Exploring Modular Task Decomposition in Cross-domain Named Entity Recognition

Xinghua Zhang
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
zhangxinghua@iie.ac.cn

Bowen Yu
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
yubowen@iie.ac.cn

Yubin Wang*
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
wangyubin@iie.ac.cn

Tingwen Liu*
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
liutingwen@iie.ac.cn

Taoyu Su
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
sutaoyu@iie.ac.cn

Hongbo Xu
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
hbxu@iie.ac.cn

ABSTRACT

Cross-domain Named Entity Recognition (NER) aims to transfer knowledge from the source domain to the target, alleviating expensive labeling costs in the target domain. Most prior studies acquire domain-invariant features under the end-to-end sequence-labeling framework where each token is assigned a compositional label (e.g., B-LOC). However, the complexity of cross-domain transfer may be increased over this complicated labeling scheme, which leads to sub-optimal results, especially when there are significantly distinct entity categories across domains. In this paper, we aim to explore the task decomposition in cross-domain NER. Concretely, we suggest a modular learning approach in which two sub-tasks (entity span detection and type classification) are learned by separate functional modules to perform respective cross-domain transfer with corresponding strategies. Compared with the compositional labeling scheme, the label spaces are smaller and closer across domains especially in entity span detection, leading to easier transfer in each sub-task. And then we combine two sub-tasks to achieve the final result with modular interaction mechanism, and deploy the adversarial regularization for generalized and robust learning in low-resource target domains. Extensive experiments over 10 diverse domain pairs demonstrate that the proposed method is superior to state-of-the-art cross-domain NER methods in an end-to-end fashion (about average 6.4% absolute F1 score increase). Further analyses show the effectiveness of modular task decomposition and its great potential in cross-domain NER. Our code and data are available at https://github.com/AIRobotZhang/MTD.

*Corresponding author.

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3531976

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Transfer learning**.

KEYWORDS

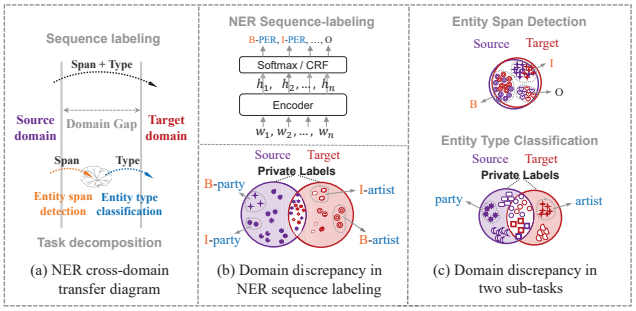
Named Entity Recognition, Cross-domain Transfer, Information Extraction, Knowledge Acquisition

ACM Reference Format:
Xinghua Zhang, Bowen Yu, Yubin Wang, Tingwen Liu, Taoyu Su, and Hongbo Xu. 2022. Exploring Modular Task Decomposition in Cross-domain Named Entity Recognition. In *Proceedings of the 45th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531976

1 INTRODUCTION

Named Entity Recognition (NER) is a major task in information extraction, benefiting web search queries [8, 10, 26, 44], question answering [19, 23] and so forth. NER aims to detect entity spans and classify these spans into predefined categories, such as person, location and organization. Due to expensive labor costs in manual labeling, cross-domain NER has attracted increasing research interests, devoting to transferring knowledge from the source domain to low-resource target ones with only few labeled data.

In order to perform cross-domain transfer, prior competitive studies [4, 12, 45] focus on capturing domain-invariant features under the end-to-end NER sequence-labeling framework by label representation [37] and parameters transfer [13, 22]. They mainly learn both domain-specific and independent features through private and shared domain components based on the monolithic tagging scheme where each token is assigned a compositional label (e.g., B-LOC). However, this monolithic sequence labeling framework is still challenging in cross-domain NER as it not only requires one model to decide the entity span and type simultaneously with a larger label space, but also to transfer two entangled information (entity



(a) NER cross-domain transfer diagram (b) Domain discrepancy in NER sequence labeling (c) Domain discrepancy in two sub-tasks

Figure 1: Task decomposition holds the smaller domain gap in each sub-task than sequence labeling. Arrows in (a) only show the cross-domain transfer diagram instead of training procedure. Two sub-tasks in our framework are parallel.

span and type) across domains, as shown in Figure 1(a). In addition, as shown in Figure 1(b), there are more private labels between the source and target domain under the sequence labeling framework due to the compositional labeling, which causes the great obstacle for cross-domain transfer. Therefore, the acquisition of domain invariance under the monolithic sequence labeling is sub-optimal in cross-domain NER due to complex labeling scheme especially with distinct entity categories across domains. Overall, the standard sequence-labeling paradigm is not ideal in cross-domain NER, as it enlarges the label space and increases the complexity of cross-domain transfer under this end-to-end monolithic learning process.

It is widely known that NER task can be viewed as a combination of two sub-tasks, *entity span detection* and *entity type classification*. Our goal in this paper is to explore modular task decomposition for more effective cross-domain transfer, which as a result simplifies the learning process, maintains the appropriate transfer complexity and needs less annotation effort. Intuitively, the individual sub-tasks are significantly easier to transfer since entity span detection sub-task shares the same label set across domains and entity type classification sub-task owns less private labels between the source and target domain than standard NER sequence labeling, as depicted in Figure 1(c). Thus, each sub-task holds the smaller cross-domain gap and simpler label space across domains compared with monolithic learning process, suggesting that easier domain transfer and less annotations are required for reliable performance.

Methodologically, we use two different encoders to extract their distinct contextual representations from two sub-tasks for separate outcomes, and then combine them to achieve the final result. Specifically, we devise the corresponding cross-domain strategies in each sub-task for transferring entity span and type information separately: (1) *Entity span detection* sub-task which has a common label set across domains, seeks to locate entities in the text. For simplicity, we share all model parameters including the final output layer across domains for transfer. (2) For *entity type classification* sub-task, it determines the entity category for each token. As different domains have distinct pre-defined categories which leads to the obvious domain discrepancy and transfer barrier, we only share encoder parameters but with specific classification head for individual domains. To compensate efficiently for the domain discrepancy in this sub-task, we construct the intermediate augmented domain by a fixed ratio-based mixup on the top of encoder representations between the source and target domain, and then send the intermediate features into a new classification head for minimizing the gap across domains.

NER task decomposition makes the explicit interaction of sub-tasks possible compared with monolithic NER sequence labeling. Two sub-tasks both have the non-entity label "O", and linguistic features are not specific to tasks in low-layer neural network. Thus, we explore these shared information to mutually reinforce two sub-tasks and then propose a modular interaction mechanism including dual-loss re-weighting and linguistic consistency learning. Besides, adversarial regularization is deployed for generalized and robust training in low-resource target domains. The major contributions of this paper are summarized as follows:

- Instead of prior monolithic sequence labeling paradigm, we explore the modular task decomposition in cross-domain NER, which effectively contributes to the transfer with tailor-designed cross-domain strategies in each sub-task (e.g., the shared output layer and the intermediate augmented domain), inspiring a new direction for cross-domain NER.
- The modular interaction mechanism is designed for the mutual reinforcement of sub-tasks, and adversarial regularization guarantees the robust learning on limited labeled data of target domains.
- We evaluate our method on 10 diverse domain pairs and results indicate that modular task decomposition leads to consistent improvements (about average 6.4% absolute F1 score). Further analyses show that our method with 40% target domain data can achieve the comparable performance as the previous SOTAs with 100% data, which confirms the notable superiority of our approach in low-resource scenario.

2 RELATED WORK

We survey related work along two dimensions: (1) cross-domain named entity recognition, (2) task decomposition in NLP and CV. Furthermore, We would like to elaborate their differences from our proposed method.

2.1 Cross-domain Named Entity Recognition

BiLSTM-CRF [15] and BERT [6] based methods become the paradigm in NER due to their promising performances and the end-to-end learning process. They regard NER as the sequence labeling task, where each word in a sentence is assigned a compositional label (e.g., B-LOC, I-LOC). However, most of these works rely on expensive labeling costs. Besides few-shot NER [5, 9, 34] and distantly supervised NER [7, 31], cross-domain NER which can handle the scarcity issue of NER samples in target domains by transfer [22] attracts increasing interests, like the popular cross-domain research of recommendation [3, 18] and so on. Most existing studies are based on the sequence-labeling framework for NER cross-domain transfer. And there are two mainline methods for cross-domain NER: label representation based and parameters transfer methods.

In line of label representation based approaches, Kim et al. [14] used label embeddings as the features to map label types across different domains for cross-domain transfer. Wang et al. [37] proposed a label-aware double transfer learning framework with a variant of

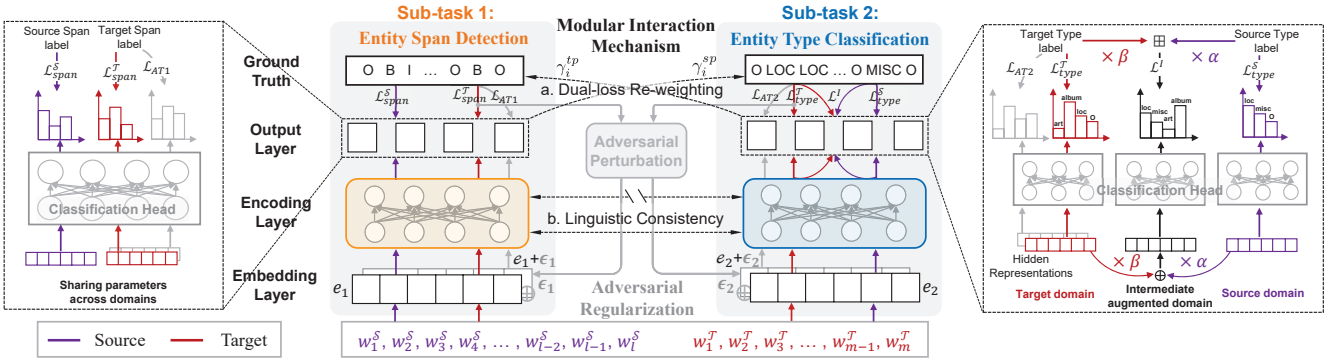


Figure 2: Overview of Modular Task Decomposition. NER task is divided into two sub-tasks using separate functional modules. Shared output layer (*left*) and intermediate augmented domain (*right*) in two sub-tasks respectively are proposed to lessen the gap across domains. Modular interaction mechanism with dual-loss re-weighting and linguistic consistency learning is for mutual reinforcement of sub-tasks. Considering limited labeled data in target domains, adversarial regularization is employed.

maximum mean discrepancy (MMD). Liu et al. [21] introduced a coarse-to-fine two-step pipeline approach based on entity type label description representations, mainly focuses on zero-shot settings and pays less attention to domain discrepancies.

For parameters transfer, some methods [20, 27, 35, 40, 41, 46] focus on learning both domain-independent and specific features through shared and private domain components with knowledge distillation [40] or domain prediction task [46], etc. Jia et al. [12] considered language model task as a bridge for NER domain transfer. Jia and Zhang [13] presented a multi-cell compositional network to model each entity type using separate cell state, learning domain-invariant features in the entity level. Liu et al. [22] released a new dataset *CrossNER*, containing five diverse domain datasets with specialized entity categories for different domains, and proposed competitive baselines for cross-domain transfer. Chen et al. [4] studied cross-domain data augmentation for NER task by domain mapping. They transformed the data representation from high resource to a low-resource domain by learning the text pattern (e.g., style). Zhang et al. [45] proposed a progressive domain adaptation knowledge distillation method for cross-domain NER. They designed the adaptive data augmentation based on data with the same label across domains. And then a multi-grained MMD and knowledge distillation are applied to perform domain adaptation.

2.2 Task Decomposition

In the natural language processing (NLP) and computer vision (CV) communities, decomposing the compositional task into single sub-tasks is a popular paradigm for coping with the *existing issues* of original monolithic task. In *joint entity and relation extraction*, Yu et al. [43] decomposed the joint extraction task into head-entity extraction, and tail-entity and relation extraction to reduce the redundant entity pairs and consider the important inner structure in the process of extracting entities and relations. In standard *named entity recognition*, Shen et al. [32] proposed a two-stage entity identifier which firstly generated span proposals by boundary regression to locate the entities, and then label the span with the corresponding entity categories. Li et al. [16] proposed a boundary-aware bidirectional neural networks by firstly detecting entity spans with pointer network and then performing span classification. Recently, Wang et al. [36] focused on few-shot and zero-shot NER and proposed the SpanNER framework, which firstly detected the entity span and then learned from the natural language descriptions of entity classes. By this decomposed framework, SpanNER can enable the identification of never-seen entity classes with label description. These works mainly decomposed the NER task to tackle existing problems in NER systems, e.g., nested entity, long entity, boundary tag sparsity, lacking of global decoding information. For *multi-task learning*, prior studies decomposed the main task into sub-tasks as auxiliary tasks to enhance the representations for main task. Aguilar et al. [1] proposed to learn NER main task together with named entity segmentation and categorization task, which extracted relevant features to improve the main task. Li et al. [17] introduced a novel interaction network to support information sharing between entity boundary detection and type prediction tasks to enhance the performance of the NER main task. In *object detection*, Perez-Rua et al. [29] decomposed the one-stage CentreNet [47] into class-generic and specific parts for adaptation to the Incremental Few-Shot Detection problem. Xie et al. [39] divided the task into two stages and proposed an oriented region proposal network for reducing the expensive computation during generating proposals.

In this paper, we decompose the monolithic NER task into two sub-tasks for dealing with the great transfer obstacle in prior cross-domain NER methods. Together with our tailor-designed transfer strategies in each sub-task, our modular task decomposition framework can perform more effective transfer and achieve the new SOTA in cross-domain NER. To the best of our knowledge, there is currently no specific research for exploring the task decomposition in cross-domain NER. Our work mainly inspires a new perspective on cross-domain transfer and shows task decomposition is more suitable than monolithic learning process in cross-domain NER, which is completely different from existing work introduced above.

3 METHODOLOGY

In this section, we first introduce the studied problem (Sec. 3.1) and then describe our proposed framework for cross-domain NER.

Figure 2 illustrates the architecture of modular task decomposition. NER task is decomposed into **entity span detection** (Sec. 3.2) and **entity type classification** (Sec. 3.3) sub-task for respective cross-domain transfer by separate functional modules. To mutually reinforce each other for sub-tasks, we develop a **modular interaction mechanism** (Sec. 3.4) where the dual-loss re-weighting and linguistic consistency learning are designed. As the fact of limited labeled data in target domains, **target-domain adversarial regularization** (Sec. 3.5) is applied for smoothing decision boundaries, achieving robust training. Finally, the training optimization and inference procedure (Sec. 3.6) are presented.

3.1 Problem Formulation

We denote $X = w_1, w_2, ..., w_N$ as a sentence, where w_j is the j-th word in sentence X. An entity $\mathbf{e}$ in the sentence X is a span of the sentence: $\mathbf{e} = \{(w_{start}, w_{start+1}, ..., w_{end}), l^e\}$, where $l^e \in C$ is an entity type (category), e.g., *person*, *location*. C is a set of entity type. *Named entity recognition* (NER) aims to find the entity $\mathbf{e}$ in the sentence.

For the cross-domain NER, there are $N_{\mathcal{S}}$ labeled sentences in the source domain $\mathcal{S}$ and $N_{\mathcal{T}}$ labeled sentences in the target domain $\mathcal{T}$. $N_{\mathcal{U}}$ unlabeled sentences in the target domain $\mathcal{T}$ may be available but are not necessary. The sets of entity type (category) in the source and target domain are $C_{\mathcal{S}}$ and $C_{\mathcal{T}}$ respectively. Our goal is to transfer information from the source domain $\mathcal{S}$ to the target domain $\mathcal{T}$. Specifically, we focus on transfer from the high-resource domain to low-resource domain, i.e., $N_{\mathcal{T}} \ll N_{\mathcal{S}}$. And entity categories are different in the source and target domain, i.e., $C_{\mathcal{S}} \neq C_{\mathcal{T}}$. The cross-domain experiment in this paper is more challenging and meets the real-world cross-domain scenario.

3.2 Entity Span Detection Sub-task

This sub-task adopts BERT [6] as backbone and shares the output layer across domains for lessening the domain discrepancy in entity span detection.

3.2.1 Embedding and Encoding Layer. Given an input sentence $X = <w_1, w_2, ..., w_n>$ from the source or target domain, we can extract the specific hidden sequence representations $H_{esd} = <h_1, h_2, ..., h_n> \in \mathbb{R}^{n \times r}$ of all words as:

$$H_{esd} = \text{BERT}(X) \tag{1}$$

The hidden representations can be notated as $H^{\mathcal{S}}_{esd} \in \mathbb{R}^{l \times r}$, $H^{\mathcal{T}}_{esd} \in \mathbb{R}^{m \times r}$ for source (l tokens) and target (m tokens) domains respectively. r is the last hidden layer dimension.

3.2.2 Output Layer. Given the hidden representations of X in the last layer of encoder: $H_{esd} = <h_1, h_2, ..., h_n>$, we use the *Softmax* function to model tagging decisions and then define the entity span distribution for each word w_i:

$$p(\text{span}_k|w_i) = \frac{\exp\{\text{w}_k^\top h_i + b_k\}}{\sum_{j=1}^{c_1} \exp\{\text{w}_j^\top h_i + b_j\}} \tag{2}$$

where $[\text{w}_k; b_k]$ are parameters of classification head specific to the k-th entity span class span_k. Then the probability that w_i belongs to the k-th class is $p(\text{span}_k|w_i)$. As shown in Figure 2 (*left*), $[\text{w}_k; b_k]$ are shared between the source and target domain to bridge the domain gap since the number of classes c_1 in entity span detection sub-task is domain-independent. Here c_1 is set to 3, i.e., $\text{span}_k \in \{\text{B, I, O}\}$. The cross entropy loss is used for training on X:

$$\mathcal{L}_{span} = -\frac{1}{|X|} \sum_{i=1}^{n} \sum_{k=1}^{c_1} y_{i,k} \log(p(\text{span}_k|w_i)) \tag{3}$$

where $y_{i,k}$ is the k-th element in y_i, and y_i is the one-hot span label of w_i. Specifically, the training losses in source and target domains are marked as $\mathcal{L}^{\mathcal{S}}_{span}$ and $\mathcal{L}^{\mathcal{T}}_{span}$ respectively.

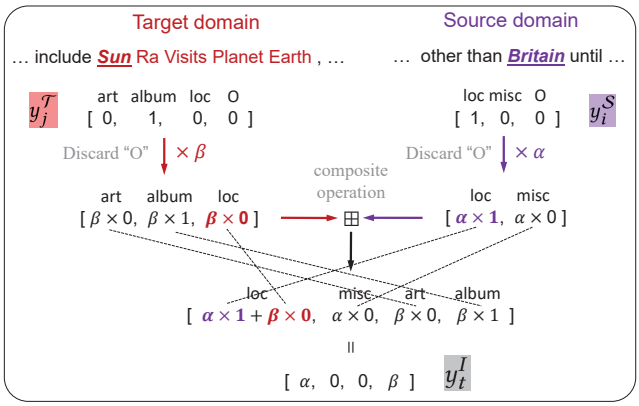


Figure 3: An example of the composite operation ⊞ for label mixup in the intermediate augmented domain.

3.3 Entity Type Classification Sub-task

In this sub-task, another BERT backbone is exploited, and a fixed ratio-based mixup in the output layer is proposed to construct an intermediate augmented domain for handling the obvious discrepancy between the source and target domain in entity type classification.

3.3.1 Embedding and Encoding Layer. Similarly, the last hidden representations in the source and target domain are notated as $H^{\mathcal{S}}_{etc} \in \mathbb{R}^{l \times r}$, $H^{\mathcal{T}}_{etc} \in \mathbb{R}^{m \times r}$ respectively.

3.3.2 Output Layer. Given the hidden representations in the last layer: $H_{etc} = <h_1, h_2, ..., h_n>$, we also use the *Softmax* function to model the entity type distribution for word w_i:

$$p(\text{type}_t|w_i) = \frac{\exp\{\text{w}_t^\top h_i + b_t\}}{\sum_{j=1}^{c_2} \exp\{\text{w}_j^\top h_i + b_j\}} \tag{4}$$

where $[\text{w}_t; b_t]$ are parameters of classification head specific to the t-th entity type class type_t. Then the probability that w_i belongs to the t-th class is $p(\text{type}_t|w_i)$. c_2 is the number of entity categories. As shown in Figure 2 (*right*), $[\text{w}_t; b_t]$ are not shared but specific to the source and target domain because of distinct entity categories across different domains (e.g., $\text{type}_t \in \{\text{location, ..., O}\}$ in the source domain, $\text{type}_t \in \{\text{artist, ..., O}\}$ in the target domain). Then this sub-task training loss on X is as follows:

$$\mathcal{L}_{type} = -\frac{1}{|X|} \sum_{i=1}^{n} \sum_{t=1}^{c_2} y_{i,t} \log(p(\text{type}_t|w_i)) \tag{5}$$

where $y_{i,t}$ is the t-th element in y_i, and y_i is the one-hot type label of w_i. Concretely, the training losses in the source and target domain are $\mathcal{L}^{\mathcal{S}}_{type}$ and $\mathcal{L}^{\mathcal{T}}_{type}$.

To reconcile the gap across domains in this sub-task, we construct the intermediate augmented domain by a fixed ratio-based mixup, inspired by [2, 33]. They utilized the mixup to construct virtual samples for semi-supervised learning. While we propose to use two fixed mixup ratios α and β to fuse the entity information from the source and target domain. Then we can get samples which exist in the intermediate domain between the source and target domain, bridging two domains for cross-domain transfer. It is worth noting that we only fuse entity tokens and non-entity tokens are ignored. Because tokens in the entity carry more domain information. In order to preserve the linguistic correctness of the mixed information, we fuse the entities not in the original text but instead in the hidden representation level. We enumerate every possible entity token pair between the source and target domain in a mini-batch. Given a pair of entity token hidden representations and their corresponding one-hot labels in the source and target domain: $(h^{\mathcal{S}}_i, y^{\mathcal{S}}_i)$ and $(h^{\mathcal{T}}_j, y^{\mathcal{T}}_j)$, our mixup settings are defined as:

$$\begin{aligned} h^I_t &= \alpha \times h^{\mathcal{S}}_i \oplus \beta \times h^{\mathcal{T}}_j \\ y^I_t &= \alpha \times y^{\mathcal{S}}_i \boxplus \beta \times y^{\mathcal{T}}_j \end{aligned} \tag{6}$$

where h^I_t, y^I_t are the fused hidden representation and label in the intermediate domain. $\oplus$ is the vector addition, and $\boxplus$ means a composite operation. Figure 3 shows an example of fusing a token label pair. *Britain* is a location entity from the source domain and *Sun Ra Visits Planet Earth* is an album entity from the target domain. So five token pairs can be formed between the two entities, and we take the token pair (*Britain*, *Sun*) for example. One-hot label $y^{\mathcal{S}}_i$ = [1, 0, 0] (w.r.t., [**location**, misc, O]) and $y^{\mathcal{T}}_j$ = [0, 1, 0, 0] (w.r.t., [artist, album, **location**, O]), non-entity type O is discarded before mixup. For the composite operation $\boxplus$, we define it as: merging values of shared categories between the source and target domain and concatenating the private categories. In figure 3, *location (loc)* is a shared entity category.

Similarly, we get the k-th entity type probability $p(\text{type}_k|h^I_t)$ for the fused representation using Equation 4, but with specific classification head parameters $[\text{w}_t; b_t]$. Because the token label y^I_t after mixup is soft, we use the common soft label loss – *Kullback-Leibler* divergence for training in the intermediate augmented domain as shown in Equation 7, where $y^I_{t,k}$ is the k-th element in y^I_t.

$$\mathcal{L}^I = \sum\nolimits_k y^I_{t,k} \log \frac{y^I_{t,k}}{p(\text{type}_k|h^I_t)} \tag{7}$$

3.4 Modular Interaction Mechanism

Task decomposition makes the explicit interaction between two sub-tasks possible. We propose the modular interaction mechanism which comprises of dual-loss re-weighting and linguistic consistency to reinforce the sub-task correlation.

3.4.1 Dual-loss Re-weighting. In entity type classification sub-task, we should pay more attention to the entity words (tokens) and exploit the entity boundary information from entity span detection sub-task. Thus we use the entity probability of each token w_i in another sub-task to strength the importance of entity words in the loss $\mathcal{L}_{type}$ with γ^{sp}_i as:

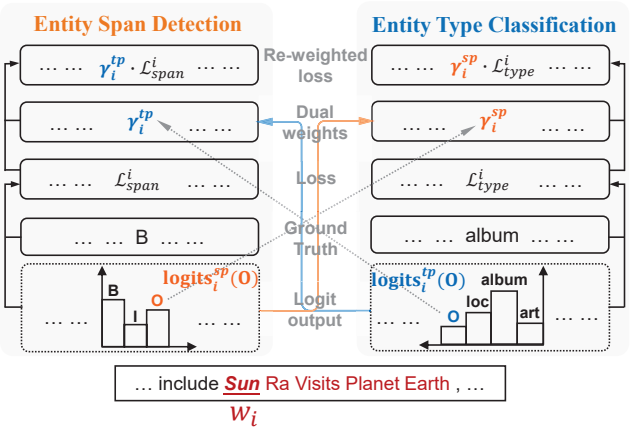


Figure 4: An example of dual-loss re-weighting.

$$\begin{aligned} \mathcal{L}_{type} &= -\frac{1}{|S|}\sum_{i=1}^{n}\gamma^{sp}_i \sum_{t=1}^{c_2} y_{i,t}\log(p(\text{type}_t|w_i)) \\ \gamma^{sp}_i &= 1 - \frac{e^{\text{logits}^{sp}_i(\text{O})/\tau}}{\sum_{k\in\{\text{B,I,O}\}} e^{\text{logits}^{sp}_i(k)/\tau}} + \xi \end{aligned} \tag{8}$$

where ξ is a constant number dominating the weighted magnitude, $\text{logits}^{sp}_i(\text{O})$ is the logit value of class "O" in entity span detection sub-task and τ is a temperature parameter that controls the distribution smoothness. Intuitively, a smaller $\text{logits}^{sp}_i(\text{O})$ means the higher probability of being an entity token, assigns the greater weight γ^{sp}_i on the entity typing loss of w_i. Similarly, for entity span detection sub-task, we expect to recall more entity spans by greater weights on entity words with the heterogeneous information from entity type classification sub-task, thus re-weighting the entity span loss of each token w_i in $\mathcal{L}_{span}$ as follows:

$$\begin{aligned} \mathcal{L}_{span} &= -\frac{1}{|X|}\sum_{i=1}^{n}\gamma^{tp}_i \sum_{k=1}^{c_1} y_{i,k}\log(p(\text{span}_k|w_i)) \\ \gamma^{tp}_i &= 1 - \frac{e^{\text{logits}^{tp}_i(\text{O})/\tau}}{\sum_{k\in\{\text{location},\ldots,\text{O}\}} e^{\text{logits}^{tp}_i(k)/\tau}} + \xi \end{aligned} \tag{9}$$

where $\text{logits}^{tp}_i(\text{O})$ is the logit value of class "O" in entity type classification sub-task. Thus, the smaller non-entity logit value is equivalent to the greater weight γ^{tp}_i on the entity span loss of w_i. Finally, $\mathcal{L}^{\mathcal{S}}_{type}$, $\mathcal{L}^{\mathcal{T}}_{type}$ and $\mathcal{L}^{\mathcal{S}}_{span}$, $\mathcal{L}^{\mathcal{T}}_{span}$ can be modified by instantiating the $\mathcal{L}_{type}$ and $\mathcal{L}_{span}$ in Equation 8, 9 respectively.

Figure 4 gives an example of dual-loss re-weighting for the entity token *Sun*. The losses in two sub-tasks are firstly calculated based on the logit output and ground truth. Then the loss weights in each sub-task are computed with the Equation 8 (9) based on the logit value of non-entity category O from another sub-task. Finally, we can optimize the model based on the re-weighted loss.

3.4.2 Linguistic Consistency. Neural network low-layer features appear not to be specific to a particular task but general [42], such as morphology and syntax in NLP. To learn the linguistic features consistently, we share the low layers between two sub-tasks and make high-level representations specific. Because we use the pre-trained language model BERT as backbone, and the implementation of sharing partial layers between two BERTs is challenging. We approximate it by forcing the two BERTs to express similar features in the low layers with mean squared error loss as:

$$\mathcal{L}_{Sha} = \sum_{k=1}^{L} \text{MSE}(H^k_{span}, H^k_{type}) \tag{10}$$

where L is the number of shared layers from the low to high layer, H^k_{span} and H^k_{type} are the k-th layer hidden representations in two BERTs corresponding to two sub-tasks.

3.5 Target-domain Adversarial Regularization

Due to low-resource target domains, we expect to avoid overfitting and learn the robust model. Here we leverage adversarial training to smooth the decision boundaries, unleashing the full potential of limited data. Concretely, we add the perturbations ϵ to the BERT embeddings e of each token (word) in the target domain. Then the labels of two sub-tasks are employed for regularizing the output under perturbation:

$$\begin{aligned}
\mathcal{L}_{AT1} &= -\frac{1}{|S|}\sum_{i=1}^{m}\sum_{k=1}^{c_1} y_{i,k}\log(p(\text{span}_k|e_{1i}+\epsilon_{1i})) \\
\mathcal{L}_{AT2} &= -\frac{1}{|S|}\sum_{i=1}^{m}\sum_{t=1}^{c_2} y_{i,t}\log(p(\text{type}_t|e_{2i}+\epsilon_{2i}))
\end{aligned} \tag{11}$$

In fact, the perturbation $\epsilon_{1(2)i}$ is in the direction with maximum model output change, which is further defined as:

$$\epsilon_{1(2)i} = \underset{\tilde{\epsilon}_i, ||\tilde{\epsilon}_i||_2 \le \mu}{\arg\max}\ \mathcal{L}_{AT1(2)} \tag{12}$$

Solving the above maximum problem means searching for the worst perturbation while trying to minimize the loss of the model. A general solution for Equation 12 is developed by a linear approximation [11, 25] of adversarial perturbation vector $\epsilon_{1(2)i}$ with a L_2 norm constraint as follows:

$$\epsilon_{1(2)i} \approx \mu \frac{g_{1(2)i}}{||g_{1(2)i}||_2} \tag{13}$$

where $g_{1(2)i} = \nabla_{e_{1(2)i}} \mathcal{L}^{\mathcal{T}}_{span(type)}$ which is efficiently computed by back-propagation. μ is the size of the perturbation. With such a perturbation, $\mathcal{L}_{AT1}$ and $\mathcal{L}_{AT2}$ can be achieved.

3.6 Optimization and Inference

3.6.1 Training Objectives. For each training step, we sample from the source and target domain data, and train the models jointly in the supervised manner by minimizing the total loss:

$$\begin{aligned}
\mathcal{L} = &\mathcal{L}^{\mathcal{S}}_{span} + \mathcal{L}^{\mathcal{T}}_{span} + \mathcal{L}^{\mathcal{S}}_{type} + \mathcal{L}^{\mathcal{T}}_{type} \\
&+ \lambda(\mathcal{L}^{I} + \mathcal{L}_{Sha} + \mathcal{L}_{AT1} + \mathcal{L}_{AT2})
\end{aligned} \tag{14}$$

where λ is the weight coefficient.

Table 1: The statistics of cross-domain NER datasets.

Domain	Dataset	#Train	#Dev	#Test	#Category
Source	CoNLL2003 (Newswire)	14041	–	–	4
	Twitter (Social Media)	4290	–	–	4
Target	Politics	200	541	651	9
	Science	200	450	543	17
	Music	100	380	465	13
	Literature	100	400	416	12
	AI	100	350	431	14

3.6.2 Inference. The predicted entity is a span $<w_i, ..., w_j>$ with the category where entity span detection sub-task provides the entity boundary $\tilde{y}^{span}_s = \arg\max_k \text{logits}^{sp}_s(k), i \le s \le j$, and entity type classification sub-task offers the entity category $\tilde{y}^{type}_j = \arg\max_t \text{logits}^{tp}_j(t)$, $\tilde{y}^{type}_j \neq \text{O}$. That is, span is tagged with the label of rightmost token. Because implementation of using rightmost label is easy. Its performance is also close to the majority label of tokens or leftmost one in our experiment, perhaps due to the encoding ability of BERT. Specifically, the span can not be regarded as an entity when two sub-tasks conflict. Supposing the entity span detection detects a span as an entity but another sub-task classifies it into 'O', then the span is not an entity in our method.

4 EXPERIMENTS

We aim to answer the following research questions: (**RQ1**) Does our proposed framework outperform state-of-the-art methods significantly in cross-domain NER? (**RQ2**) How much can cross-domain transfer in our method gain compared with the monolithic sequence-labeling framework? That is, can our method contribute to the transfer more effectively in cross-domain NER?

4.1 Experimental Settings

4.1.1 Datasets. We conduct experiments on 10 domain pairs which transfer from two source domains to five target domains. The two source domains are CoNLL2003 (English) [30] (a *Newswire* domain dataset) and Twitter dataset [24] (*Social Media* domain). The five target domains are *Politics, Natural Science, Music, Literature* and *Artificial Intelligence* domain datasets released by Liu et al. [22]. The detailed statistics of datasets are shown in Table 1. We can see that two source domains are high-resource and five target domains are low-resource with 100 or 200 training sentences. Because we aim to transfer from the high-resource source domain to low-resource target domain, we ignore dev and test datasets in source domains. The two source domains (*Newswire* and *Social Media*) only have 4 general entity categories (i.e., *person, location, organization and miscellaneous*) which most public NER datasets contain, so source domain datasets are easily available. And target domains own only few labeled data with 9–17 entity categories. Overall, the cross-domain setting of this paper is more applicable in the real world.

Table 2: F1 scores on 10 domain pairs which transfer from two source domains (Newswire, Social Media) to 5 target domains respectively. *Extra Data* **indicates unlabeled target domain data.** *Bold* **marks the highest and** *Blue* **shows the absolute increase compared with prior cross-domain NER baselines.**

Extra Data	Method	CoNLL2003 (Newswire) →					Twitter (Social Media) →				
		Politics	Science	Music	Litera.	AI	Politics	Science	Music	Litera.	AI
No.	BiLSTM-CRF [15]	56.60	49.97	44.79	43.03	43.56	53.64	47.33	48.85	45.23	44.08
	Coach [21]	61.50	52.09	51.66	48.35	45.15	55.03	50.22	49.91	44.88	42.98
	LM-NER [12]	68.44	64.31	63.56	59.59	53.70	66.99	64.23	61.48	59.09	50.46
	BERT-JF [22]	68.85	65.03	67.59	62.57	58.57	67.52	64.51	67.74	61.38	57.05
	BERT-PF [22]	68.71	64.94	68.30	63.63	58.88	68.60	62.23	68.06	61.91	54.72
	MultiCell-LM [13]	70.56	66.42	70.52	66.96	58.28	66.59	63.79	66.54	59.02	53.82
	Style-NER [4]	68.78	63.95	65.43	60.94	58.73	67.33	63.14	67.12	62.06	57.76
	Ours	**76.70**	**72.35**	**76.10**	**69.22**	**68.93**	**74.62**	**71.37**	**74.41**	**69.67**	**64.55**
	Improv.	(+6.14)	(+5.93)	(+5.58)	(+2.26)	(+10.05)	(+6.02)	(+6.86)	(+6.35)	(+7.61)	(+6.79)
Yes. +DAPT [22]	MultiCell-LM [13]	71.45	67.68	74.19	68.63	61.64	69.13	66.76	74.22	64.88	62.41
	Style-NER [4]	71.74	69.11	68.44	62.63	61.76	70.94	68.28	74.40	67.05	63.33
	BERT-JF [22]	72.76	68.28	74.30	65.18	63.07	70.78	67.31	68.13	62.69	59.17
	BERT-PF [22]	72.05	68.78	75.71	69.04	62.56	70.11	66.87	73.88	66.61	61.12
	Ours	**79.52**	**74.82**	**79.80**	**71.15**	**70.41**	**75.49**	**72.81**	**77.43**	**70.14**	**66.18**
	Improv.	(+6.76)	(+5.71)	(+4.09)	(+2.11)	(+7.34)	(+4.55)	(+4.53)	(+3.03)	(+3.09)	(+2.85)

4.1.2 Baselines and Evaluation Metric. We compare the proposed architecture with the following competitive cross-domain NER baselines: (1) **BiLSTM-CRF** [15] combines source domain data and the upsampled target ones to jointly train the model with word and character embeddings. (2) **Coach** [21] performs the coarse-to-fine detection process to handle unseen types, which is a pipeline framework. (3) **LM-NER** [12] integrates the language modeling (LM) and NER tasks in both source and target domains for cross-domain transfer. (4) **MultiCell-LM** [13] investigates a multi-cell compositional LSTM structure on the top of BERT for learning domain-invariant features in the entity level. (5) In Liu et al. [22], **BERT-JF** jointly fine-tunes BERT on both source and target domain data with upsampling in the target domain. **BERT-PF** first pre-trains BERT on the source domain data, and then fine-tunes it to the target ones. (6) **Style-NER** [4] investigates the possibility of leveraging data from high-resource domains by projecting it into the low-resource domains for cross-domain NER. For the completeness of experiments, following Liu et al. [22], before performing cross-domain transfer, we also continue pre-training the BERT on the unlabeled target domain-related corpus (i.e., **DAPT**) for further experiments. The unlabeled datasets are released by Liu et al. [22] with millions of samples. We use the F1 score as the evaluation metric based on exact span matching.

4.1.3 Implementation Details. For fair comparison with competitive baselines (e.g., MultiCell-LM, Style-NER), our method is based on BERT-base [6]. We tune all hyper-parameters according to the results on *dev* sets with grid-search. For each mini-batch, we sample 16 sentences from the source and target domain dataset respectively. The learning rate is 1e-5, maximum training epoches is 30 and the seed of random numbers is set to 0. The fixed mixup ratios (α, β) are set to (0.3, 0.7) by tuning from {(0.1, 0.9), ..., (0.9, 0.1)}. ξ in entity span detection and type classification sub-task is set to 0.5 and 0 respectively. τ is tuned from {0.1, 1.0, 10} in two sub-tasks and finally is set to 0.1 on all datasets except that *Politics* is 10 in entity span detection sub-task. L is set to 3 by tuning from 0 to 12. We tune μ from {0.2, 0.5, 1.0, 2.0} and set 0.5 on *Politics*, *AI* dataset, others are 1.0. λ is set to 0.1. We implement our code with Pytorch based on huggingface Transformers [38][1]. The baseline (except Style-NER) results of the first five domain pairs in Table 2 are all from Liu et al. [22]. For other experimental results, we run the official codes to produce them. Besides partial experimental results in Table 2 and Table 6, other analyses are not using DAPT.

4.2 Experimental Results (RQ1)

Table 2 shows the results of our proposed method compared with baselines and highlights the best F1 score in bold. We conduct two groups of experiments without (No.) or with (Yes. +DAPT) extra unlabeled domain-related corpus. In each group of them, we perform cross-domain transfer experiments from two source domains (CoNLL2003, Twitter) to five target domains respectively. Obviously, our modular task decomposition significantly outperforms the state-of-the-art method with large margins on all experiments.

For the case where no extra unlabeled corpus is used (i.e., only few labeled data in the target domain), we can observe that our proposed method improves the F1 score with an average increase of 6.36% compared with the previous SOTA, where the highest improvement is 10.05% (*CoNLL2003* → *AI*) and the least one also achieves 2.26% absolute percentage points (*CoNLL2003* → *Literature*). It demonstrates the effectiveness and strong generalization of our modular task decomposition for cross-domain NER. In the *AI* domain, the significant increase may result from the larger gap

[1] https://huggingface.co/transformers/

Table 3: Ablation study of our method on dev sets. Scores are averaged over five domains (CoNLL2003 as source domain).

Method	Dev F1
Ours	**74.32**
w/o Shared output layer in *ESD*	73.80
w/o Intermediate augmented domain in *ETC*	73.18
w/o Modular interaction	73.15
– Dual-loss re-weighting	73.56
– Linguistic consistency	73.87
w/o Adversarial regularization	73.89
w/o All of above components	72.15
SpanNER [36]	71.86
BERT-JF [22]	66.55

between *CoNLL2003* and *AI*. And it is difficult for prior studies to learn domain-invariant features only based on end-to-end sequence-labeling scheme. As for the less improvement in the *Literature* domain, the reason is that the entities with *writer* category occupy higher proportion in the dataset, which can be easily confused with *person* category existing in both *CoNLL2003* and *Literature*.

When the extra unlabeled corpus is available, our method also gains significant improvements (4.41% F1 score increase). In addition, it is worth noting that our proposed modular task decomposition without extra data even outperforms other competitive baselines using extra domain-related corpus. For 10 domain pairs, no baselines can always occupy an absolute advantage on them while our method can keep the superiority consistently.

4.2.1 Ablation Study. To evaluate the influence of each component in our method, we conduct the ablation study for further exploration (see Table 3). From these ablations, we can observe that: (1) The shared output layer across domains is effective for domain adaptation in entity span detection sub-task (ESD). As the source and target domain hold the same label space in ESD, the sharing operation can preserve more information from the source domain. (2) The intermediate augmented domain contributes to 1.14% gains of F1 score, we attribute the gains to the bridge role of the constructed domain between the source and target domain in entity type classification (ETC). Because the entity categories from two domains can both be perceived in the intermediate domain. (3) Removing modular interaction mechanism leads to 1.17% declines on F1, which indicates that it is important to enhance the interaction between two sub-tasks by exchanging heterogeneous information and linguistic features with dual-loss re-weighting and linguistic consistency learning. (4) Adversarial regularization in target domains improves the generalization and contributes to 0.43% increase due to decision boundary smoothing. (5) With only two separate sub-tasks (w/o All of above components), the performance decreases by 2.17%, which reflects the comprehensive effect of the designed components. (6) Furthermore, we also run BERT-JF [22] (a sequence labeling framework) on dev sets, which achieves 66.55%. That is, task decomposition can contribute to 5.60% (66.55% to 72.15%) and our tailor-designed modular strategies based on this decomposition paradigm can further improve 2.17% (72.15% to 74.32%). This shows the task decomposition is more suitable as the basic framework in cross-domain NER and verifies our motivation.

Table 4: Avg F1 score over 5 target domains (CoNLL2003 as the source) with different number of parameters. B/s refers to the processed number of batches per second during test.

Method	F1 (Avg)	#Param	Speed
BERT-JF ($\text{BERT}_{\text{BASE}}$)	64.52	108.9M	15 B/s
BERT-JF ($\text{BERT}_{\text{LARGE}}$)	67.88	334.7M	6.8 B/s
MultiCell-LM ($\text{BERT}_{\text{BASE}}$)	66.55	119.5M	2.6 B/s
MultiCell-LM ($\text{BERT}_{\text{LARGE}}$)	67.13	344.7M	2.1 B/s
Ours (shared $\text{BERT}_{\text{BASE}}$)	71.38	109.3M	9.1 B/s
Ours (two $\text{BERT}_{\text{BASE}}$)	**72.66**	216.6M	6.7 B/s

In addition, we also perform experiments on recent low-shot NER method SpanNER [36], which mainly focuses on few-shot and zero-shot learning. SpanNER argues that treating each class as one-hot vector cannot capture the semantic meaning of those labels. Thus, they design the decomposed framework to utilize the label descriptions for the detection of novel entity classes. We adapt SpanNER to our scenario settings with upsampling in the target domain. Experimental result of SpanNER in Table 3 further confirms the effectiveness of task decomposition, consolidating our argument. Lower F1 score of SpanNER compared with our method originates from its pipeline framework and label description quality that cannot be guaranteed.

4.2.2 Parameter Analysis. To check whether our improvements mainly come from more parameters, we show the F1 scores and corresponding parameters of the state-of-the-art baselines and ours in Table 4. We can observe that our method which uses two $\text{BERT}_{\text{BASE}}$ encoders with 216.6M parameters still significantly outperforms the BERT-JF (334.7M) and MultiCell-LM (344.7M) with $\text{BERT}_{\text{LARGE}}$. Furthermore, we share the BERT encoder for two sub-tasks, and the F1 score decreases to 71.38%. The reason may be that sharing encoders completely in two sub-tasks hinders learning subtask-specific information. However, it is still significantly higher than other baselines. Overall, our proposed method does not mainly gain from more parameters but the NER task decomposition with the designed components which leads to easier and more effective cross-domain transfer in each sub-task. In addition, we present the test speed of different methods under the same batch size and experimental environment. We can see that the efficiency of our method is acceptable. In fact, the two sub-tasks in our method can be processed in parallel, which will further accelerate the efficiency.

4.3 Experimental Analyses

*4.3.1 Gains from the Cross-domain Transfer (***RQ2***).* As shown in Table 5, we would like to show the performance gain (↑) from the cross-domain transfer. Therefore, we report the average gain (↑) on five *low-resource* target domains before and after using the source data CoNLL2003. The blue numbers mean only using the target domain data and red ones represent using both the source and target data by cross-domain transfer. We see that our method gains more increase from the source domain data (4.17% absolute F1 score) than

Table 5: Gains (F1 score) of target domain from source domain by transfer. ↑ means the increase after using the source.

Without/With Source domain	CoNLL2003 → five *low-resource* domains (*Avg*)	Twitter → *high-resource* BioMedical
MultiCell-LM	↑2.40 (64.15/66.55)	↑1.42 (78.76/80.18)
BERT-JF	↑2.66 (61.86/64.52)	↑1.55 (79.17/80.72)
Style-NER	↑2.13 (61.44/63.57)	↑2.20 (79.60/81.80)
Ours	↑**4.17** (68.49/72.66)	↑**2.60** (80.83/83.43)

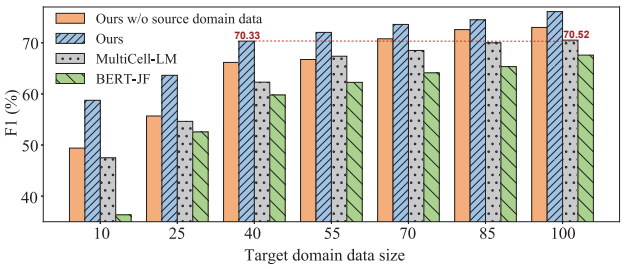


Figure 5: F1 score vs. data size in music domain (CoNLL2003 as source domain, averaged over three samplings).

previous sequence labeling based SOTAs. This shows the notable efficacy of our modular task decomposition in cross-domain transfer. Because NER task decomposition reduces the transfer complexity due to the less private labels across domains and only transferring single entity information (span or type) in each sub-task, contributing to more effective transfer with our devised modular strategies. Interestingly, our method without using the source domain data even significantly surpasses previous SOTAs using source data in Table 5 (Figure 5 also shows this point). That is because task decomposition holds the smaller label space and simpler task form in each sub-task which benefits the *low-resource* NER a lot.

To further explore the above interesting phenomenon, we perform cross-domain transfer in the *high-resource* scenario, where Twitter [24] with 4290 training sentences is the source domain and BioMedical [28] with 3033 training, 1003 dev and 1906 test sentences is the target. We can see that our method still obtains 2.60% gains, and the advantage of our method without using source domain over others with the source data is reasonably reduced due to the rich labeled data in target domain. Overall, we can achieve more effective transfer despite of low and high-resource scenario. Our method reaches the more significant advantage over prior SOTAs in low-resource scenario, benefits from *better performance on low-resource data* and *more effective transfer* which shows the great potential of modular task decomposition in cross-domain NER.

4.3.2 Performance vs. Target Domain Data Size. Figure 5 shows the performance of our approach and two monolithic sequence labeling based methods (e.g., MultiCell-LM and BERT-JF) with different number of target domain data. Style-NER is not shown due to its poor target data generation with only few sentences. We can observe that our proposed method significantly outperforms other baselines, especially when there are few samples in the target

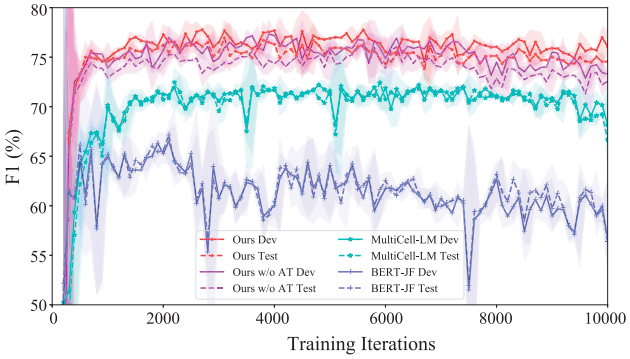


Figure 6: F1 score vs. iterations in the music domain.

domain. We also train our model directly on the target domain data without the source domain training samples. We can see that our method gains more from source domain samples as the target domain data size decreases, which shows the necessity and effectiveness of cross-domain transfer in low-resource target domains. As our model can obtain the abilities of entity span detection and type classification from the source domain which benefits low-resource target domains. Furthermore, our method using 40% target data can achieve the comparable performance as the previous SOTA using the full data, which shows our notable superiority in the extremely low-resource scenario due to smaller spaces and more effective transfer in task decomposition paradigm together with the designed modular constituents.

4.3.3 Learning Curves during Training. Considering only few labeled data in the target domain, we evaluate the robustness and stability of our method during training. Figure 6 shows the F1 score vs. training iterations on the music development and test datasets. Compared with BERT-JF, MultiCell-LM and our method remain more stable. Moreover, our method does not show the obvious overfitting and consistently achieves better performance than other baselines as the training goes. When we remove the target-domain adversarial regularization (w/o AT) from our method, the learning curves on development and test sets are at a low level in comparison with before removal, and tends to decrease as the training goes. Thus, Figure 6 not only confirms the significant improvements of our architecture but also shows the robust training process and powerful generalization ability.

4.3.4 Visualization of Feature Spaces. The hidden representations reflect the effect of cross-domain transfer and determine the final task performances. Figure 7 visualizes the hidden vectors of the monolithic NER sequence-labeling baseline, and our entity span detection (ESD) and type classification (ETC) sub-tasks. Because of smaller and closer label spaces across domains in our task decomposition, especially ESD sub-task, we can see that our two sub-tasks can better capture similar feature distributions across domains, in comparison with the sequence labeling based method which has the complex labeling scheme. Two sub-tasks show the closer representations across domains on the same class and the relatively clear decision boundaries, which indicates the effectiveness of modular task decomposition in cross-domain NER.

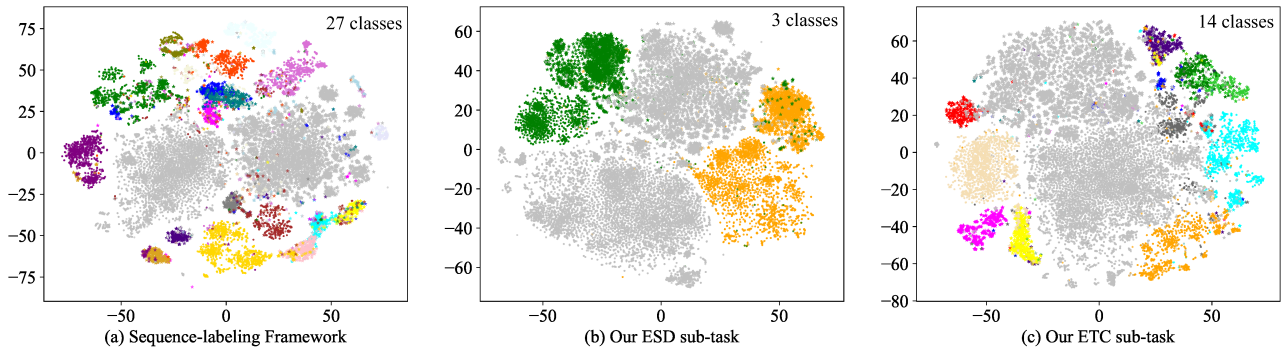


Figure 7: t-SNE visualization of hidden representations on the music dev set. • marks the source domain and ★ marks the target domain. Different classes are represented by different colours. Sequence-labeling framework takes BERT-JF as an example.

Table 6: The respective performance of two sub-tasks (entity span detection, ESD and entity type classification, ETC) without/with extra domain-related data.

No/Yes Extra.	**P**	**R**	**F1**
ESD	88.10/89.13	87.51/87.30	87.80/88.19
ETC	81.22/83.48	78.22/80.56	79.68/81.99
Ours (ESD & ETC)	73.60/76.38	71.78/73.97	72.66/75.14

4.3.5 Error Analysis. Although the proposed method outperforms the state-of-the-art systems, we would like to analyze the factors restricting improvements in task decomposition. Therefore, we show the average performance of each sub-task over five diverse target domains in Table 6 (CoNLL2003 as the source domain). ETC sub-task uses the ground-truth entity spans for evaluation with the predicted entity categories. We observe that the F1 score of ESD sub-task has achieved 87.80% and ETC reaches 79.68%. Thus, the bottleneck lies in ETC sub-task because of distinct label sets across domains and larger label spaces compared with ESD, hindering the cross-domain and few labeled learning. Because of rich domain information for better initialization and fast domain adaptation, using domain-related corpus further improves the F1 scores especially in ETC, but the improvements of all metrics are limited in ESD. The reason may be that the model performance has reached a certain level and then gains little from better initialization with extra data.

5 CONCLUSION AND FUTURE WORK

In this paper, we explore the modular task decomposition to boost the performance of cross-domain NER. Specifically, we decompose the monolithic NER task into two sub-tasks: entity span detection and type classification, and suggest the separate functional modules for respective cross-domain transfer with the shared output layer and intermediate augmented domain. Furthermore, we present a modular interaction mechanism for mutual reinforcement of sub-tasks, and deploy the adversarial regularization to improve the generalization and robustness in low-resource target domains. Experimental results and analyses confirm the effectiveness of our method. For future work, cross-domain transfer with distinct label sets is still worth exploring for improving the entity type classification which is a key sub-task in task decomposition.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their insightful comments and constructive suggestions. This work is supported by the National Key Research and Development Program of China (grant No.2021YFB3100600), the Strategic Priority Research Program of Chinese Academy of Sciences (grant No.XDC02040400) and the Youth Innovation Promotion Association of CAS (grant No.2021153).

REFERENCES

[1] Gustavo Aguilar, Suraj Maharjan, A. Pastor Lopez-Monroy, and Thamar Solorio. 2017. A Multi-task Approach for Named Entity Recognition in Social Media Data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Association for Computational Linguistics, 148–153.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Avital Oliver, Nicolas Papernot, and Colin Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Curran Associates, Inc.

[3] Jiangxia Cao, Jiawei Sheng, Xin Cong, Tingwen Liu, and Bin Wang. 2022. Cross-Domain Recommendation to Cold-Start Users via Variational Information Bottleneck. In *Proceedings of the 38th IEEE International Conference on Data Engineering (ICDE 2022)*.

[4] Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data Augmentation for Cross-Domain Named Entity Recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 5346–5356.

[5] Xin Cong, Shiyao Cui, Bowen Yu, Tingwen Liu, Yubin Wang, and Bin Wang. 2021. Few-Shot Event Detection with Prototypical Amortized Conditional Random Field. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 28–40.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings ofthe 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 4171–4186.

[7] Zheng Fang, Yanan Cao, Tai Li, Ruipeng Jia, Fang Fang, Yanmin Shang, and Yuhai Lu. 2021. TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-aware Network. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 198–207.

[8] Besnik Fetahu, Shervin Malmasi, Anjie Fang, and Oleg Rokhlenko. 2021. Gazetteer Enhanced Named Entity Recognition for Code-Mixed WebQueries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 1677–1681.

[9] Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th*

ACM/SIGAPP Symposium on Applied Computing. Association for Computing Machinery, 993–1000.

[10] 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*. Association for Computing Machinery, 267–274.

[11] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial Personalized Ranking for Recommendation. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 355–364.

[12] 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*. Association for Computational Linguistics, 2464–2474.

[13] 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*. Association for Computational Linguistics, 5906–5917.

[14] 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*. Association for Computational Linguistics, 473–482.

[15] 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*. Association for Computational Linguistics, 260–270.

[16] Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, and Jing Xu. 2021. Effective Named Entity Recognition with Boundary-aware Bidirectional Neural Networks. In *Proceedings of the Web Conference 2021 (WWW'21)*. Association for Computing Machinery, 1695–1703.

[17] Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, Jing Xu, Guoxiu He, and Meihuizi Jia. 2021. Modularized Interaction Network for Named Entity Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 200–209.

[18] Siqing Li, Liuyi Yao, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Tonglei Guo, Bolin Ding, and Ji-Rong Wen. 2021. Debiasing Learning based Cross-domain Recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)*. Association for Computing Machinery, 3190–3199.

[19] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-Relation Extraction as Multi-turn Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1340–1350.

[20] 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*. Association for Computational Linguistics, 2012–2022.

[21] Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020. Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 19–25.

[22] 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*. Association for the Advancement of Artificial Intelligence, 13452–13460.

[23] 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*. Association for Computational Linguistics, 7052–7063.

[24] 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*. Association for Computational Linguistics, 1990–1999.

[25] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations*.

[26] Shekoofeh Mokhtari, Ahmad Mahmoody, Dragomir Yankov, and Ning Xie. 2019. Tagging Address Queries in Maps Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9547–9551.

[27] Hoang Van Nguyen, Francesco Gelli, and Soujanya Poria. 2021. DOZEN: Cross-Domain Zero Shot Named Entity Recognition with Knowledge Graph. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, 1642–1646.

[28] Claire Nédellec, Robert Bossy, Jin-Dong Kim, and et al. 2013. Overview of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*.

[29] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. 2020. Incremental Few-Shot Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13846–13855.

[30] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 142–147.

[31] Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning Named Entity Tagger using Domain-Specific Dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2054–2064.

[32] Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and Label: A Two-stage Identifier for Nested 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*. Association for Computational Linguistics, 2782–2794.

[33] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. Curran Associates, Inc., 596–608.

[34] Meihan Tong, Shuai Wang, Bin Xu, Yixin Cao, Minghui Liu, Lei Hou, and Juanzi Li. 2021. Learning from Miscellaneous Other-Class Words for Few-shot 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*. Association for Computational Linguistics, 6236–6247.

[35] Jing Wang, Mayank Kulkarni, and Daniel Preotiuc-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*. Association for Computational Linguistics, 8476–8488.

[36] Yaqing Wang, Haoda Chu, Chao Zhang, and Jing Gao. 2021. Learning from Language Description: Low-shot Named Entity Recognition via Decomposed Framework. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 1618–1630.

[37] Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, and et al. 2018. Label-aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition. In *Proceedings of NAACL-HLT 2018*. Association for Computational Linguistics, 1–15.

[38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, and et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 38–45.

[39] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. 2021. Oriented R-CNN for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3520–3529.

[40] Huiyun Yang, Shujian Huang, Xin-Yu Dai, and Jiajun Chen. 2019. Fine-grained Knowledge Fusion for Sequence Labeling Domain Adaptation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 4197–4206.

[41] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. In *ICLR 2017*.

[42] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Curran Associates, Inc., 3320–3328.

[43] Bowen Yu, Zhenyu Zhang, Xiaobo Shu, Tingwen Liu, and et al. 2020. Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy. In *24th European Conference on Artificial Intelligence (ECAI)*.

[44] Ningyu Zhang, Qianghuai Jia, Shumin Deng, Xiang Chen, Hongbin Ye, Hui Chen, Huaixiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, and Huajun Chen. 2021. AliCG: Fine-grained and Evolvable Conceptual Graph Construction for Semantic Search at Alibaba. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)*. Association for Computing Machinery, 3895–3905.

[45] Tao Zhang, Congying Xia, Philip S. Yu, Zhiwei Liu, and Shu Zhao. 2021. 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*. Association for Computational Linguistics, 5441–5451.

[46] Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3461–3471.

[47] Xingyi Zhou, Dequan Wang, and Philipp Krahenbuhl. 2019. Objects as points. In *arXiv:1904.07850*.