Unsupervised Domain Adaptation with Contrastive Learning for Cross-domain Chinese NER

Anonymous ACL submission

Abstract

Understanding and recognizing the entities of Chinese articles highly relies on the fully supervised learning based on the domain-specific annotation corpus. However, this paradigm fails to generalize over other unlabeled domain data which consists of different entities semantics and domain knowledge. To address this domain shift issue, we propose the framework of unsupervised **D**omain **A**daptation with **C**ontrastive learning for Chinese **NER** (***DAC-NER***). We follow Domain Separation Network (DSN) framework to leverage private-share pattern to capture domain-specific and domain-invariant knowledge. Specifically, we enhance the Chinese word by injecting external lexical knowledge base into the context-aware word embeddings, and then combine with sentence-level semantics to represent the domain knowledge. To learn the domain-invariant knowledge, we replace the conventional adversarial method with novel contrastive regularization to further improve the generalization abilities. Extensive experiments conducted over the labeled source domain MSRA and the unlabeled target domain Social Media and News show that our approach outperforms state-of-the-arts, and achieves the improvement of F1 score by 8.7% over the baseline[1].

1 Introduction

Chinese Named Entity Recognition (NER) is one of the most challenging tasks of Natural Language Processing (NLP), which involves the detection and category of named entities (Gui et al., 2019; Xi et al., 2021; Zhao et al., 2021). In real-word scenarios, Chinese NER task has abundant annotations in formal domain, which supports for the supervised learning. However, it may perform poorly when applied to a new domain without any labeled data, which can be regarded as a data shift problem (Kim et al., 2017; Peng and Dredze, 2017).

[1]All the datasets are publicly available. Source codes are provided in attachments and will be released upon acceptance.

A straightforward method is to pre-train the large-scale labeled data from the source domain, and then fine-tune it with the target domain data (Huang et al., 2015; Keith et al., 2017). However, it highly relies on the human annotations, and the performance of these strategy is still fluctuated and even degraded when directly fine-tuning on a unseen target domain (Zhao et al., 2021) To alleviate this problem, unsupervised domain adaptation (UDA) has been proposed to capture the domain-invariant knowledge from data-rich source domain and unlabeled target data to make the model easier adapt to target domain (Bousmalis et al., 2016; Jia et al., 2019; Naik and Rosé, 2020; Ngo et al., 2021). For example, Bousmalis et al. (2016) presents Domain Separation Network (DSN) to capture the orthogonal representations from shared and private modules; Naik and Rosé (2020) leverages adversarial learning for domain label discrimination to capture domain-invariant semantics.

There have been many studies on English NER(Kim et al., 2017; Jia et al., 2019; Naik and Rosé, 2020). The main idea is to make the token-level representation learning from source domain better adapt to the target domain. In contrast, Chinese named entities are hard to recognize due to the rich semantic knowledge and ambiguous boundaries information of the Chinese token. Recently methods include (Xu et al., 2018b; Yang et al., 2018; Sheng et al., 2020) enhance the representation of Chinese word segmentation across domains. However, the generalization performance of these methods is still poor. We observe that there are still domain gaps at the both word-level and sentence-level between the source domain and the target domain.

To address these issues, we introduce a novel framework named *DAC-NER* for unsupervised **D**omain **A**daption with **C**ontrastive learning for Chinese **NER**. We further extend DSN (Bousmalis et al., 2016) to learn the domain-invariant infor-

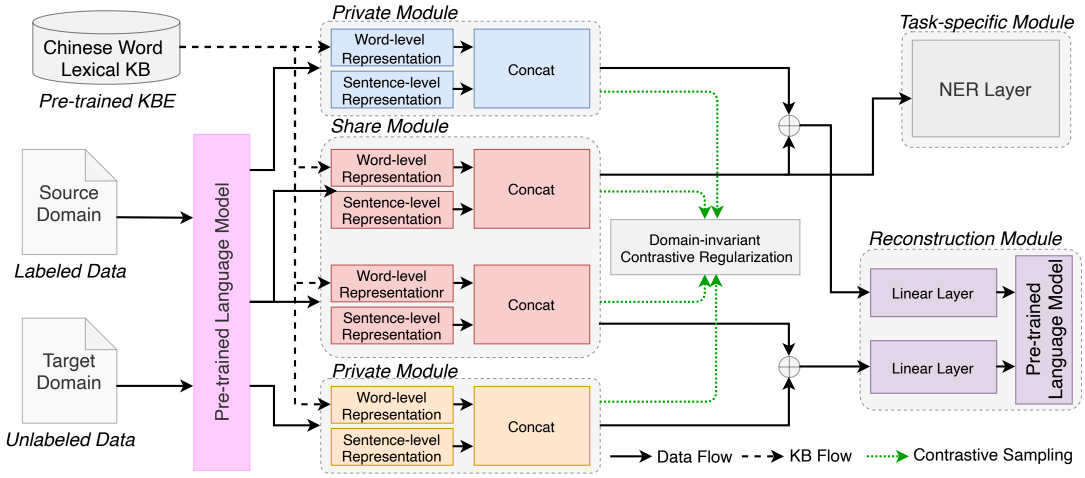

Figure 1: The architecture of our proposed *DAC-NER*. (Best viewed in color.)

mation from both word-level and sentence-level to make the model better adapt to target domain. Concretely, given labeled data from source domain and unlabeled data from target domain, we first leverage Pre-trained Language Model (PLM) to represent the rich semantics from the given sentence. To further improve the generalization of the model, we propose *Knowledge-aware Domain Injector* (KDI) to capture both word-level and sentence-level domain knowledge. In details, we first inject pre-trained Chinese lexical Knowledge Base Embeddings (KBE) with context-aware embeddings from the PLM to enhance the semantics of Chinese word segmentation. Then, we combine it with the sentence-level semantics from the representation of `[CLS]` token. During the training stage, we propose contrastive regularization learning for capturing the domain-invariant meta-knowledge, which is the novel technique for domain adaptation. Specially, we leverage dropout sampling mechanism to generate more augment representations. In addition, we follow (Bousmalis et al., 2016) to introduce a shared PLM decoder as a regular method to enable the separated private and share features to be reconstructed back.

The contributions of this paper are three fold:

- We propose an UDA framework based on contrastive learning. To our knowledge, it is the first to apply contrastive learning to cross-domain adaptation for Chinese NER.

- In *DAC-NER*, a novel Knowledge-aware Domain Injector is proposed to capture the domain-invariant knowledge both from word-level and sentence-level semantics.

- Experimental results over three datasets suggest that our framework consistently outperform state-of-the-arts by a relative large margin.

2 *DAC-NER*: The Proposed Framework

We start with a brief overview of the *DAC-NER*, and followed by the details of the framework. The architecture of *DAC-NER* is shown in Figure 1.

2.1 Brief Overview

We introduce some basic notations. Given two training set $\mathcal{D}_s$ and $\mathcal{D}_t$, where $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{N_s}$ denotes the source domain dataset with N_s labeled training samples, $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ denotes the target domain dataset with N_t unlabeled training samples. The goal of our framework is to transfer the semantics knowledge from the labeled source domain data to unlabeled target domain data, and to recognize the named entities over target domain evaluating samples. Specifically, we first seek to learn the word-level and sentence-level domain specific knowledge from source domain and target domain, respectively. Particularly, we introduce two word-level representation mapping $\mathcal{E}_{wr}$ and $\mathcal{E}_{kn}$. $\mathcal{E}_{wr} : x \to \mathbf{w}$ is the word embedding mapping that maps the word w to the vector $\mathbf{w}$ retrieved from the embedding table of PLM. $\mathcal{E}_{kn} : x \to \mathbf{e}$ is the knowledge base mapping that maps the word w to the vector $\mathbf{e}$ pre-trained by the ConVE algorithm(Dettmers et al., 2018).

To capture the domain-invariant knowledge, we follow the strategy of DSN (Bousmalis et al., 2016) to introduce a share module named domain-invariant injector and private module named domain-specific injector. Additionally, we also find that the adversarial training process in DSN is so hard, which fails to generalize on unseen domain over Chinese task. In contrast, we propose a novel contrastive regularization to learn the domain-invariant knowledge, which is more suitable for the model generalization.

2.2 Main Architecture

To further enhance the abilities of adaptation, we introduce the universal encoder PLM for the representation learning. Let $\mathbf{x}^s = x_1^s, x_2^s, \cdots, x_n^s$ and $\mathbf{x}^t = x_1^t, x_2^t, \cdots, x_n^t$ denote the two input sentences [2] that randomly select from source domain and target domain, respectively. $\mathbf{x}^s \in \mathcal{D}_s$, $\mathbf{x}^t \in \mathcal{D}_t$. n_s, n_t represent the length of sentence from source domain and target domain, respectively. We first encode each sentence [3] by a parameter-shared PLM to obtain the context-aware word embeddings: $\mathbf{w}_i^s = \mathcal{E}_{wr}(x_i^s)$, $\mathbf{w}_i^t = \mathcal{E}_{wr}(x_i^t)$, where $\mathbf{w}_i^s$, $\mathbf{w}_i^t$ is the pre-trained word embeddings of x_i^s and x_i^t, respectively. Based on these semantic-rich representations, we propose series module to capture domain knowledge.

In addition, We find the share-private paradigm (Bousmalis et al., 2016) where the model learns the share semantics knowledge from cross-domain and the independent domain-specific knowledge are suitable for UDA (Kim et al., 2017). In this paper, we follow this structure and propose *Knowledge-aware Domain Injection* (*KDI*) to further enhance the domain adaptation for Chinese NER. Specifically, *KDI* requires the share module to capture the public knowledge both from both source domain and target domain with the same parameters. Tow other independent private modules are designed to learn domain-specific representations. We denote the parameters of the share module, the private module over source domain data and the private module over target domain data as $\boldsymbol{\Theta}^c$, $\boldsymbol{\Theta}^{ps}$ and $\boldsymbol{\Theta}^{pt}$, respectively.

2.3 Knowledge-aware Domain Injection

To enhance the model capability and features, we simultaneously capture word-level and sentence-level representations.

Word-level Representations. To enhance the Chinese word representations, we propose to inject pre-trained knowledge base embeddings with word embeddings. For example, given a Chinese word x_i^s from source domain, we retrieve the entities from the knowledge base [4] that have the same lemma with the word, and the averaged entity embeddings are stored as their knowledge base embeddings. Formally, we generate the knowledge base representation $\mathbf{e}_i^s$ of x_i^s:

$$\mathbf{e}_i^s = Mean(\mathcal{E}_{kn}(e_j)|lemma(x_i^s) = lemma(e_j)), \tag{1}$$

where $lemma$ is the lemmatization operator (Dai et al., 2021). The calculation of knowledge base representations of word x_i^t from target domain is the same as Equal 1.

We then use the gating mechanism to inject the knowledge retrieved from the knowledge base to plain word embeddings. For example, we can obtain the knowledge-enhanced word embeddings of x_i^s through the private module over source domain data, formally:

$$\mathbf{g}_i^{ps} = gate_i^{ps} \cdot \mathbf{w}_i^s + (1 - gate_i^{ps}) \cdot \mathbf{e}_i^s, \tag{2}$$

where i denotes the i-th word in sentence. $gate_i^{ps} \in [0, 1]$ is the trainable gating coefficient. In the same way, we can obtain four representations $\mathbf{g}_i^{ps}, \mathbf{g}_i^{pt}, \mathbf{g}_i^{cs}, \mathbf{g}_i^{ct} \in \mathbb{R}^{n \times h}$ through all modules and domains, respectively. The corresponding matrix are denoted as $\mathbf{G}^{ps}, \mathbf{G}^{pt}, \mathbf{G}^{cs}, \mathbf{G}^{ct} \in \mathbb{R}^{n \times h}$, h is the hidden dimension. The trainable gating coefficients are $gate_i^{ps} \in \boldsymbol{\Theta}^{ps}, gate_i^{pt} \in \boldsymbol{\Theta}^{pt}, gate_i^{cs} \in \boldsymbol{\Theta}^c, gate_i^{ct} \in \boldsymbol{\Theta}^c$.

Sentence-level Representations. As the discussion above, Chinese words have different semantics when demonstrate over cross-domain, which highly guided by the sentence-level representations. We select GRUs (Cho et al., 2014) to encode the context-ware word embeddings [5]. We also take the example of calculation through private module over source domain data, formally:

$$\overrightarrow{\mathbf{s}_j^{ps}} = \overrightarrow{\text{GRU}}^{ps}(\overrightarrow{\mathbf{s}_{j-1}^{ps}}, \mathbf{w}_j^s), \tag{3}$$

$$\overleftarrow{\mathbf{s}_j} = \overleftarrow{\text{GRU}}^{ps}(\overleftarrow{\mathbf{s}_{j+1}^{ps}}, \mathbf{w}_j^s), \tag{4}$$

$$\mathbf{s}_j^{ps} = [\overrightarrow{\mathbf{s}_j^{ps}}; \overleftarrow{\mathbf{s}_j^{ps}}], \tag{5}$$

[2] In the experimental settings, we consider that the length of two sentence is the same.

[3] We add series special tokens to support for encoding by PLM, such as `[CLS]` and `[SEP]`.

[4] We use SKCC (Wang and Yu, 2003) as our knowledge base, the pre-training algorithm is ConVE (Dettmers et al., 2018).

[5] A straightforward way is to use the embeddings of `[CLS]` from PLM as the sentence representations. However, it is not effect for share-private paradigm because that it has no parameters.

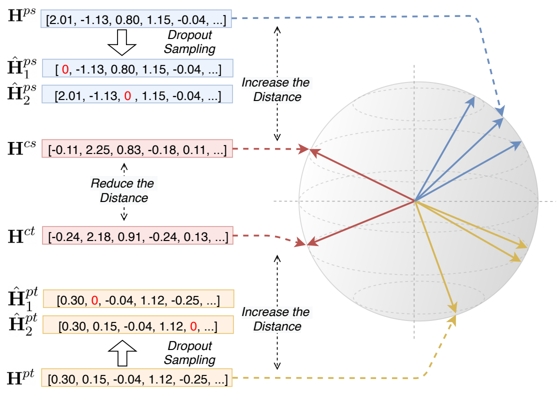

Figure 2: An example of contrastive regularization with dropout sampling.

where $[\cdot;\cdot]$ represents the vector concatenate. At last, we will obtain all sentence-level representations $\mathbf{s_j}^{ps}, \mathbf{s_j}^{pt}, \mathbf{s_j}^{cs}, \mathbf{s_j}^{ct}$ through all modules and domains. The corresponding matrix are denoted as $\mathbf{S}^{ps}, \mathbf{S}^{pt}, \mathbf{S}^{cs}, \mathbf{S}^{ct} \in \mathbb{R}^{n \times h}$.

2.4 Domain-knowledge Projection

To enhance the representations, we project the word-level embeddings and sentence-level embeddings into domain representations. Specifically, we have:

$$\mathbf{H}^{ps} = \texttt{Relu}(\mathbf{W}^{ps}[\mathbf{S}^{ps}, \mathbf{G}^{ps}] + \mathbf{b}^{ps}), \quad (6)$$

$$\mathbf{H}^{pt} = \texttt{Relu}(\mathbf{W}^{pt}[\mathbf{S}^{pt}, \mathbf{G}^{pt}] + \mathbf{b}^{pt}), \quad (7)$$

$$\mathbf{H}^{cs} = \texttt{Relu}(\mathbf{W}^{cs}[\mathbf{S}^{cs}, \mathbf{G}^{cs}] + \mathbf{b}^{cs}), \quad (8)$$

$$\mathbf{H}^{ct} = \texttt{Relu}(\mathbf{W}^{ct}[\mathbf{S}^{ct}, \mathbf{G}^{ct}] + \mathbf{b}^{ct}), \quad (9)$$

where $\mathbf{W}^{ps}$, $\mathbf{W}^{pt}$, $\mathbf{W}^{cs}$, $\mathbf{W}^{ct}$ are the trainable matrix, $\mathbf{b}^{ps}$, $\mathbf{b}^{pt}$, $\mathbf{b}^{cs}$, $\mathbf{b}^{ct}$ are the trainable bias parameters. `Relu` is the activation function.

2.5 Training Schemes of *DAC-NER*

In the previous sections, we describe the model architecture of *DAC-NER* aims to capture the enhanced domain knowledge. In this part, we outline show the training schemes for domain adaptation for Chinese NER.

Named Entity Recognition (NER). In the training stage, we only have labeled data from source domain. We follow (Kim et al., 2017) to view NER as a sequence labeling problem. At the head of PLM, we choose Conditional Random Field (CRF) as the predictor. Specifically, we choose the representations $\mathbf{H}^{ps}$ of source domain data through the share module. The task-specific loss can be written as:

$$\mathcal{L}_{NER} = -\sum_{(\mathbf{x}_i^t, \mathbf{y}_i^s) \in \mathcal{D}_s} \log p(\mathbf{y}_i^s | \mathbf{x}_i^t, \mathbf{\Theta}_1), \quad (10)$$

where $\mathbf{\Theta}_1$ is the parameters of CRF layer.

Domain-invariant Contrastive Regularization (DCR). The main significant part for UDA is to capture the domain-invariant meta-knowledge from cross-domain. Previous methods (Bousmalis et al., 2016; Huang et al., 2015; Zhang and Yang, 2018) generally leverage adversarial learning to let the model unable to classify the domain class. However, we find this method only captures the instance-level meta-knowledge, it is not suitable for word-level because that the semantics of each word in cross-domain are different.

In this paper, we leverage contrastive learning for capturing domain-invariant knowledge. The representations $\mathbf{H}^{cs}$ of source domain sentence through share module fully contains the transferable knowledge (Bousmalis et al., 2016), which can be used as an anchor. Intuitively, if the representations of target domain data learned by the model are similar to $\mathbf{H}^{cs}$, the model will be easy to adapt the domain-specific knowledge to unlabeled target domain due to the similar meta-knowledge. Specifically, the goal is to reduce the semantics distance between $\mathbf{H}^{cs}$ and $\mathbf{H}^{ct}$, and increase the semantics gap between $\mathbf{H}^{cs}$ and $\mathbf{H}^{ps}$, $\mathbf{H}^{ct}$ and $\mathbf{H}^{pt}$, respectively.

As shown in Figure 2, to further enhance the generalization, we use dropout mechanism (Xu et al., 2018a) to randomly reset the value in each representations as 0. We denote $\mathcal{V}(\mathbf{x})$ as the dropout function to map the input vector $\mathbf{x}$ into series of pseudo vectors. In addition, we select cosine similarity function $\mathcal{S}(\mathbf{x}, \mathbf{y})$ to represent the distance between two given vectors $\mathbf{x}$ and $\mathbf{y}$. The contrastive loss function can be calculated as:

$$\mathcal{L}_{DCR} = -\frac{1}{|\mathcal{V}(\mathbf{H}^{ps})|} \times \sum_{i=1}^{|\mathcal{V}(\mathbf{H}^{ps})|} \log\left[\frac{\exp \mathcal{S}(\mathbf{H}^{cs}, \mathbf{H}^{ct})}{\exp \mathcal{S}(\mathbf{H}^{cs}, \mathbf{H}^{ct}) + \exp \mathcal{S}(\mathbf{H}^{cs}, \hat{\mathbf{H}}_i^{ps})}\right] + \frac{1}{|\mathcal{V}(\mathbf{H}^{pt})|} \times \sum_{i=1}^{|\mathcal{V}(\mathbf{H}^{pt})|} \log\left[\frac{\exp \mathcal{S}(\mathbf{H}^{cs}, \mathbf{H}^{ct})}{\exp \mathcal{S}(\mathbf{H}^{cs}, \mathbf{H}^{ct}) + \exp \mathcal{S}(\mathbf{H}^{ct}, \hat{\mathbf{H}}_i^{pt})}\right] \quad (11)$$

where $|\cdot|$ denotes the number of dropout sampling

Dataset	#train	#dev	#test	All
MSRA	40551	4506	3442	48499
Weibo	1256	249	249	1754
People's Daily	20510	2278	4558	27346

Table 1: Dataset Statistics

vectors, and is the hyper-parameter in our experiment.

Reconstruction Learning (RL). Following previous work (Bousmalis et al., 2016; Kim et al., 2017), we reconstruct the hidden representations to the origin input sentences. Formally, we obtain the two concatenated vectors from each domain, and then feed them into one linear layer to map the hidden representations into the dimension h. We have:

$$\hat{\mathbf{H}}^s = \mathbf{W}^s[\mathbf{H}^{ps}; \mathbf{H}^{cs}], \hat{\mathbf{H}}^t = \mathbf{W}^t[\mathbf{H}^{pt}; \mathbf{H}^{ct}], \tag{12}$$

where $\mathbf{W}^s, \mathbf{W}^t$ are the trainable parameters. We use these two representations $\hat{\mathbf{H}}^s, \hat{\mathbf{H}}^t$ to re-generate the input sentences $\mathbf{x}^s, \mathbf{x}^t$ by the PLM [6], respectively. We follow Kim et al. (2017) to use averaged log loss function:

$$\begin{aligned}\mathcal{L}_{RL} = &- \sum_{(\mathbf{x}_i^s) \in \mathcal{D}_s} \frac{1}{n} \sum_{j=1}^{n} \log p(x_j^s | \hat{\mathbf{H}}^s, \mathbf{\Theta}_2) \\ &- \sum_{(\mathbf{x}_i^t) \in \mathcal{D}_t} \frac{1}{n} \sum_{j=1}^{n} \log p(x_j^t | \hat{\mathbf{H}}^t, \mathbf{\Theta}_2),\end{aligned} \tag{13}$$

where $\mathbf{\Theta_2}$ is the parameter of the reconstruction PLM.

Joint Learning. For unsupervised domain adaptation learning, we joint optimize our framework *DAC-NER* by the formula:

$$\mathcal{L} = \mathcal{L}_{NER} + \lambda \mathcal{L}_{DCR} + \mu \mathcal{L}_{RL} + \alpha ||\Theta||^2, \tag{14}$$

where $\lambda \in [0, 1]$, $\mu \in [0, 1]$ are the coefficient hyper-parameters of training loss, $\alpha > 0$ is the L2 regularization value.

3 Experiments

In this section, we conduct a series of experiments on public datasets to evaluate our approach.

3.1 Experimental Settings

Datasets. We employ 3 widely used datasets in our experiments to evaluate the *DAC-NER*, including MSRA (Levow, 2006), Weibo (Peng and Dredze, 2015, 2016) and People's Daily [7]. The distribution of the datasets as shown in Table 1. We take the MSRA as the source domain training data, which has a large amount of labeled data. We use the Weibo and Peoples' Daily as two target domain dataset, respectively. There are three types of entities in MSRA and People's Daily: PER (person), ORG (organization) and LOC (location). Weibo contains four entities: PER, ORG, LOC and GPE (geo-political). Each of these entities has two subtypes, named and nominally mentioned. We only use named types and incorporate GPE in the category of LOC.

Baseline. We use RoBERTa (Liu et al., 2019) as the backbone, which has a stronger ability to extract general context features of the text. For the baseline, we choose CRF as the NER predictor to obtain the results of the model without domain adaptation over the two target domains data. In addition, we also select the basic framework of Adversarial Domain Adaptation as our baseline.

Settings. In the main experiment, the batch size denotes to 64. Specifically, we train our model by the Adam optimizer (Kingma and Ba, 2015). The learning rate for all training stages is fixed to be 1e-5. The training epoch is set to 100, and the model is saved according to the evaluation performances on the development set. The representations dimension of the *Domain-knowledge Projection* module is set to 100. For the external Chinese lexical knowledge base, the hidden dimension of pre-trained KBE is 768, which is the same with RoBERTa. For evaluation, the metrics we selected consist of precision (P), recall (R) and corresponding F1 value (F1).

3.2 Main Results

The general experimental results are shown in Table 2. The method in bold is our approach. ADA (Naik and Rosé, 2020) represents the method which leverages adversarial training. Source and Target denote the labeled source domain training data and the unlabeled target domain training data, respectively. For the baseline, we only train the model over MSRA training data. For the other two UDA methods, there are two settings, one is with MSRA as the source domain, Weibo as the target domain, and the other is with People's

[6] During re-generating, the model structure is the same as the universal encoder PLM, but the parameters are different.

[7] URL: https://libraries.indiana.edu/peoples-daily-database

Methods	Source	Target	MSRA			Weibo			People's Daily		
			P	R	F1	P	R	F1	P	R	F1
Baseline	MSRA	-	96.9	95.6	96.3	47.3	31.1	37.5	86.3	90.0	88.1
ADA	MSRA	Weibo	92.5	91.2	91.8	50.5	30.6	38.1	85.2	86.4	85.8
(Naik and Rosé, 2020)	MSRA	People's Daily	93.3	92.7	93.0	52.0	32.1	39.7	92.6	85.5	88.9
DAC-NER	MSRA	Weibo	94.6	95.3	95.0	**52.5**	**41.3**	**46.2**	**90.1**	**92.3**	**91.2**
	MSRA	People's Daily	95.1	95.6	95.3	**50.3**	**64.1**	**56.4**	**96.2**	**91.8**	**93.9**

Table 2: Comparison among baseline, ADA and *DAC-NER* over all testing sets in terms of precision (P), recall (R) and F1 value (F1) (%). The method in bold refers to our approach.

Methods	MSRA			Weibo		
	P	R	F1	P	R	F1
DAC-NER	**94.6**	**95.3**	**95.0**	**52.5**	**41.3**	**46.2**
w/o. RL	93.0	93.7	93.4	51.2	37.9	43.6
w/o. DCR	94.3	94.0	94.2	49.6	36.1	41.8
w/o. KBE	94.2	93.5	93.8	50.9	31.5	38.9
w/o. SenRe	92.8	94.2	93.5	48.3	34.6	40.3

Table 3: Ablations on Weibo. RL: Reconstruction Learning, DCR: Domain-invariant Contrastive Regularization, KBE: external Knowledge Base Embedding, SenRe: Sentence-level Representation.

Daily as the target domain. Experimental results show that: 1) Our framework outperform all state-of-the-art baselines over all dataset. 2) We find that the baseline performs poorly over target domain data which only trains over source domain. In contrast, the method trained with UDA consistently improve the results due to the domain adaptation. 3) Our method enables to extract domain invariant features without any labeled data on target domain. Specifically, our method improves the F1 value by 8.7% over the target data Weibo. 4) When using relatively more unlabeled People's Daily data as the target domain, the results show that our method can also improve the performance on Weibo, demonstrating that our method can extract domain-invariant features that have strong generalization ability and are beneficial to NER classification tasks.

3.3 Ablation Study

In this part, we investigate the influence of different modules in our approach *DAC-NER*. We conducted ablation experiments with four configurations: without Reconstruction Learning (RL) module, without Domain-invariant Contrastive Regularization (DCR) module, without external Knowledge Base Embedding (KBE) and without sentence-level representation (SenRe). As shown in the Table 3, the two columns on the right are the evaluation results on the source domain MSRA and target domain Weibo testing sets. When the KBE is removed, the F1 score drops by 7.3%, and the absence of other modules has relatively little effect on the result, which shows the efficiency of injected external lexical knowledge base into the context-aware word embedding.

3.4 Visualization

We utilize visualization experiments to demonstrate that our framework can generate high-quality domain-invariant representations on the sentence-level. We use t-SNE (Van der Maaten and Hinton, 2008) to show the data distributions of four different representations, including $\mathbf{H}^{ps}$, $\mathbf{H}^{pt}$, $\mathbf{H}^{cs}$ and $\mathbf{H}^{ct}$. As shown in Figure 4. We find that the domain-invariant representations of all domains data are aggregate together, and the corresponding domain-specific representations are independent. It demonstrates the effectiveness of *DAC-NER* in leveraging the sentence-level knowledge and knowledge-enhanced word embeddings to generate high-quality sentence embeddings for unsupervised domain adaptation Chinese NER.

3.5 Case Study

To gain a deeper understanding of the improvements achieved using *DAC-NER*, we conduct a manual analysis of out-of-domain examples which the baseline incorrectly marked but our approach correctly recognized. Figure 3 demonstrates three examples from MSRA, Weibo and People's Daily datasets, showing the comparison between the baseline and our *DAC-NER*. In the MSRA case, the baseline detected the wrong entity end index, but our two adaptive models can both detect it correctly. This shows that *DAC-NER* improves the robustness of the model to some extent by learning domain-invariant representations from the source and target domains. In the Weibo case, we can see that only the model adaptive to Weibo data can correctly identify the ORG entity "Tudou.com", and there is no error in recognizing "Niucha Fifth Avenue" as LOC. This shows that our method has captured the domain context knowledge of social media data, because of the semantic features in the *Knowledge-*

<table>
<tr><th></th><th>MSRA</th><th>Weibo</th><th>People's Daily</th></tr>
<tr><td>Sentence</td><td>祝中国国民党革命委员会第九次全国代表大会圆满成功！
I wish the Ninth National Congress of the Revolutionary Committee of the Chinese Kuomintang a complete success!</td><td>牛叉第五大道超小朋友分享自土豆网音乐
Niucha Fifth Avenue Super Kid share music from Tudou.com</td><td>灾后仅10多分钟，中保财险员工就赶到厂里，与职工们一道进行抗灾施救和查勘理赔工作。
Only more than 10 minutes after the disaster, the employees of China Insurance Property & Casualty Insurance rushed to the factory and worked with them on disaster relief and investigation and settlement of claims.</td></tr>
<tr><td>Gold tags</td><td>祝<|中国国民党革命委员会第九次全国代表大会@ORG|>圆满成功！</td><td>牛叉第五大道超小朋友分享自<|土豆网@ORG|>音乐</td><td>灾后仅10多分钟，<|中保财险@ORG|>员工就赶到厂里，与职工们一道进行抗灾施救和查勘理赔工作。</td></tr>
<tr><td>Baseline</td><td>祝<|中国国民党革命委员会@ORG|>第九次全国代表大会圆满成功！</td><td><|牛叉第五大道@LOC|>超小朋友分享自土豆网音乐</td><td>灾后仅10多分钟，<|中保@ORG|>财险员工就赶到厂里，与职工们一道进行抗灾施救和查勘理赔工作。</td></tr>
<tr><td>DAC-NER (adapt Weibo)</td><td>祝<|中国国民党革命委员会第九次全国代表大会@ORG|>圆满成功！</td><td>牛叉第五大道超小朋友分享自<|土豆网@ORG|>音乐</td><td>灾后仅10多分钟，<|中保财险@ORG|>员工就赶到厂里，与职工们一道进行抗灾施救和查勘理赔工作。</td></tr>
<tr><td>DAC-NER (adapt PD)</td><td>祝<|中国国民党革命委员会第九次全国代表大会@ORG|>圆满成功！</td><td><|牛叉第五大道@LOC|>超小朋友分享自土豆网音乐</td><td>灾后仅10多分钟，<|中保财险@ORG|>员工就赶到厂里，与职工们一道进行抗灾施救和查勘理赔工作。</td></tr>
</table>

Figure 3: Case studies of different settings on MSRA, Weibo and People's Daily, where correct entities are in green and incorrect entities are in red.

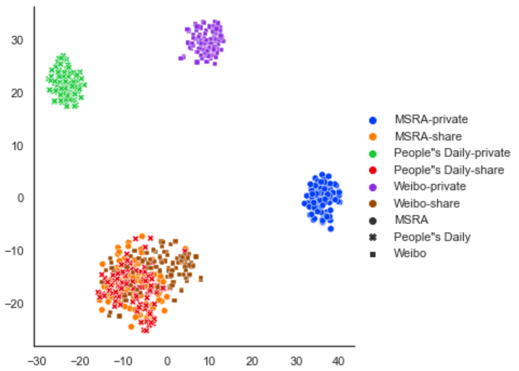

Figure 4: t-SNE visualization of the input representation over all dataset. (Best viewed in color.)

aware Domain Injection module are incorporated, making the representation more context-aware. In the People's Daily case, our two adaptive models both recognize it correctly. In summary, *DAC-NER* can improve the generalization ability on other unlabeled domain data composed of different entity semantics and domain knowledge.

4 Related Work

In this section, we summarize the related work on Chinese NER, unsupervised domain adaptation and contrastive learning.

Chinese NER. Previous work has shown that character-based approaches perform better for Chinese NER than word-based approaches because of the freedom from Chinese word segmentation errors(He and Wang, 2008; Liu et al., 2010; Li et al., 2014). To deal with a dataset with relatively little labeled data, Jia et al. (2020) uses a semisupervised method by using a pre-trained LM. Recently, there is an increasing interest to augment such contextualized representation with external knowledge. These methods focus on augmenting BERT by integrating KG embeddings such as TransE(Bordes et al., 2013). And ERNIE(Sun et al., 2019a,b) enhances BERT through knowledge integration. In our work, we integrate external knowledge to token-embedding for generating more generalize and robust repreasentation.

Unsupervised Domain Adaptation. Existing models for cross-domain Chinese NER rely on numerous unlabeled corpus or labeled NER training data in target domains(Peng and Dredze, 2017; Xu et al., 2018b; Yang et al., 2018). The first UDA method is Jia et al. (2019), requiring an external unlabeled data corpus in both the source and target domains to conduct the unsupervised cross-domain NER task, and such resources are difficult to obtain, especially for low-resource target domains. Liu et al. (2020) is the first to conduct zero-resource cross-domain NER, however, this approach does not meet the situation of Chinese NER, which has a lot of noise and ambiguity.

Some existing methods(Shi et al., 2018; Yang et al., 2018; Naik and Rosé, 2020) leverage adversarial domain adaptation for capturing domain-invariant knowledge. Naik and Rosé (2020) manages to reduce event extraction models' reliance on lexical features. Specifically, Yang et al. (2018) uses a common Bi-LSTM and a private Bi-LSTM for representing annotator generic and -specific information. In adversarial part, they exploit both the source sentences and the crowd-annotated NE labels as basic inputs for the worker discrimination. Our approach is in line with existing works using domain separation network for introducing domain-invariant features.

Contrastive Learning. There are a lot of researches on contrastive learning recently (Le-Khac et al., 2020), including Computer Vision field (He et al., 2019; Chen et al., 2020) and NLP field (Giorgi et al., 2020). A recent work (Wang et al., 2021) on UDA builds upon contrastive self-supervised learning to align features, using a clustering method to produce pseudo-labels for the target domain. They are instance-level, specifically, given an anchor image from one domain and minimize its distances to cross-domain samples from the same class relative to those from different categories. In contrast, our method aligns features based on CL in the shared representation space, which is simple and easy to optimize. To the best of our knowledge, we are the first to use CL in the representation-level.

5 Conclusion

In this paper, we present the *DAC-NER* framework for unsupervised domain adaptation Chinese NER. We propose to inject external Chinese lexical knowledge into context-aware embeddings, and combine with sentence-level representations to enhance the Chinese word segmentation. In addition, we utilize contrastive regularization learning to learn the domain-invariant knowledge for better adaptation. Experiments on a variety dataset show that our framework outperform state-of-the-art baselines. In the future, we will extend our framework into semi-supervised learning and few-shot learning.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 343–351.

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

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734. ACL.

Damai Dai, Hua Zheng, Zhifang Sui, and Baobao Chang. 2021. Incorporating connections beyond knowledge embeddings: A plug-and-play module to enhance commonsense reasoning in machine reading comprehension. *CoRR*, abs/2103.14443.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. In *AAAI*, pages 1811–1818. AAAI Press.

John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. Cnn-based chinese ner with lexicon rethinking. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4982–4988. International Joint Conferences on Artificial Intelligence Organization.

Jingzhou He and Houfeng Wang. 2008. Chinese named entity recognition and word segmentation based on character. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020. Entity enhanced BERT pre-training for Chinese NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6384–6396, Online. Association for Computational Linguistics.

Katherine A. Keith, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor. 2017. Identifying civilians killed by police with distantly supervised entity-event extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1547–1557. Association for Computational Linguistics.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Adversarial adaptation of synthetic or stale data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1297–1307. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Haibo Li, Masato Hagiwara, Qi Li, and Heng Ji. 2014. Comparison of the impact of word segmentation on name tagging for Chinese and Japanese. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2532–2536, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Zhangxun Liu, Conghui Zhu, and Tiejun Zhao. 2010. Chinese named entity recognition with a sequence labeling approach: Based on characters, or based on words? In *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, pages 634–640, Berlin, Heidelberg. Springer Berlin Heidelberg.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020. Zero-resource cross-domain named entity recognition. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 1–6, Online. Association for Computational Linguistics.

Aakanksha Naik and Carolyn Penstein Rosé. 2020. Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7618–7624. Association for Computational Linguistics.

Nghia Trung Ngo, Duy Phung, and Thien Huu Nguyen. 2021. Unsupervised domain adaptation for event detection using domain-specific adapters. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4015–4025. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. *arXiv preprint arXiv:1603.00786*.

Nanyun Peng and Mark Dredze. 2017. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 91–100. Association for Computational Linguistics.

Jiabao Sheng, Aishan Wumaier, and Zhe Li. 2020. POISE: efficient cross-domain chinese named entity recognization via transfer learning. *Symmetry*, 12(10):1673.

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and Heyan Huang. 2018. Genre separation network with adversarial training for cross-genre relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1023, Brussels, Belgium. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019a. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019b. ERNIE 2.0: A continual pre-training framework for language understanding. *CoRR*, abs/1907.12412.

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

Hui Wang and Shiwen Yu. 2003. A large-scale lexical semantic knowledge-base of Chinese. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, pages 243–250, Sentosa, Singapore. COLIPS PUBLICATIONS.

Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. 2021. Cross-domain contrastive learning for unsupervised domain adaptation. *CoRR*, abs/2106.05528.

Qisen Xi, Yizhi Ren, Siyu Yao, Guohua Wu, Gongxun Miao, and Zhen Zhang. 2021. Chinese named entity recognition: Applications and challenges. In Yan Jia, Zhaoquan Gu, and Aiping Li, editors, *MDATA: A New Knowledge Representation Model - Theory, Methods and Applications*, volume 12647 of *Lecture Notes in Computer Science*, pages 51–81. Springer.

Hengru Xu, Shen Li, Renfen Hu, Si Li, and Sheng Gao. 2018a. From random to supervised: A novel dropout mechanism integrated with global information. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 573–582. Association for Computational Linguistics.

Jingjing Xu, Hangfeng He, Xu Sun, Xuancheng Ren, and Sujian Li. 2018b. Cross-domain and semisupervised named entity recognition in chinese social media: A unified model. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(11):2142–2152.

YaoSheng Yang, Meishan Zhang, Wenliang Chen, Wei Zhang, Haofen Wang, and Min Zhang. 2018. Adversarial learning for chinese NER from crowd annotations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1627–1635. AAAI Press.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. *CoRR*, abs/1805.02023.

Shan Zhao, Minghao Hu, Zhiping Cai, Haiwen Chen, and Fang Liu. 2021. Dynamic modeling cross- and self-lattice attention network for chinese NER. 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 14515–14523. AAAI Press.