our main contributions are summarized as follows:

- To bridge domain gaps for cross-domain few-shot NER, we analyze this issue from two new perspectives, i.e., *entity annotations* and *entity structures* and define *word-to-word* and *word-to-tag* relations to model them, respectively.

- We propose a method called Structure and Label Constrained Data Augmentation (SLC-DA), introducing a label-constrained pre-train task and structure-constrained optimization objectives in the data augmentation process.

- We achieved state-of-the-art results in the cross-domain few-shot NER task. We also achieved competitive results by transferring from a flat entity dataset to a nested entity dataset for the first time.

2 Background

2.1 Few-shot NER

Since there are usually multiple entity instances in a single sentence for NER task, which is different from text classification tasks, we cannot following the few-shot settings of text classification to construct the support set by sampling K sentences for each entity class, which can lead to an imbalance in the data. Therefore, for few-shot NER, the dataset consists of N entity classes, with K annotated instances per class, denoted as the N-way K-shot setting. To address the issue of imbalanced data distribution across different entity categories, we employ a specific strategy. Firstly, we calculate the total number of entities for each category. Then, we prioritize the selection of categories with fewer entities. This approach ensures the rationality of the support set size. Statistics of datasets used in experiments can be found in A.4.

2.2 Domain Transfer for Few-shot NER

Domain transfer usually transfers knowledge from the source domain to the target domain.For the few-shot NER task, the NER model is first trained on the source domain and then fine-tuned on the target domain. During the source domain training, the train, development, and test sets are fully utilized. The data splits for train, development, and test are described in Section 4.2. During the target domain training, the training data consist of the generated augmented data (data augmentation based on support set), the development set is the support set, and we adopt the original test set from the target domain for testing.

3 Method

In this section, we present our proposed method, called Structure and Label Constrained Data Augmentation (SLC-DA) for cross-domain few-shot NER. Figure 2 depicts the overview of our method, which includes two modules: structure constrained data augmentation and label constrained pre-train task. We illustrate the details of how we learn entity structure and label relation.

3.1 Structure Constrained Data Augmentation

To enhance the quality of generated NER data, we propose to augment data with structure constrained optimization objective by learning and preserving entity structures.

In the structure constrained data augmentation module, we first combine source domain data and target domain support set to pretrain the data augmentation model. Then, we capture the entity structure by modeling the word-to-word-relation. Subsequently, the structure constrained data augmentation is used to generate more entities that conform to the target domain entity structure and replace the original entities to compose the augmentation data.

Let D_{source} and D_{target} denote the source domain dataset and target domain support set. Given a N tokens sentence $X = [x_1, x_2, ..., x_N]$ with corresponding NER labels $L = [l_1, l_2, ..., l_N]$, we encode the sentence to $H = [h_1, h_2, ..., h_N]$. Then we randomly mask entity tokens to generate a new sequence X' and $M = [m_1, m_2, ..., m_N]$, where $m_i = 1$ if x_i is masked else $m_i = 0$.

For a masked entity token $x_i^{'}$, we use masked language model (MLM) to generate the entity $x_i^{''}$ which is the most similar to the entity x_i, and the new sequence $X^{''}$ is generated by minimizing the loss

$$L_{MLM}(x_i^{''}) = -m_i log p_\theta(x_i|X^{'}), \quad (1)$$

where θ represents model parameters.

Following Qian et al. (2021), we exploit Gaussian embedding, which is innately more expressive than point embedding. For token embedding h_i, we compute its Gaussian Embedding $G_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ as follows:

$$\mu_i = ReLU(h_i), \quad (2)$$

$$\Sigma_i = ELU(ReLU(h_i)) + (1 + \epsilon), \quad (3)$$

where μ_i denotes the semantics of x_i, covariance matrix Σ_i represents the uncertainty, ELU represents exponential linear unit, and ϵ is set toe^{-14} for numerical stability. We use KL-divergence to calculate the similarity of entity structure

$$D_{KL}(i,j) = \frac{1}{2}(log\frac{|\Sigma_i|}{|\Sigma_j|} - D + tr(\Sigma_i^{-1}\Sigma_j) + (\mu_i - \mu_j)^T\Sigma_\mu^{-1}(\mu_i - \mu_j). \quad (4)$$

Since the KL-divergence is asymmetric, we obtain the similarity by calculating the KL-divergence in both directions

$$d(i,j) = \frac{1}{2}(D_{KL}(i,j) + D_{KL}(j,i)). \quad (5)$$

We define the optimization objective as follows to minimize the structure difference between newly

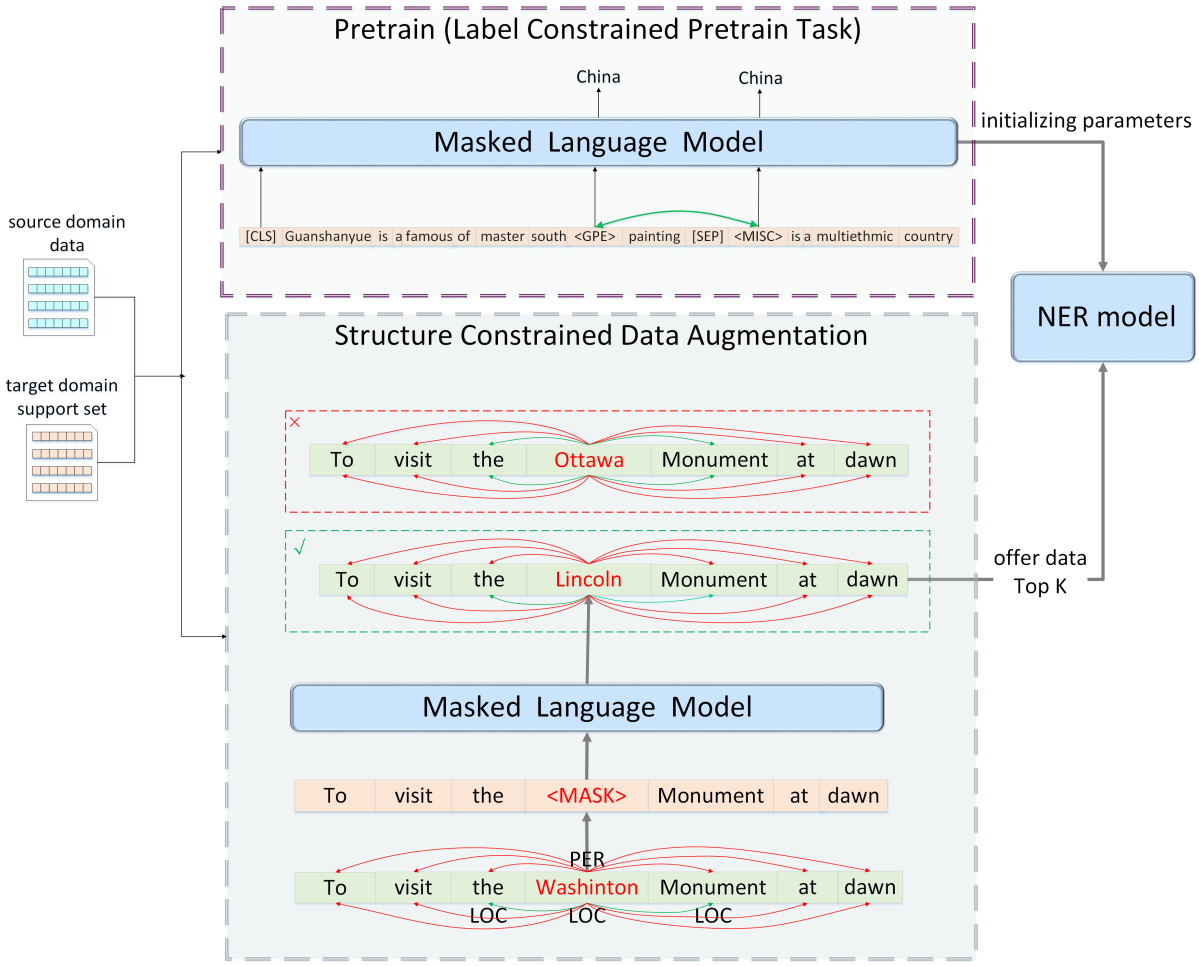

Figure 2: An overview of our proposed approach, which comprises structure constrained data augmentation (bottom) and label constrained pre-train task (top). First, we introduce the word-to-tag relation for label constrained pre-train task. Then, we compute the word-to-word relation for structure constrained data augmentation to generate augmented data for the target domain support set. Among the sentences generated for each sentence, we'll pick the top-K sentences that meet our satisfaction criteria, which we'll denote as $\surd$. Finally, we merge the generated data with the source domain data to train the NER model.

generated entities and original ones.

$$L_{structure}(x_i) = log\frac{\sum_{j=1,j\neq i}^{n}|(d(x_i,x_j)-d(x_i^{'},x_j)|}{n-1}. \quad (6)$$

To sum up, for sequence X, the total loss is formulated as

$$L_{total} = L_{MLM} + L_{structure}. \quad (7)$$

In this way, we generate entities that conform to the target domain entity structure. So we can combine the generated data with the source domain data to train the NER model.

3.2 Label Constrained Pre-train Task

In the label constrained pre-training module, to alleviate label inconsistency between different domains, we design several label constraint strategies to align predefined labels between source and target domains. First, we extract all entities and corresponding sentences from the support set, and find sentences containing these entities in the source domain data set. Then, we form a sentence pair containing the same entity with the sentences in the support set. Let the pretrain language model (PLM) learn different labels of the same entity in different domains and learn the relationship between these two labels. This PLM is subsequently trained with the train data and structure constrained of the source domain to become a NER model. Finally, when inference, we utilize a pre-trained label-constrained model to compute the mapping relationships between labels in the source and target domains. This allows us to bridge the gap between the two domains. We also predefine the label

mapping to bridge the source and target domain.

To alleviate the label inconsistency among different domains, we propose a novel label-constrained pre-training task to align the inconsistent predefined labels between the source and the target domains in training and prediction process. Based on pre-trained contextual representations, we design a label mapping strategy by calculating the similarity of various predefined labels to align inconsistent labels between the source and target domains in training and prediction process.

In the training process, we first filter out all the entities E and their labels L from the sentences S in the target domain denoted as $[e_1, e_2, ..., e_N]$, $[l_1, l_2, ..., l_N]$, and $[s_1, s_2, ..., s_N]$, respectively. s_i is a sequence of m tokens $[x_1, x_2, ..., x_m]$. We then select sentences containing these entities and labels from the source domain data as $[s_1^{'}, s_2^{'}, ..., s_p^{'}]$ and $[l_1^{'}, l_2^{'}, ..., l_p{}^{'}]$ and match all sentences which have the same entity up as $[e_1; s_1; s_1^{'}, e_2; s_2; s_2^{'}, ..., e_n; s_n; s_p^{'}]$. Then we swap entities in these sentences with their corresponding labels and generate the representations of two labels l_i and $l_i^{'}$ and compute the KL-Distribution between them as:

$$L(i) = -log\frac{exp(-d(l_i, l_i'))}{\sum_{j=1}^{n} exp(-d(l_i, l_j)} \quad (8)$$

and achieve label alignment by learning the relationship between the labels of these entities in the source and the target domains. Finally, we apply the saved parameters to initialize the NER model.

In the prediction process, we propose a simple but efficient post processing method to align labels from different domains. Since the NER model is pre-trained on the source domain, it will be affected by predefined labels of the source domain when identifying entities, making some predictions of predefined labels not part of the target domain. Therefore, during inference, we post-process results predicted by the model.

We employ label-constrained pre-training task to obtain contextual representations for different entity labels and then compute the mapping relationships of entity categories between the source and target domains (including the "other" category). Specifically, according to the official annotation guidelines for each dataset, we generate descriptive statements for each entity category and calculate KL divergence based on the representations of description sentences for each entity category between the source and target domains. This process allows us to derive the mapping relationships between entity categories in the source domain and entity categories in the target domain.

4 Experiments

We validate our proposed method in the flat entity and nested entity settings. The details of the experiments are elaborated in this section. We use precision (P), recall (R), and F (F1) as evaluation metrics. All experimental results are the average score over five runs with random seeds.

4.1 Datasets

We validate our proposed method on various domain datasets. We use OntoNotes 5.0 (General) (Pradhan et al., 2013) standard training set as our source domain training data, and use CoNLL 2003 (News) (Sang and Meulder, 2003), Wnut 17 (Social) (Derczynski et al., 2017), I2B2(Medical) (Stubbs and Uzuner, 2015), GUM(Mixed) (Zeldes, 2017) as our flat NER setting target domain, ACE2004 (Event) (Doddington et al., 2004), ACE2005 (Event) (Walker and Consortium, 2005) as our nested NER target domain. For the source domain, we use the OntoNotes train/development/test splits released for the CoNLL 2012 shared task. For the target domains, we consider all datasets except for OntoNotes, and then extract the support set as mentioned in Section 2.2. The statistics of datasets used in experiments can be found in A.4.

4.2 Baselines

We compare the performances of SLC-DA on different datasets in the flat and nested entity settings with the following cross-domain Few-Shot NER models. 1) **Direct-Transfer**: which trains the NER model on the source domain data and evaluates it on the target domain support set. 2) **MELM** (Zhou et al., 2022): our baseline method, which exploits the Masked Entity Language Model (MELM) to generate the augmented NER data. 3) **ProtoBERT** (Snell et al., 2017; Fritzler et al., 2019; Hou et al., 2020): an implementation of Prototypical Network based on BERT. 4) **NNShot** and **Structshot** (Yang and Katiyar, 2020):a nearest neighbor based network and a viterbi decoding variant nearest neighbor based network. 5) **CONTaiNER** (Das et al., 2022): a model based on contrastive learning to learn the relationship between

	Flat entity settings									*Nested entity settings*			
	CoNLL		**WNUT**		**I2B2**		**GUM**		**Avg**	**ACE04**		**ACE05**	
Models	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot		1-shot	5-shot	1-shot	5-shot
Direct Transfer	15.5	15.5	0	0	0	0	0	0	3.9	34.1	34.1	25.2	25.2
MELM‡	17.8	20.8	1.0	9.1	0	4.4	0.4	1.1	6.8	34.1	31.8	24.2	24.2
Proto†	49.9	61.3	17.4	22.8	13.4	17.9	17.8	19.5	27.5	-	-	-	-
NNShot†	61.2	74.1	22.7	27.3	15.3	22.0	10.5	15.9	31.1	-	-	-	-
StructShot†	62.4	74.8	24.2	30.4	21.4	30.3	7.8	13.3	33.1	-	-	-	-
CONTaiNER†	61.2	75.8	27.5	32.5	21.5	36.7	18.5	25.2	37.4	-	-	-	-
ProML†	69.2	79.1	43.9	**53.4**	25.0	**58.2**	15.3	37.0	47.6	-	-	-	-
SLC-DA (ours)	**79.2**	**81.7**	**44.1**	46.1	**35.3**	49.1	**19.2**	**41.3**	**49.5**	**39.9**	**41.2**	**32.6**	**40.2**

Table 1: Main results of SLC-DA and comparison methods for cross-domain few-shot NER. We use F (F1) as evaluation metrics. '†' represents the results are cited from the initial paper, and '‡' represents the results re-implemented by us. All experimental results are the average score over 5 runs with random seeds.

entities of different categories. 6) **ProML** (Chen et al., 2022): designed multiple prompt schemas are to enhance label semantics.

4.3 Main Results

Table 1 presents the results of flat and nested NER. Compared with these strong baselines, SLC-DA leads to significant improvements and achieves state-of-the-art performances in flat NER setting.

We also report competitive results in the newly proposed nested-entity cross-domain setting. Particularly, in the flat NER setting, our method improves 1.9% on average compared to SOTA, and improves 12.1% on average compared to CONTaiNER. In the nested NER setting, our method improves 6.45% and 11.2% on average compared to SOTA, 7.6% and 12.2% compared to baseline for ACE04 and ACE05, respectively. All these results well demonstrate the effective of our method. The reason we did not report all baselines for the ACE04/ACE05 datasets is because these datasets contain nested named entities, which pose a challenge for traditional baseline methods designed for flat entities. These baseline methods, such as container-based approaches, are not suitable for accurately handling nested entities. Therefore, including their results in the evaluation would not provide a fair comparison.

The experimental results demonstrate that our method and the metric-based method can achieve good performance when there is small domain difference between the source and target domains (where the target domain is the CoNLL dataset in the news domain). However, our method has significant advantages over other methods when there is a large domain difference and limited target domain data, as demonstrated by the experiments on I2B2 (medical) and GUM (mixed) datasets. This indicates that our data augmentation method can help the NER model smoothly transfer from the source domain to the target domain. In addition, due to the different usage of the target domain support set, our method only uses the generated augmented data and source domain data in the training set, while other metric-based comparison methods use the support set as the training set. As k-shot increases, our method performs slightly worse than ProML method on WNUT and I2B2 datasets but better than other methods.

MELM in the nested NER setting is even less effective than direct migration. The main reason is that the labels of ACE04 and ACE05 are identical to the labels of OntoNotes in five cases, but ACE04 and ACE05 have nested entities, resulting in the failure of data augmentation. By contrast, our SLC-DA can learn the entity structure of target domain and thus generate appropriate augmented data, benefiting the NER model to learn the knowledge effectively and further improve the ability of recognizing entity in the target domain. Parameter settings can be found in A.1.

In our study, the migration from OntoNotes to CoNLL dataset cannot strictly be considered as a cross-domain setup, since the OntoNotes dataset includes news data as one of its sources. However, to maintain consistency with existing literature and experimental settings, we conducted experiments under this particular setup.

5 Analysis and Discussion

5.1 Ablation Study

We conduct ablation studies to explore the effect of structure-constrained and label-constrained mod-

Flat entity settings												
Models	**CoNLL**						**WNUT**					
	1-shot			5-shot			1-shot			5-shot		
	P	R	F	P	R	F	P	R	F	P	R	F
Direct Transfer	39.5	11.8	15.5	39.5	11.8	15.5	0	0	0	0	0	0
SLC-DA	82.9	**76.4**	**79.2**	**85.0**	**78.6**	**81.7**	**70.5**	**32.1**	**44.1**	**68.7**	**34.6**	**46.1**
- w/o structure-constrained	65.9	52.3	58.5	68.3	54.3	60.5	36.5	23.6	28.7	35.5	23.9	28.6
- w/o label-constrained	**83.5**	75.0	79.0	81.3	77.6	79.4	55.1	29.8	38.7	51.6	30.8	38.5
Nested entity settings												
Models	**ACE04**						**ACE05**					
	1-shot			5-shot			1-shot			5-shot		
	P	R	F	P	R	F	P	R	F	P	R	F
Direct Transfer	69.8	22.5	34.1	**69.8**	22.5	34.1	**63.1**	15.77	25.2	**63.1**	15.77	25.2
SLC-DA	**72.2**	**27.6**	**39.9**	67.0	**29.7**	**41.2**	55.6	**23.1**	**32.6**	61.0	**30.0**	**40.2**
- w/o structure-constrained	70.6	21.7	33.2	66.5	23.4	34.6	60.8	9.3	16.1	60.8	13.6	22.2
- w/o label-constrained	68.2	25.5	37.2	68.3	28.1	39.8	56.9	19.8	29.4	58.6	28.5	38.3

Table 2: Ablation study on SLC-DA method for cross domain few-shot NER.

ules on the overall performance. The results of flat NER and nested NER are reported in Table 2.

w/o structure-constrained: Structure constrained is not used in data augmentation, and entities predicted by language model are directly used as newly generated data. When ablating the structure-constrained module, Table 2 shows that the performances of SLC-DA drop dramatically for both flat and nested NER in the 1-shot and 5-shot settings. Particularly, for flat NER in both 1-shot and 5-shot settings, when removing the 'structure-constrained' module, the F1-scores drop by over 20% and 10% on the CoNLL and WNUT datasets, respectively. For nested NER, the F1-scores drop by over 6% and 15% on ACE04 and ACE05 datasets, respectively. Overall, these results prove that our structure-constrained data augmentation module plays an important role in SLC-DA and it is necessary to exploiting entity structure information in data augmentation methods.

w/o label-constrained: The NER model are directly initialized by pretrained *bert-base-cased* and no longer learns the relationship between entity labels from the source and target domain. When ablating the label-constrained module, Table 2 illustrates that the performances of SLC-DA drop slightly for both flat and nested NER in the 1-shot and 5-shot settings. Concretely, for flat NER in both 1-shot and 5-shot settings, when removing the 'label-constrained' module, the F1-scores drop by approximately 1.5% and 1% on the CoNLL and WNUT datasets, respectively. For nested NER, the F1-scores drop by over 2% and 2.5% on ACE04 and ACE05 datasets, separately. Subsequently, these analyses demonstrate that the label-constrained data augmentation module also have a consistent effect on the performance of SLC-DA and modeling relations among different labels of the same entities contributes to data augmentation.

In summary, both 'structure-constrained' and 'label-constrained' module have important effects on performances of our proposed method. However, compared with the operation 'w/o structure-constrained', removing the 'label constrained' module from SLC-DA results in a more marginal decrease of performances on both flat and nested NER, illustrating that 'structure-constrained' module is more influential than 'label-constrained' in the SLC-DA method. We conjecture that there are two possible reasons: I) 'structure-constrained' module directly participates in the process of augmented data generation while 'label-constrained' does not, since 'structure-constrained' is one of the optimization objective of data generation process while 'label-constrained' is only used as the initialization of parameters. II) the scale of training data for 'structure-constrained' module is larger than that for 'label-constrained' module, leading to the differences of model's ability on capturing entity structure and label-relation information.

5.2 Results on Different Labels

Table 3 shows the results of our SLC-DA model on different label entities in CoNLL, WNUT, ACE04

Dataset	Entity Type	MELM			Structure			Label			SLC-DA		
		P	R	F	P	R	F	P	R	F	P	R	F
CoNLL	PER	**100.0**	0.6	1.2	87.9	84.1	86.0	82.4	57.2	67.6	90.5	**87.2**	**88.8**
	ORG	72.4	78.4	75.2	68.7	79.4	73.6	59.3	40.9	48.4	**76.3**	**79.5**	**77.8**
	LOC	70.4	3.7	7.1	**90.1**	74.2	81.3	69.7	62.5	65.9	88.8	**75.2**	**81.4**
	MISC	**100.0**	0.6	1.2	84.9	**66.3**	74.5	57.8	59.5	58.7	87.0	65.1	**74.5**
WNUT	corporation	13.2	**10.6**	**11.7**	10.5	3.0	4.7	11.1	2.4	4.0	**40.0**	6.1	10.5
	creative-work	20.0	4.2	7.0	**39.7**	20.4	27.0	28.1	16.2	20.5	38.3	**25.4**	**30.5**
	group	**41.7**	3.0	5.6	21.7	18.2	19.8	11.4	**36.3**	17.3	22.0	20.0	**21.0**
	location	50.0	3.3	6.2	60.2	35.3	44.5	45.8	32.7	38.1	**62.8**	**36.0**	**45.8**
	person	47.7	7.4	12.9	71.6	47.6	57.1	59.8	35.0	44.1	**78.8**	**47.6**	**59.3**
	product	23.8	3.9	6.7	**40.7**	8.7	14.3	20.0	6.3	9.6	34.2	**11.0**	**16.7**
ACE04	GPE	**79.0**	29.9	43.4	73.2	46.0	56.5	75.1	39.4	51.6	73.6	**47.7**	**57.9**
	ORG	69.6	29.9	41.8	67.1	30.6	42.0	**70.2**	29.0	41.0	67.7	**34.2**	**45.5**
	PER	64.5	15.3	24.7	66.4	19.9	30.6	61.8	15.9	25.3	**64.7**	**22.8**	**33.7**
	FAC	**66.7**	5.4	9.9	50.0	6.3	11.1	60.0	2.7	5.1	45.0	**8.0**	**13.6**
	VEH	60.8	9.3	16.1	56.9	19.8	29.4	**62.8**	15.0	24.2	55.6	**23.1**	**32.6**
	LOC	**43.5**	9.5	15.6	43.2	15.2	22.5	36.8	13.3	19.6	40.9	**17.1**	**24.2**
	WEA	0	0	0	68.2	25.5	37.2	70.6	21.7	33.2	**72.2**	**27.6**	**39.9**
ACE05	GPE	**76.2**	31.7	44.8	74.5	40.9	52.8	73.4	31.0	43.6	69.5	**43.3**	**53.5**
	ORG	65.4	21.8	32.7	63.3	28.3	39.1	**71.4**	25.5	37.6	61.6	**28.7**	**39.1**
	PER	52.4	8.9	15.2	54.8	29.4	38.3	53.8	9.5	16.1	**59.3**	**30.8**	**40.5**
	FAC	33.3	0.7	1.4	52.4	8.1	14.0	50.0	4.4	8.1	**73.9**	**12.5**	**21.4**
	VEH	15.8	3.0	5.0	20.0	2.9	5.1	**66.7**	3.9	7.4	42.9	**5.9**	**10.3**
	LOC	30.8	7.4	11.9	52.2	22.2	31.2	29.4	9.3	14.1	**66.7**	**29.6**	**41.0**
	WEA	0	0	0	13.3	4.0	6.2	9.1	2.0	3.3	**25.0**	**4.0**	**6.9**

Table 3: Performance of SLC-DA and comparison methods on each entity type of all datasets in the 5-shot settings. 'Structure' denotes only structure constrained is used in data augmentation, and 'Label' denotes only label constrained is used in data augmentation.

and ACE05 in the 5-shot settings. Our method achieves the best F1-score (87.6%) on almost all labels and an extremely highest recall (86.6%) compared with MELM. The results well demonstrate the effective of our method.

Upon observing the results of the ablation experiments, it can be seen that the MELM model only performed well on the 'ORG' category in the CoNLL dataset, which is because this category is included in the source domain OntoNotes. For the other categories, the MELM model showed high precision but low recall and F1 scores, indicating that the model could not identify most entities belonging to the target domain label. On the other hand, our SLC-DA model achieved better results in all categories except for 'corporation' in the WNUT dataset, with an increase in recall proving that our method helped the NER model learn to identify entities belonging to the target domain, and thus, demonstrated that our approach can help NER models more smoothly and effectively transfer from the source to the target domain.

In term of the abnormal result of label 'corporation', we conjecture it is because 'corporation' and 'group' are overlapped by label 'ORG' in OntoNotes. Since there are more entities with 'group' label than 'corporation' label, the model better learns the mapping relationship between 'group' and 'ORG'. Consequently, some data that should be labeled as 'corporation' is labeled as 'group'. These results in the precision decrease in label 'group' and the recall drop in label 'corporation' by comparison.

5.3 Case Study

In Table 4, we present some cases by comparing words generated by SLC-DA and MELM to verify the effectiveness of our method. It can be seen that our method can generate appropriate entities according to the entity structure when encountering unseen entities in the source domain. In addition, when the target domain contains more difficult

Text	I switched the channel last night to [Independence Day]$_{creative\_work}$.
MELM	American Football, Between Football,Hard Trek
SLC-DA	[Seven Days]$_{creative\_work}$, [Saturday Night]$_{creative\_work}$ [The Trek]$_{creative\_work}$,
Text	...that were imposed in 1998 to punish and isolate the regime of [[Yugoslav]$_{GPE}$ President Slobodan Milosevic]$_{PER}$.
MELM	[Yugoslavia President Slobodan Milosevic]$_{PER}$ [communist President Slobodan Milosevic]$_{PER}$ [former President Slobodan Milosevic]$_{PER}$
SLC-DA	[[Serbian]$_{GPE}$ President Slobodan Milosevic]$_{PER}$ [[Montenegro]$_{GPE}$ President Slobodan Milosevic]$_{PER}$ [[Kosovo]$_{GPE}$ President Slobodan Milosevic]$_{PER}$
Text	What they say are [the [Bush]$_{PER}$ administration]$_{ORG}$ ' s neglectful attitudes about their problems
MELM	United [Houston]$_{GPE}$ Workers [Social United Administration]$_{ORG}$ [The United House administration]$_{ORG}$
SLC-DA	[the [Reagan]$_{PER}$ administration]$_{ORG}$ [the [Carter]$_{PER}$ administration]$_{ORG}$ [the [Clinton]$_{PER}$ administration]$_{ORG}$

Table 4: Case study of augmented data, blue represents all entities and red represents nested entities.

nested entities, our method can also generate appropriate entities. And the results on different labels can be found in 5.2.

6 Related Work

6.1 Cross-domain Few-shot NER

Recently, few-shot NER tasks have attracted a lot of researches. For example, Snell et al. (2017); Fritzler et al. (2019); Hou et al. (2020)applied prototype networks to the few-shot NER task, while Yang and Katiyar (2020) proposed StructShot, a model that learns class-specific features and extends intermediate representations to new domains. Das et al. (2022) proposed Container, a model based on contrastive learning to learn the relationship between entities of different categories for better understanding of new unseen classes. Liu et al. (2022) devises two prompting mechanisms for better training data generation, which is heavily influenced by the prompt strategies and needs heavy computation. However, these methods suffer from serious label inconsistency issues in cross-domain scenarios. The knowledge they learned in the source domain cannot be directly applied to the target domain. Moreover, due to the complexity of NER tasks, these works need to design complex learning strategies to be applied to few-shot NER tasks. Unlike previous works, our work focuses on the knowledge of entity internal structure. As far as we know, we are the first to introduce entity structure and label information in data augmentation to solve cross-domain few-shot NER tasks, which may inspire the following exploration to study internal structure knowledge of entities.

6.2 Data Augmentation

Data augmentation is a popular solution for few-shot learning tasks, which is also studied in cross-domain few-shot NER. However, noise is inevitably introduced in the process of introducing augmented data. As a token-level task, NER is vulnerable to the noise caused by augmented data. Dai and Adel (2020) uses label-wise token replacement, synonym replacement, and mention replacement to augmented data but does not increase the diversity of entities. Ding et al. (2020) and Zhou et al. (2022) respectively train language model and mask language model to fuse the alignment information of entities and labels to constrain the newly generated words to match labels, but inevitably introduce a lot of noise. While current methods aim to align the newly generated entities with the original labels, they face limitations in cross-domain few-shot NER. This is because each domain may have distinct entity structures and labels, rendering the generated data incompatible with the target domain's entity structure. Consequently, conventional data augmentation methods are not directly applicable to cross-domain few-shot NER.

7 Conclusion

In this paper, we propose a **S**tructure and **L**abel **C**onstrained **D**ata **A**ugmentation (SLC-DA) for cross-domain few-shot NER, by introducing entity structure and label information from various domains in the data augmentation process, to obtain high-quality synthetic data in the target domain. Experimental results on both flat and nested few-shot NER tasks show that our method can significantly improve the quality of generated data and help NER model find more target domain entities.

Limitations

Named entities are typically classified as flat, nested, or discontinuous entities, with significant structural differences between the three types. This makes it challenging for existing methods to effectively transfer from flat NER datasets to either nested or discontinuous NER datasets. While our experiments in this paper validate that our method can effectively transfer NER models from flat to nested datasets, we have yet to demonstrate its efficacy in transferring to discontinuous NER datasets. This is an avenue for future research, as our method's effectiveness on discontinuous datasets remains to be explored.

Acknowledgments

The research work descried in this paper has been supported by the National Nature Science Foundation of China (No. 61976016, 61976015, and 61876198) and the National Key R&D Program of China (2020AAA0108001). The authors also would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Yanru Chen, Yanan Zheng, and Zhilin Yang. 2022. Prompt-based metric learning for few-shot ner.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3861–3867. International Committee on Computational Linguistics.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 140–147. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2020. DAGA: data augmentation with a generation approach for low-resource tagging tasks. *CoRR*, abs/2011.01549.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC 2019, Limassol, Cyprus, April 8-12, 2019*, pages 993–1000. ACM.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference- 6: A brief history. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*, pages 466–471.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1381–1393. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Jian Liu, Yufeng Chen, and Jinan Xu. 2022. Low-resource NER by data augmentation with prompting.

In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4252–4258. ijcai.org.

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

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 143–152. ACL.

Chen Qian, Fuli Feng, Lijie Wen, and Tat-Seng Chua. 2021. Conceptualized and contextualized gaussian embedding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13683–13691. AAAI Press.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

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

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *J. Biomed. Informatics*, 58:S20–S29.

C. Walker and Linguistic Data Consortium. 2005. *ACE 2005 Multilingual Training Corpus*. LDC corpora. Linguistic Data Consortium.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6365–6375. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Lang. Resour. Evaluation*, 51(3):581–612.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. MELM: data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2251–2262. Association for Computational Linguistics.

A Example Appendix

A.1 Parameter Settings

We elaborate experimental settings of SLC-DA and NER models.

SLC-DA: We use bert-base-cased (Devlin et al., 2019) with a language model head for our structure constrained data augmentation model. The model is trained for 20 epochs on source domain training data and target domain support set, using Adam optimizer with batch size set to 16 and learning rate set to 1e-5.

We calculate the loss for each word generated by the masked language model and select the best top-K. For a sentence containing n entities, this results in generating K^n new sentences (with k=5 in the presented experimental results). The additional time required by our method compared to MELM is not substantial because of the few-shot setting of the target domain's support set. Let's take the example of the WNUT 2017 dataset with a 5-shot setting. The support set consists of 216 sentences and 334 entity words. MELM requires approximately 4 minutes for data augmentation, while our method takes around 13 minutes.

NER Model: For *flat NER setting*, we use bert-base-cased (Devlin et al., 2019) with CRF (Lample et al., 2016) head as the NER model. The model is trained for 5 epochs on source domain training data and target domain support set, using Adam optimizer (Loshchilov and Hutter, 2019) with batch size set to 32 and learning rate set to 5e-5. For *nested NER setting*, we use the MRC-NER (Li et al., 2020) model as the NER model. The model is trained for 7 epochs on source domain training data and finetuned for 20 epochs on the SLC-DA generated data, using Adam optimizer with batch size set to 16 and learning rate set to 2e-5. For *Direct Transfer setting*, the model is trained only on the source domain training data.

The server used for running our program is NVIDIA Tesla P100-SXM2. The average training time per epoch for our structure constrained data augmentation component is approximately 0.5

Models	CoNLL	WNUT	I2B2	GUM
Direct Transfer	15.5	15.5	0	0
MELM‡	17.5	11.1	5.4	1.2
CONTaiNER‡	77.6	35.4	41.0	28.1
ProML‡	81.5	**55.0**	**61.2**	39.6
SLC-DA (ours)	**83.2**	50.4	56.1	**59.5**

Table 5: Overall performances of all systems on four datasets in 10-shot settings for few-shot NER. We use F1 scores as evaluation metrics. '‡' represents the results re-implemented by us. Each experimental result is the average of performance over 5 runs with random seeds.

Models	CoNLL		WNUT		I2B2		GUM	
	50shot	full	50shot	full	50shot	full	50shot	full
Direct Transfer	15.5	15.5	0	0	0	0	0	0
MELM‡	22.8	91.2	13.0	45.6	9.6	97.9	1.4	31.7
SLC-DA (ours)	**85.1**	**91.7**	**56.2**	**47.5**	**61.9**	**98.0**	**73.8**	**60.3**

Table 6: Performances of systems in 50-shot and full set settings for few-shot NER.

hours. For our label constrained data augmentation component, the training time per epoch is influenced by the difference in size between the source and target domains, ranging from 0.5 to 3 hours. For training the NER model, taking the OntoNotes5.0 dataset of approximately 70k samples as an example, the training time is around 1 hour.

A.2 Experiments on a few more samples

Dataset	Support Data		Augmented Data	
	10-shot	50-shot	10-shot	50-shot
CoNLL	401	1422	65,870	123,376
WNUT	423	697	53,622	75,561
I2B2	894	1793	134,944	196,003
GUM	866	2107	239,766	590,523

Table 7: Statistics of support and augmented data used in our experiments.

We provide experimental results in the 10-shot, 50-shot and full set settings, as shown in Table 5 and 6. The statistics of the support sets and augmented data used in our experiments is shown in Table 7.

In the 10-shot settings, our SLC-DA method still outperforms all other comparisons on CoNLL and GUM datasets. Although ProML method performs better than us on WNUT and I2B2 datasets, our approach based on data augmentation explore a various direction and can be combined with ProML to achieve more improvements.

Dataset	Entity Type	Dataset	Entity Type
OntoNotes	NORP	I2B2	IDNUM
	ORG		MEDICALRECORD
	PERSON		PHONE
	DATE		ZIP
	GPE		age
	FAC		city
	CARDINAL		country
	TIME		date
	ORDINAL		device
	EVENT		email
	QUANTITY		fax
	PERCENT		hospital
	LOC		organization
	WORK_OF_ART		person
	MONEY		profession
	LAW		state
	PRODUCT		street
	LANGUAGE		username
CoNLL	PER	WNUT	corporation
	ORG		creative-work
	LOC		group
	MISC		location
			person
			product
GUM	abstract	ACE04 /ACE05	GPE
	animal		ORG
	event		PER
	object		FAC
	organization		VEH
	person		LOC
	place		WEA
	plant		
	quantity		
	substance		
	time		

Table 8: Statistics of predefined labels in all datasets from both source and target domains.

A.3 Statistics of Predefined Labels

Table 8 displays all predefined labels from various datasets in both source and target domains. Concretely, 'OntoNotes' dataset from the source domain contains 18 predefined entity types. For the target domains, there are 18, 11, 7, 6, and 4 entity categories in 'I2B2', 'GUM', 'ACE04/ACE05', 'WNUT' and 'CoNLL' dataset, respectively. Note that we collect their corresponding label descriptions from the annotation guidelines.

A.4 Statistics of datasets used in experiments

The datasets utilized in our study are open-source and consist of various entity types. Further information on the datasets can be found in Table 9. The table lists the entity types present in the datasets, along with the percentage of nested entities relative to the total number of entities.

We report the size of the augment data generated by the support set for each data set in the table 10.

Dataset	Domian	Class	Nested Entity/Enity	1-shot entity	5-shot entity	1-shot sentence	5-shot sentence
OntoNotes	General	18	-	-	-	-	-
CoNLL	News	4	-	79	259	25	87
WNUT	Social	6	-	82	334	51	216
I2B2	Medical	18	-	238	1133	162	706
GUM	Mixed	11	-	170	868	81	276
ACE04	Event	7	12k/27k	202	352	61	102
ACE05	Event	7	12k/30k	162	407	43	115

Table 9: Statistics of support set used in experiments

Dataset	1-shot sentence	5-shot sentence
OntoNotes	-	-
CoNLL	8618	48809
WNUT	3734	14770
I2B2	5546	49066
GUM	8303	130857
ACE04	18198	48855
ACE05	54371	92278

Table 10: Statistics of augment data used in experiments

Dataset	CoNLL	WNUT	I2B2	GUM
zero-shot	68.4	33.3	15.6	16.3
5-shot	76.2	40.0	9.5	13.2
10-shot	68.4	40.0	8.6	10.5

Table 11: Results of GPT3.5 in different settings.

A.5 Result of GPT3.5

As shown in Table 11, we provide the result of GPT3.5 on zero-shot, 5-shot and 10-shot setting. When constructing the input prompt using the support set, we used the following template: First, input a prompt: "As a good linguist, you are asked to identify and label the named entities in the given sentences. There are some examples, please remember them:" Secondly, input the support-generated prompt one by one, for example: "Sentence is: John lives in New York City. Entities are: <John, PER> <New York City, LOC>."

We observed that the performance of GPT 3.5 varies on different datasets, performing well on common entities but worse on scarce entities(e.g., medical dataset I2B2). Additionally, compared results of '5-shot' and '10-shot', we find that GPT 3.5 does not perform better with the increase of support samples due to the match degree between inputs and prompts, which highlights the importance of selecting proper prompts for the application of LLMs to few-shot tasks.

Compared to GPT 3.5, our method achieves better performance in the few-shot NER scenario. Furthermore, our core idea can be easily applied to LLMs by modeling structure and label relationships with appropriate prompts, which can help select input data and improve their performance.