Improving Named Entity Recognition via Bridge-based Domain Adaptation

Jingyun Xu[1,2], **Changmeng Zheng**[3], **Yi Cai**[1,2]†, **Tat-Seng Chua**[4]
[1] School of Software Engineering, South China University of Technology, China
[2] Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China
[3] Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
[4] School of Computing, National University of Singapore, Singapore
jingyun.x@qq.com, csczheng@comp.polyu.edu.hk,
ycai@scut.edu.cn, dcscts@nus.edu.sg

Abstract

Recent studies have shown remarkable success in cross-domain named entity recognition (cross-domain NER). Despite the promising results, existing methods mainly utilize pre-training language models like BERT to represent words. As such, the original chaotic representations may challenge them to distinguish entity types of entities, leading to entity type misclassification. To this end, we attempt to utilize contrastive learning to refine the original representations and propose a model-agnostic framework named MoCL for cross-domain NER. Additionally, we respectively combine MoCL with two distinctive cross-domain NER methods and two pre-training language models to explore its generalization ability. Empirical results on six domains show the effectiveness and good generalization ability of MoCL.

1 Introduction

Given a sentence, named entity recognition (NER) aims to extract entities and classify them into pre-defined entity types (Zhu and Li, 2022; Wang et al., 2020). As shown in Table 1, given the sentence S1, a NER model needs to extract the entity "Nova" and classify it into the entity type *person*. Most existing NER models rely on massive annotated data, making it hard to directly apply them to data-limited domains. To this end, many researchers started to explore cross-domain named entity recognition (cross-domain NER) methods (Yang et al., 2022; Chen et al., 2022). This paper focuses on the supervised setting, which generalizes effective representations learned from the source domain to the target domain with small annotated samples of the target domain (DAUME III, 2007).

According to the tagging scheme, previously supervised cross-domain NER approaches can be grouped into two types: (1) compositional labeling-based methods that utilize the monolithic tags to train models, where each token is labeled by a composition tag (*e.g.*, *B-person*) (Liu et al., 2021; Zheng et al., 2022); (2) modular learning-based approaches that decompose the composition tag into two tags, where each token is labeled by an entity boundary tag (*e.g.*, *B*) and an entity type tag (*e.g.*, *person*) (Zhang et al., 2022a).

Input Sentence	Ground Truth	Prediction‡
S1: **Nova** was selected as the official voice of the **2013 Central American Games**	**Nova**: *person* **2013 Central American Games**: *event*	**Nova**: *musicalartist* **2013 Central American Games**: *event*

Table 1: An example of entity type misclassification from the CrossNER *music* dataset (Liu et al., 2021). Entities are shown in Bold. The entity types shown in blue are correct while the red one is wrong.

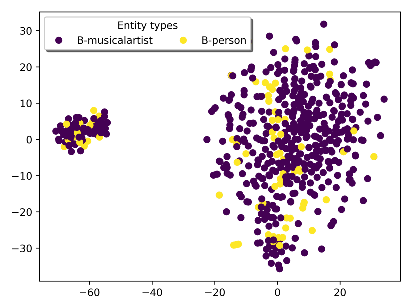


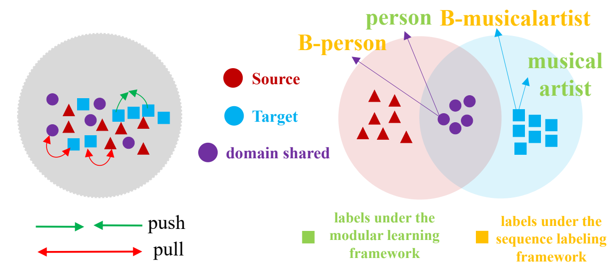
Figure 1: The t-SNE visualization of the representations of entities from the CrossNER *music* dataset under the compositional labeling-based framework in the BERT embedding space (Kenton and Toutanova, 2019).

Despite the promising results, both types of approaches mainly leverage pre-training language models like BERT to represent words. As such, the original chaotic representations (Li et al., 2020) may bring challenges for models to distinguish en-

†Corresponding author.
‡The results are predicted by a state-of-the-art model (Zhang et al., 2022a).

Findings of the Association for Computational Linguistics: ACL 2023, pages 3869–3882
July 9-14, 2023 ©2023 Association for Computational Linguistics

tities with different entity types of corresponding domains, leading to *entity type misclassification*. Let us consider S1 again. Through visualization shown in Figure 1, we observe that the representations of entities with entity types “person” of the source domain and “musicalartist” of the target domain are mixed. As such, as shown in Table 1, even the state-of-the-art method may struggle to successfully distinguish entities belonging to the entity types “person” and “musicalartist”, and hence wrongly classify the entity “Nova” into the incorrect entity type “musicalartist” rather than the correct entity type “person”. Recently, contrastive learning has achieved remarkable success in computer vision, which could generate discriminative representations based on queries and keys (He et al., 2020; Chen et al., 2020). Motivated by this, we attempt to utilize contrastive learning to solve entity-type misclassification faced by the above two kinds of methods by refining the original chaotic representations.



(a) orignial representations (b) refined representations

Figure 2: The illustration of our proposed framework MoCL. Different shapes and colors (*e.g.*, red, blue, and purple) represent the entity types and domains of entities, respectively. The tags adopted by the existing two types of mainstream models (*i.e.*, compositional-labeling (Liu et al., 2021) and modular learning-based) are colored in yellow and green, respectively. For simplicity, we only draw the labels of entity type classification in the modular learning-based approach (Zhang et al., 2022a). **Left**: the original representation of entities, where the entities of the different entity types from the source domain and the target domain are mixed. **Right**: the refined representations after applying MoCL, where entities of the different entity types from the source domain and target domain are separated.

In this paper, we propose a **m**omentum **c**ontrastive **l**earning-based model-agnostic framework named **MoCL** for cross-domain NER. To guide the learning processing of momentum contrastive learning, we first design two approaches to generate keys[‡] required by contrastive learning and name them **Entity Bridge (EB)** and **Label Bridge (LB)** since they work as bridges to enable knowledge transfer from the data-resource source domain to the data-limited target domain. Then based on the generated keys, as shown in Figure 2 (a), MoCL would explicitly pull closer entity representations belonging to the same entity type. Besides, it would simultaneously push away entity representations belonging to different entity types. Thus, as shown in Figure 2 (b), the distances between entities of different entity types become larger while the distances between entities of the same entity type are reduced, resulting in discriminative representations. To summarize, we make the following contributions:

- To the best of our knowledge, we are the first to utilize contrastive learning to refine the original chaotic representations in cross-domain NER. A model-agnostic framework MoCL is proposed and we respectively combine it with two distinct models and two different pre-training language models to explore its generalization ability.

- In order to guide the process of contrastive learning, we explore two methods to generate keys, namely Entity Bridge (EB) and Label Bridge (LB). With the combination of both bridges, MoCL could capture the relations of entities at different granularities, which have been shown effective for NER (Ma et al., 2022a; Chen et al., 2021a).

- Experimental results show the effectiveness of MoCL and the visualization analysis shows it could provide better separation among different entity types in the embedding space.

2 Model

This paper proposes a contrastive learning-based framework MoCL for cross-domain NER, which facilitates the ability to discriminate entities with different entity types. MoCL mainly consists of two modules: **the Base Cross-NER model** and **the Contrastive Learning framework**. We first introduce the **Base cross-domain NER Model** (Section 2.1) and then describe the **Contrastive Learning Framework** (Section 2.2). Finally, we present the training procedure (Section 2.3). The whole architecture of MoCL is shown in Figure 3.

[‡]Here we denote keys are refined sentences of the originally given sentence.

2.1 Base cross-domain NER Model

The **Base cross-domain NER model** involves a **Base cross-domain NER Encoder** (Section 2.1.1) and an **Output Layer** (Section 2.1.2), is constructed to perform the task of cross-domain NER. The base cross-domain NER Model can be **implemented by different existing cross-NER approaches** (Liu et al., 2021; Zhang et al., 2022a).

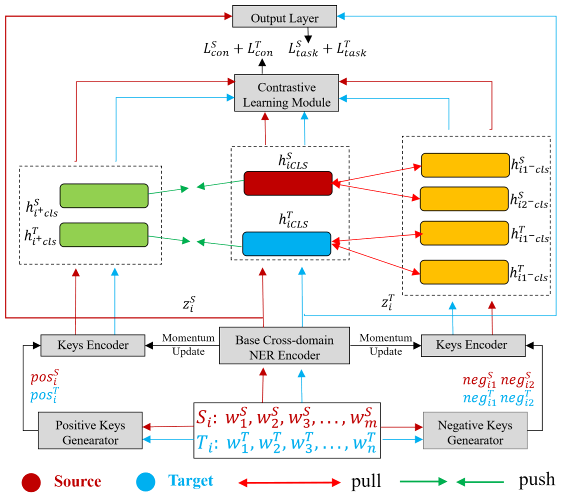


Figure 3: The architecture of MoCL. The rectangles with colors indicate representations of different sentences (red: h^S_{iCLS} of S_i, blue: h^T_{iCLS} of T_i, green: h^S_{i+CLS} of pos^S_i, h^T_{i+CLS} of pos^T_i, and orange: h^S_{i1-CLS} of neg^S_{i1}, h^S_{i2-CLS} of neg^S_{i2}, h^T_{i1-CLS} of neg^T_{i1}, h^t_{i2-CLS} of neg^t_{i2}). MoCL decreases the embedding distance between sentences and their positive keys (shown in the direction of green full lines and arrows outside the sentence embeddings) while pushing away negative keys (shown in the direction of red full blue lines and arrows outside the sentence embeddings).

2.1.1 Base cross-domain NER Encoder

For the cross-domain NER, there are a large set of annotated sentences $S = (S_1, S_2, \ldots, S_{N_s})$ from a source domain and a set of limited sentences $T = (T_1, T_2, \ldots, T_{N_t})$ from a target domain, where D_i denotes the i^{th} sentence of the domain D, and the lengths of the number of sentences are N_s and N_t respectively. Given two sentences $S_i = (w^S_{i1}, w^S_{i2}, \ldots, w^S_{im})$ and $T_i = (w^T_{i1}, w^T_{i2}, \ldots, w^T_{in})$, one from each domain side, here l (m for source domain and n for the target domain, respectively) denote the sentence length ($i.e.$, the total number of words). Each sentence can be constructed as "$[CLS]D_i[SEP]$", where [CLS] and [SEP] denote two special symbols (Kenton and Toutanova, 2019). Then, we feed them into the Base cross-domain NER Encoder, which can be implemented by a pre-trained model like BERT to respectively obtain their hidden representations, denoted as $z^S_i = (h^S_{iCLS}, h^S_{i1}, h^S_{i2}, \ldots, h^S_{im}, h^S_{iSEP})$ and $z^T_i = (h^T_{iCLS}, h^T_{i1}, h^T_{i2}, \ldots, h^T_{in}, h^T_{iSEP})$.

2.1.2 Output Layer

Sequentially, $(h^S_{i1}, h^S_{i2}, \ldots, h^S_{im})$ and $(h^S_{i1}, h^S_{i2}, \ldots, h^S_{in})$ are delivered to an output layer to obtain the types of entities. Then the probability that the j^{th} word in i^{th} sentence of domain D be categorized to the k^{th} entity type $type_k$, denoted by $p(type_k|h^D_{ij})$, can be computed by Softmax function:

$$p(type_k|w^D_{ij}) = \frac{exp\{w^D_k h^D_{ij} + b^D_k\}}{\sum_{g=1}^{c^D} exp\{w^D_g h^D_{ij} + b^D_g\}}. \quad (1)$$

where c^D, w^D_g and b^D_g denotes the number of entity types, the weight and bias parameters in the domain D (source or target), respectively. We then utilize cross-entropy loss to train on the corresponding sentence (S or T) as follows:

$$\mathcal{L}^D_{task} = -\sum_{i=1}^{N_D} \frac{1}{|D_i|} \sum_{j=1}^{l} \sum_{k=1}^{c^D} y_{j,k} log(p(type_k|w^D_{ij})) \quad (2)$$

where $y_{j,k}$ denotes the k^{th} element in y_i, which is an one-hot label indicating the entity type of w^D_{ij}. In terms of the source domain, the training loss is L^S_{task}. When it comes to the target domain, the training loss is L^T_{task}.

2.2 The Contrastive Learning Framework

The **Contrastive Learning Framework** mainly contains three components: 1) the **Keys Generators** (Section 2.2.1); 2) the **Keys Encoder** (Section 2.2.2), and 3) a **Contrastive Learning Module** (Section 2.2.3). The Keys Generators ($i.e.$, Positive Keys Generator and Negative Keys Generator) **can be implemented by our proposed three bridges**. The Contrastive Learning module is designed to allow the model to distinguish entities with respect to their entity types based on the output from the two encoders ($i.e.$, the Base cross-domain NER Encoder, the Key Encoder).

2.2.1 Keys Generator

Motivated by the power of contrastive learning to learn discriminative representations in computer vision, we consider applying it in cross-domain NER. In computer vision, the typical ways to construct keys and queries are such that the query is an original image, and its positive keys are obtained by

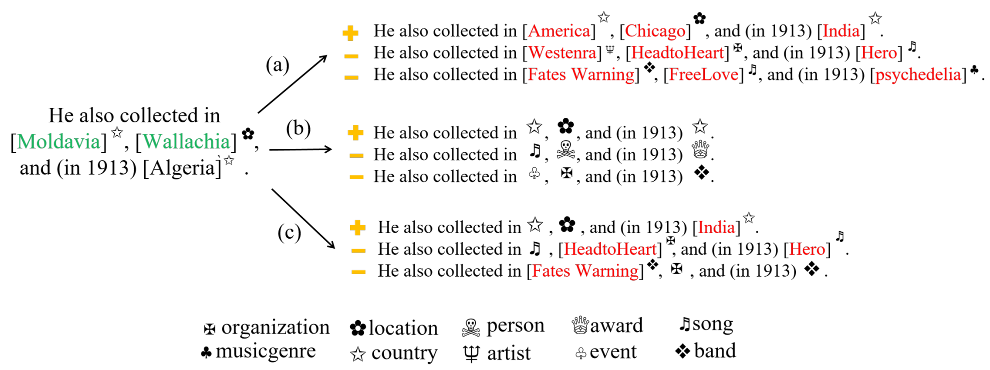


Figure 4: Comparative illustration of three key generation strategies. Plus and minus refer to positive and negative examples. Original entities and their replacements are shown in green and red, respectively. (a) Entity Bridge (b) Label Bridge (c) The combination of (a) and (b).

applying operations like revolving or cutting the same image. In contrast, negative keys are other images (He et al., 2020). However, vanilla momentum contrastive learning is not directly applicable in cross-domain NER. If directly taking the typical data augmentation in cross-NER, some information may be closed in the immediate keys. For instance, given the sentence “He worked in Consortium”, the generated positive key may be “The workplace of him is Consortium” while the generated negative key may be “The workplace of him is not Consortium”. In this way, the model mainly learns “Consortium is a location” instead of “Consortium” belongs to the entity type *ORG*. In fact, the typical way to construct keys and queries neglects the fine-grained entity-type information.

To address this limitation, we explore three different key generation strategies, which can make better use of the fine-grained entity-type information. The detailed key generation strategies are illustrated in Figure 4.

- **Entity Bridge.** Given a sentence, an intuitive way to generate keys is based on entities. In particular, we first use the entities and their entity types from the whole training set to construct a dictionary and we did not use additional dictionaries or knowledge bases. Given a sentence, we would ergodic all entities in a sentence and modify them one by one. In particular, based on one given entity, and its entity type, we would randomly select another entity from the constructed dictionary to replace the entity. For example, suppose there are five entities (entity types are shown in the form of ()) in the training set: XXX (person), YYY (person), and ZZZ (song), AAA (location), BBB (song). the dictionary would be person: XXX, YYY, location: AAA, song: BBB. Then given the sentence A= "XXX and YYY are running". If we want to modify XXX, we would randomly select another entity type except person from the dictionary (e.g., song and location). Suppose the selected entity type is song. Then we would randomly select one entity from entities whose entity type is song (e.g., ZZZ and BBB). Suppose the selected entity is BBB. Then the generated negative key would be BBB is running. Similarly, as shown in Figure 4 (a), we can replace “Moldavia” in the original sentence from the source domain with “America”, where the “America” is a different entity with the same entity type *country* as “Moldavia”. Similarly, we can replace “Wallachia” with “Chicago”, “Algeria” with “India”, from which we can get a positive Key “He also collected in America, Chicago, and (in 1913) India.”. Besides, a negative entity could be generated by replacing “Moldavia” with “Westenra”, an entity belonging to another entity type *award*, which is randomly sampled from the dataset. Finally, we can obtain a negative Key “He also collected in Westenra, HeadtoHeart, and (in 1913) Hero.”.

- **Label Bridge.** In order to leverage the label information of text, which has been shown effective in the cross-domain NER task (Hu et al., 2022), we propose a key generation strategy called Label Bridge. As shown in Figure 4 (b), to get a positive key, we replace the entity “Moldavia” in the original sentence with its entity type *country*, and we can produce a negative key by replacing “Moldavia” with another entity type *song*, a different entity type randomly sampled from the dataset.

- **The combination.** In order to simultaneously utilize the entity and label information, we adopt either the entity bridge or the label bridge randomly with equal likelihood, which has been shown in Figure 4 (c). In terms of the positive key, we can replace "Algeria" with an entity "India", "Moldavia" with its entity type *country*, and "Wallachia" with its entity type *location*. When it comes to the negative key, it can be generated by replacing "Moldavia" with another entity type *song* and replacing "Wallachia" with another entity *HeadtoHeart*.

After applying one of the above strategies, given original sentences S_i and T_i mentioned in Section 2.1.1, a new positive key pos_i^s for S_i and a new positive key pos_i^T for T_i will be generated, respectively. Meanwhile, we generate $N^{\ddagger}$ different negative keys neg_{i1}^S and $neg_{i2}S$ for S_i and two different negative keys neg_{i1}^T and neg_{i2}^T for T_i to make better use of mutual information (Oord et al., 2018).

2.2.2 Keys Encoder

By leveraging one of the above three key generation strategies, we can obtain a total of six keys for each original sentence. Each key X can be constructed as "$[CLS]X[SEP]$", where [CLS] and [SEP] denote two special symbols (Kenton and Toutanova, 2019). Then, we feed them into the keys Encoder, which can be implemented by a pre-trained model like BERT to respectively obtain the corresponding sentence representations, denoted as h_{i+CLS}^S of pos_i^s, h_{i+CLS}^T of pos_i^T, h_{i1-CLS}^S of neg_{i1}^S, h_{i1-CLS}^T of neg_{i1}^T,h_{i2-CLS}^S of neg_{i1}^S, h_{i2-CLS}^T of neg_{i1}^T.

2.2.3 the Contrastive Learning Module

Based on the generated keys, we apply contrastive learning to cross-domain NER by minimizing the distance between representations of entities with the same type and maximizing the distance between representations of entities belonging to different types in order to improve the applicability of the model in the target domain.

Given the above sentence representations, we can calculate the contrast loss for each original sentence and its sampled sentences by:

$$L_{con}^D = -\sum_{i=1}^{N_D} \frac{1}{|D_i|} * log \frac{s(q,k^+)}{s(q,k^+) + \sum_{j=1}^{2} s(q,k^-)} \tag{3}$$

$$s(q,k^+) = s(h_{iCLS}^D, h_{i+CLS}^D)/\tau \tag{4}$$

$$s(q,k^-) = s(h_{iCLS}^D, h_{ij-CLS}^D)/\tau \tag{5}$$

Here s denotes the function to calculate the similarity score by applying the dot product operation between two given embeddings, while τ is a scalar temperature parameter (Wang and Isola, 2020). Based on the similarity score, models could minimize the distance between positive keys and maximize the distance between the negative keys, achieving alignments among entities.

2.3 Model training

Following (He et al., 2020), we utilize momentum update to maintain the stability and to keep the consistency of representations between the Base cross-domain NER Encoder and the keys Encoder. In particular, by having the weights of the networks slowly track the learned networks, which means the keys encoder updates slowly, this can greatly improve the stability during training. Momentum updates can be formulated as:

$$\theta \leftarrow m\theta + (1-m)\theta' \tag{6}$$

where m is a momentum coefficient, which is a relatively large number between 0 and 1, and θ and θ' is the parameter of the Base cross-domain NER Encoder and the keys Encoder respectively.

Eventually, we attempt to minimize the combined loss to train our model by:

$$\mathcal{L} = \mathcal{L}_{task}^S + +\mathcal{L}_{task}^T + \gamma(L_{con}^S + L_{con}^T) \tag{7}$$

where γ is a weight coefficient.

3 Experimental Setups

3.1 Datasets & Evaluation Metrics

We use two datasets for experiments, including one domain Social Media of the dataset Twitter (Lu et al., 2018), and five domains in the dataset CrossNER (Liu et al., 2021). We take Social Media as the source domain and five domains in CrossNER as target domains (Liu et al., 2021). Table 2 shows detailed statistics of each domain and their corresponding entity types are shown in Table 3.

[‡] According to preliminary experiment results, which is described in Section 4.2 A3, we set N to 2.

Table 2: Statistics on the seven domains in our experiments.

	Domain	#Train	#Dev	#Test
Source	Social Media	4290	-	-
Target	Politics	200	541	651
	Science	200	450	543
	Music	200	380	465
	Literature	100	400	416
	AI	100	350	431

Following (Liu et al., 2021) and (Zhang et al., 2022a), we use F1-score to evaluate the performance of models. In particular, an entity is considered to be correct only if its range and entity type are both correct.

3.2 Experimental Settings

We combine MoCL with two pre-training language models, including BERT[‡] (Kenton and Toutanova, 2019) (*i.e.*, the basic setting) and the domain-adaptive pre-training language model (*i.e.*, the DAPT setting) of each target domain (Liu et al., 2021). Following (Liu et al., 2021) and (Zhang et al., 2022a), for the basic setting, we initialize the textual representation by BERT and set the dimension to 768. While for the DAPT setting, following (Liu et al., 2021) we use BERT and unlabeled domain-specific corpus to train a domain-adaptive pre-training language model for each domain[‡]. As for two competitive baseline models **BERT-JF** and **MTD**, we respectively follow the same settings from the implementation of (Liu et al., 2021)[‡] and (Zhang et al., 2022a)[‡] for a fair comparison.[‡]. In order to get the keys required by contrastive learning, we first utilize the training set in each domain to construct dictionaries of each entity type. Then given the sentence from the source domain or target domain, we apply one of the three key generation strategies to generate keys based on the constructed dictionaries. Moreover, we set $\tau = 0.07$ (Eq. 4/5), $m = 0.999$ (Eq. 8). While γ is tuned from 0.1, 0.2, 0.3, 0.5, 0.7, 0.9 1.0 in different settings and finally is set to 0.7 (Politics), 1 (Science, Music), 0.1 (Literature, AI) under the basic setting, 0.1 (Politics, AI, Music), 0.7 (Science), 0.3 (Literature) under the DAPT setting. We implement our model with the PyTorch framework and conduct experiments at Tesla P100 and V100.

[‡]https://huggingface.co/bert-base-cased

[‡]We will release the checkpoints of all domain-adaptive pre-training language models to facilitate further research.

[‡]https://github.com/zliucr/CrossNER

[‡]https://github.com/AIRobotZhang/MTD

[‡]We are highly grateful for their public codes, our code will be publicly available via GitHub.

Table 3: The corresponding entity categories for each cross-domain NER dataset.

Dataset	Entity Categories
CoNLL 2003	person, organization, location, miscellaneous
Twitter	person, organization, location, miscellaneous
Politics	person, organization, politician, political party, location, event, country, election, miscellaneous
Science	person, country, university, scientist, organization, location, miscellaneous, enzyme, protein, discipline chemical element, event, academic journal, award, theory, chemical compound, astronomical object
Music	musicalartist, music genre, band, album, song, award, musical instrument, , event, country, location, organization, person, miscellaneous
Literature	person, organization, writer, award, poem, book, location, country, magazine, event, miscellaneous
AI	location, field, task, product, algorithm person, country, researcher, metrics organization, miscellaneous, university

3.3 Baseline Models

Our baselines are:

- **BiLSTM-CRF**, which combines BiLSTM and CRF to train the model (Lample et al., 2016).
- **LM-NER**, which integrates cross-domain language models (Jia and Zhang, 2020).
- **BERT-PF**, which firstly utilizes the source domain data and then uses the target domain data (Liu et al., 2021).
- **BERT-JF**, which simultaneously utilizes both the source and target domain data (Liu et al., 2021).
- **Style-NER**, a method that applies data augmentation (Chen et al., 2021b).
- **MultiCell-LM**, a method utilizes a separate cell state to model each entity type for domain adaptation (Jia and Zhang, 2020).
- **MTD**, a modular learning-based method that splits cross-domain NER into two sub-tasks (Zhang et al., 2022a).

4 EXPERIMENTAL RESULTS

4.1 Overall Performance

According to Table 4, we observe that: (1) *MTD-MoCL* achieves better performance than no alignment work *BERT-JF* with 8-11% improvements,

Setting	Extra Data	No. (Basic setting)					Yes. (DAPT setting (Liu et al., 2021))				
	Source Domain	Social Media (Twitter) ->									
	Target Domain	Politics	Science	Music	Litera.	AI	Politics	Science	Music	Litera.	AI
Baselines	BiLSTM-CRF	53.64	47.33	48.85	45.23	44.0	-	-	-	-	-
	Style-NER	-	-	-	-	-	70.94	68.28	74.40	67.05	63.33
	LM-NER	66.99	64.23	61.48	59.09	50.46	-	-	-	-	-
	BERT-JF	67.52	64.51	67.74	61.38	57.05	70.78	67.31	68.13	62.69	59.17
	BERT-PF	68.60	62.23	68.06	61.91	54.72	70.11	66.87	73.88	66.61	61.12
	MultiCell-LM	66.59	63.79	66.54	59.02	53.82	69.13	66.76	74.22	64.88	62.41
	MTD	74.62	71.37	74.41	69.67	64.55	75.49	72.81	77.43	70.14	66.18
Ours	BERT-JF-MoCL	71.35	69.01	71.19	64.91	59.98	74.38	71.05	74.41	67.13	62.76
	MTD-MoCL	**75.13**	**72.83**	**77.15**	**70.71**	**67.87**	**77.78**	**75.08**	**80.02**	**72.09**	**69.94**

Table 4: Detailed F1 scores on from the source domain *Social Media* to the five target domains. The best scores are shown in bold.

showing the effectiveness of contrastive learning. (2) *MTD-MoCL* achieves the state-of-the-art performance and beats *MTD* (a representative model of the modular learning-based approaches). Moreover, the performance of *MTD-MoCL* is relatively high when the source domain is Twitter, whose size is smaller than conll2003. This demonstrates that *MoCL* could help methods achieve better performance by refining the original representations, especially in the low-source setting.

4.2 Analysis

A1: The effectiveness of incorporating MoCL with different base cross-domain NER models. As shown in Tables 4, *BERT-JF-MoCL* also achieves better performance than *BERT-JF* (a representative model of the compositional labeling-based approaches). This shows that MoCL can not only benefit compositional labeling-based methods but also modular learning-based methods.

A2: The effectiveness of incorporating MoCL with different pre-training models. we incorporate MoCL with MTD with a domain-adaptive pre-training model (Liu et al., 2021). As shown in Table 4, both *MTD-MoCL* and *BERT-JF-MoCL* respectively outperform *MTD* and *BERT-JF* from across all domains with a noticeable margin, which shows the great generalization ability of MoCL.

N	0	1	2	3	9	18
F1-score	74.41	76.08	77.15	76.37	76.16	75.59

Table 5: Performance of MTD-MoCL with different values of N under the basic setting on the target domain *music*.

A3: Impact of the value of negative samples N. As shown in Table 5, the = 0 means *MTD*, which still can be improved. On the one hand, when N is less than 2, when N increases, the results are better. On the other hand, when N is large than 2, when N increases, the results are worse. As such, we set N to 2.

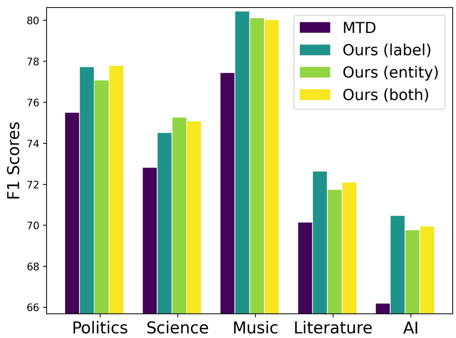
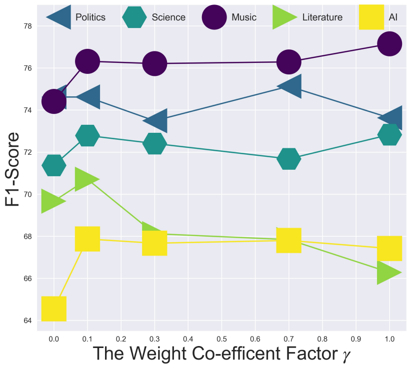


Figure 5: Experimental results of different bridges under DAPT setting.[‡]

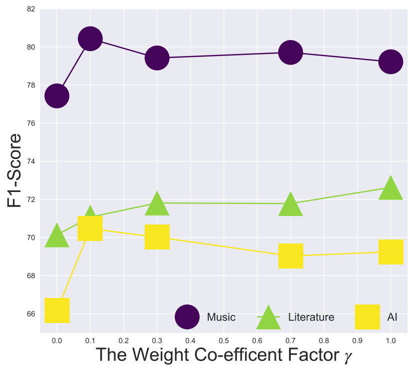
A4: The effectiveness of different bridges. we conduct experiments to investigate the impact of bridges. According to Figure 5, we find that both entity information and label information (*Ours (entity)* VS *MTD*, *Ours (label)* VS *MTD*) are beneficial for learning a better Cross-NER model. Besides, there is no winner always and the performance was improved consistently regardless of the bridges used, which indicates the absolute advantage of contrastive learning.

A5: Impact of γ in Equation 7. In Figure **??**, the γ = 0 means *MTD*, which still can be improved. The influence of γ on domains is different. For domains *Literature*, and *AI*, when γ are smaller, the proposed MTD-MoCL achieved better performance; While for domains *Science*, *Politics*, and

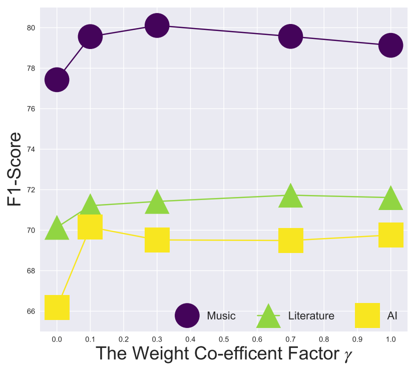
[‡]Here *Ours (entity)* means MTD-MoCL with the Entity Bridge, *Ours (label)* means MTD-MoCL with the Label Bridge, while *Ours (both)* means MTD-MoCL with the combination of both bridges.



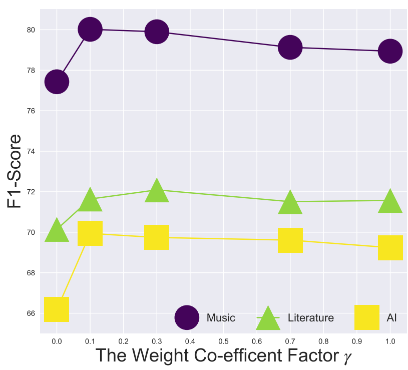
(a) Ours (both) under Basic Setting



(b) Ours (entity) under DAPT Setting

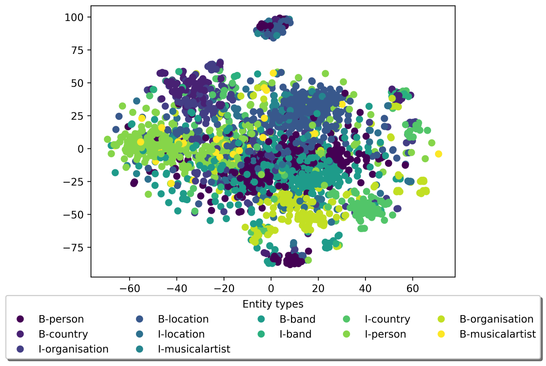


(c) Ours (label) under DAPT Setting

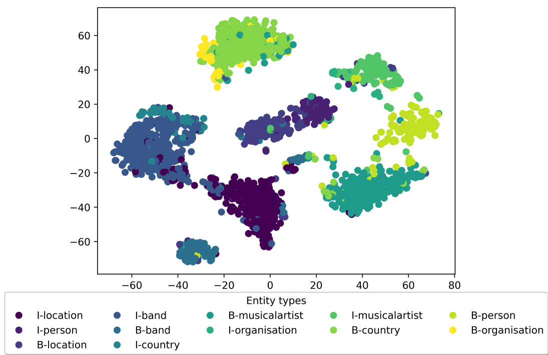


(d) Ours (both) under DAPT Setting

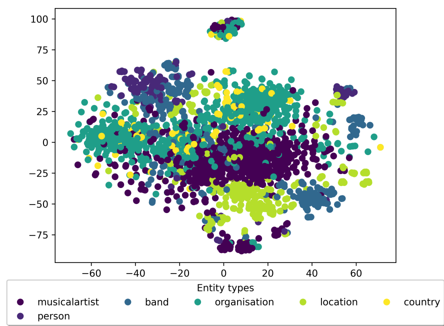
Figure 6: Performance of MTD-MoCL with different values of γ.



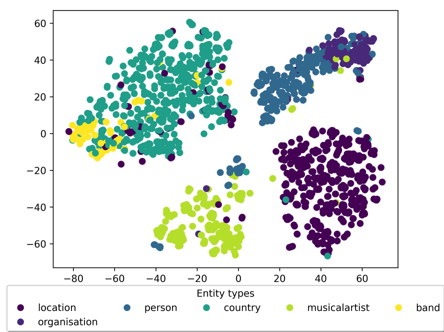
(a) BERT-JF, before



(b) BERT-JF, after



(c) MTD, before



(d) MTD, before

Figure 7: The t-SNE visualization of entity representations on the domain *music*. (a), (b), (c) and (d) are the results before and after applying MoCL with BERT-JF/MTD, respectively.

Music, when γ increases, the results of are better. We also conduct experiments to evaluate the influence of γ with different bridges under the DAPT setting and the results are shown in Figure 6. Similar to Figure 6, the $\gamma = 0$ means *MTD*, which still can be improved regardless of what kind of bridge is applied. Moreover, we observe that during the DAPT setting, the influence of contrastive learning is smaller than in the basic setting ($e.g.$ the basic between the best model and worst is smaller than that in Figure 6). We think after fine-tuning BERT on the large domain-specific corpus, models may learn some discriminative representations. However, compared with applying contrastive learning, training a domain-adaptive pre-training language model is inefficient ($e.g.$, it takes almost 30 hours to train a model for each domain).

A6: Visualization analysis. We conduct visualization analysis to explore the effects of MoCL on the representation of entities. As shown in Figure 7:(1) On one hand, we observe that the original entity representations of the same entity types under the sequence-labeling framework or the modular-learning-based framework disperse sparsely, which is consistent with the observation of (Kenton and Toutanova, 2019). After applying our proposed bridges and contrastive learning, MoCL tries to force the entities belonging to the same entity type to collapse into essentially a close cluster. (2) On the other hand, we observe that the original entity representations of similar entity types under the sequence-labeling framework or the modular-learning-based framework are prone to mix with each other, thus making them hard to be distinguished by the prediction model. In contrast, the entity representations produced by MoCL are clearly separated, which is much more discriminative.

5 Related Work

Due to the capability of extracting useful information and benefiting many NLP applications ($e.g.$, information retrieval (Fetahu et al., 2021; Guo et al., 2009) and question answering (Longpre et al., 2021)), NER appeals to many researchers (Jiang et al., 2021; Feng et al., 2018; Kim et al., 2015; Lee et al., 2018; Qu et al., 2016; Rodriguez et al., 2018; Wang et al., 2018; Zhang et al., 2021b; Yang et al., 2017; Yang and Katiyar, 2020; Fei et al., 2021). Recently, to reduce the huge cost of annotating data, researchers start to explore cross-domain NER methods. According to whether the labeled data of the target domain are used or not, these methods can be classified into unsupervised (Jia et al., 2019; Peng et al., 2021; Chen et al., 2022; Yang et al., 2022; Liu et al., 2022; Ma et al., 2022b; Zhang et al., 2021a) and supervised (Wang et al., 2020; Lin and Lu, 2018; Houlsby et al., 2019; Zheng et al., 2022). This paper focuses on the latter and according to the tagging scheme, supervised cross-domain NER methods can be classified into compositional labeling-based (Wang et al., 2020) and modular learning-based (Zhang et al., 2022a). Compared with previous studies, we attempt to improve both kinds of methods from the perspective of representation. In particular, a model-agnostic framework MoCL is introduced to refine the original chaotic representations by contrastive learning, motivated by its success in computer vision (Radford et al., 2021; Grill et al., 2020; Caron et al., 2020; Chen and He, 2021; Choi et al., 2022; Zhang et al., 2022b; Giorgi et al., 2021; Xu et al., 2022).

6 Conclusion

This paper explores utilizing contrastive learning to gain discriminative entity representations in the field of cross-domain named entity recognition. To guide contrastive learning at the entity level, we explored two bridges to capture different relations of entities at different granularities. Additionally, our framework is model-agnostic, so we respectively integrate it into two existing cross-NER baselines and two different pre-training language models to evaluate its generalization ability. The experimental results show that MoCL could help models learn discriminative representations and it has good generalization ability. In terms of the limitation, currently, we mainly evaluate MoCL under the single-source cross-domain setting. We plan to further extend it to multi-source cross-domain settings. Moreover, the interaction between named entity recognition and relation extraction can be considered to improve performance in the future.

Limitations

We propose a sequence-level contrastive learning-based model-agnostic framework MoCL to enhance entity type classification in cross-domain named entity recognition (NER). In the future, we would like to combine the different granularities of contrastive learning (i.e., token-level and sequence-level) to learn generalized representation for further improving the capability of MoCL. In addition, due to the hierarchical structure of entity types between the source domain and the target domain, it would also be beneficial to adopt Non-Euclidean space to represent words for better learning the relative hierarchical relationship between entities.

Acknowledgement

This work has been supported by the National Natural Science Foundation of China (62076100), Fundamental Research Funds for the Central Universities, SCUT (x2rjD2220050), the Science and Technology Planning Project of Guangdong Province (2020B0101100002), CAAIHuawei MindSpore Open Fund, CCF-Zhipu AI Large Model Fund.

Ethical Considerations

Our research aims to benefit the efforts in delivering domain-adaption technology to data-limited settings. As a key task in NLP, named entity recognition has broad applications, e.g., machine translation, question answering, and potentially protecting endangered languages. Compared with many previous studies, we stress the importance of diversity in the sense that our experiments cover seven domains, including five lower-resource domains from the CrossNER dataset. Hoping that our work can contribute to extending modern NLP techniques to the lower-resource named entity recognition setting. The three datasets we use are both publicly available. To our best knowledge, the data do not contain any sensitive information and have no foreseeable risk.

References

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. volume 33, pages 9912–9924.

Pei Chen, Haibo Ding, Jun Araki, and Ruihong Huang. 2021a. Explicitly capturing relations between entity mentions via graph neural networks for domain-specific named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 735–742.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021b. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356.

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.

Wei Chen, Songqiao Han, and Hailiang Huang. 2022. An empirical cross domain-specific entity recognition with domain vector. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3868–3872.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. 2022. C2l: Causally contrastive learning for robust text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10526–10534.

H DAUME III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Hao Fei, Donghong Ji, Bobo Li, Yijiang Liu, Yafeng Ren, and Fei Li. 2021. Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12785–12793.

Xiaocheng Feng, Xiachong Feng, Bing Qin, Zhangyin Feng, and Ting Liu. 2018. Improving low resource named entity recognition using cross-lingual knowledge transfer. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4071–4077.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer enhanced named entity recognition for code-mixed web queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 879–895.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799.

Jinpeng Hu, He Zhao, Dan Guo, Xiang Wan, and Tsung-Hui Chang. 2022. A label-aware autoregressive framework for cross-domain ner. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2222–2232.

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.

Chen Jia and Yue Zhang. 2020. Multi-cell compositional lstm for ner domain adaptation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5906–5917.

Deng Jiang, Haopeng Ren, Yi Cai, Jingyun Xu, Yanxia Liu, and Ho-fung Leung. 2021. Candidate region aware nested named entity recognition. *Neural Networks*, 142:340–350.

Jacob Devlin Ming-Wei Chang Kenton and Lee 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*, pages 4171–4186.

Young-Bum Kim, Karl Stratos, Ruhi Sarikaya, and Minwoo Jeong. 2015. New transfer learning techniques for disparate label sets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 473–482.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9119–9130.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022.

Luchen Liu, Xixun Lin, Peng Zhang, Lei Zhang, and Bin Wang. 2022. Learning common dependency structure for unsupervised cross-domain ner. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8347–8351. IEEE.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13452–13460.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1990–1999.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022a. Label semantics for few shot named entity recognition. In *ACL 2022*, pages 1956–1971.

Ruotian Ma, Yiding Tan, Xin Zhou, Xuanting Chen, Di Liang, Sirui Wang, Wei Wu, and Tao Gui. 2022b. Searching for optimal subword tokenization in cross-domain ner. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4289–4295.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qi Peng, Changmeng Zheng, Yi Cai, Tao Wang, Haoran Xie, and Qing Li. 2021. Unsupervised cross-domain named entity recognition using entity-aware adversarial training. *Neural Networks*, pages 68–77.

Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, and Timothy Baldwin. 2016. Named entity recognition for novel types by transfer learning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 899–905.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763.

Juan Diego Rodriguez, Adam Caldwell, and Alex Liu. 2018. Transfer learning for entity recognition of novel classes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1974–1985.

Jing Wang, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. 2020. Multi-domain named entity recognition with genre-aware and agnostic inference. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8476–8488.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–15.

Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11556–11565.

Linyi Yang, Lifan Yuan, Leyang Cui, Wenyang Gao, and Yue Zhang. 2022. Factmix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5360–5371.

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*, pages 6365–6375.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *5th International Conference on Learning Representations*.

Tao Zhang, Congying Xia, S Yu Philip, Zhiwei Liu, and Shu Zhao. 2021a. Pdaln: Progressive domain adaptation over a pre-trained model for low-resource cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5441–5451.

Xinghua Zhang, Bowen Yu, Yubin Wang, Tingwen Liu, Taoyu Su, and Hongbo Xu. 2022a. Exploring modular task decomposition in cross-domain named entity recognition. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 301–311.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022b. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11730–11738.

Ying Zhang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021b. Target-oriented fine-tuning for zero-resource named entity recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1603–1615.

Junhao Zheng, Haibin Chen, and Qianli Ma. 2022. Cross-domain named entity recognition via graph matching. In *Findings of the Association for Computational Linguistics*, pages 2670–2680.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7096–7108.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☑ A2. Did you discuss any potential risks of your work?
Left blank.

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

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

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Please refer to Section 3.2 to see more details.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Please refer to Section 3.2 to see more details.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Please refer to Section 3.2 to see more details.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Please refer to Section 3.2 to see more details.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Please refer to Section 3.2 to see more details.

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

Left blank.

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

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

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

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

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