Representation and Labeling Gap Bridging for Cross-lingual Named Entity Recognition

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

Bowen Yu
DAMO Academy, Alibaba Group
yubowen.ybw@alibaba-inc.com

Jiangxia Cao
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, UCAS
caojiangxia@iie.ac.cn

Quangang Li*
Institute of Information Engineering, Chinese Academy of Sciences

Xuebin Wang
Institute of Information Engineering, Chinese Academy of Sciences

Tingwen Liu
Institute of Information Engineering, Chinese Academy of Sciences

Hongbo Xu
Institute of Information Engineering, Chinese Academy of Sciences

ABSTRACT

Cross-lingual Named Entity Recognition (NER) aims to address the challenge of data scarcity in low-resource languages by leveraging knowledge from high-resource languages. Most current work relies on general multilingual language models to represent text, and then uses classic combined tagging (e.g., B-ORG) to annotate entities; However, this approach neglects the lack of cross-lingual alignment of entity representations in language models, and also ignores the fact that entity spans and types have varying levels of labeling difficulty in terms of transferability. To address these challenges, we propose a novel framework, referred to as DLBri, which addresses the issues of representation and labeling simultaneously. Specifically, the proposed framework utilizes progressive contrastive learning with source-to-target oriented sentence pairs to pre-finetune the language model, resulting in improved cross-lingual entity-aware representations. Additionally, a decomposition-then-combination procedure is proposed, which separately transfers entity span and type, and then combines their information, to reduce the difficulty of cross-lingual entity labeling. Extensive experiments on 13 diverse language pairs confirm the effectiveness of DLBri. The code for this framework is available at https://github.com/AIRobotZhang/DLBri.

CCS CONCEPTS

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

KEYWORDS

Low Resource, Cross-lingual Transfer, Knowledge Acquisition

*Corresponding author. Email: liquangang@iie.ac.cn
† University of Chinese Academy of Sciences

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591757

ACM Reference Format:
Xinghua Zhang, Bowen Yu, Jiangxia Cao, Quangang Li, Xuebin Wang, Tingwen Liu, and Hongbo Xu. 2023. Representation and Labeling Gap Bridging for Cross-lingual Named Entity Recognition. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591757

1 INTRODUCTION

Named Entity Recognition (NER) is one of the important and critical tasks in knowledge acquisition and information retrieval, and has been widely used in web search [9, 10, 20, 28, 56] and so forth. NER is to tag entity spans in text with their corresponding type (e.g., *person*), and has gained considerable improvements with deep learning and abundant labeled data, but this situation brings the rock-ribbed challenge to low-resource languages. Thus, cross-lingual NER was proposed and has attracted great research interest, which transfers information from high-resource language to low-resource ones.

Recent methods for addressing cross-lingual NER typically utilize a framework that involves utilizing a multilingual pre-trained encoder, such as mBERT [6], to encode text in both the source and target languages, and then applying a commonly used combined tagging scheme (e.g., B-ORG) to label each word with its position and type [21, 48, 61]. Additionally, a language adversarial discriminator [4] or cross-lingual similarity loss [21] is employed as an multi-task training objective to reduce the distance between different languages. However, it is our belief that this framework, despite producing good results, does not fully take into account the characteristics of cross-lingual NER in two crucial aspects: 1) **Cross-lingual entity representation**: current methods rely on mBERT to lead off cross-lingual tranfer, but mBERT is a language model trained on multilingual text, whose goal is general language understanding, and is therefore not well suited to understanding and aligning cross-lingual entities. For example, "卡塔尔" in Chinese and "*Qatar*" in English should have similar representations in the encoder to perform cross-lingual labeling, but this is not achieved in a general multilingual model. 2) **Cross-lingual entity**

Table 1: The drops of F1 score (%) on target-language test data finetuning mBERT on source-language training data instead of target-language training data (averaged over each dataset).

Dataset (Source → #Target) \ Subtask	Span Extraction (↓)	Type Prediction (↓)
CoNLL (EN → 3 Western)	5.79	8.10
WikiAnn (EN → 3 Non-western)	24.74	7.81
LOWNER (EN → 7 Non-western)	18.55	9.81

labeling: the combined tag used in current cross-lingual NER extracts entity spans and predicts types simultaneously, which causes entangled entity information. Although entity span and type are closely related, these two kinds of information have different cross-lingual transfer difficulties. For example, as shown in Table 1, the performance of entity span extraction drops significantly more on distant cross-lingual pairs (about 20% decline), while type prediction is less affected by the language discrepancy between the source and target (less than 10% decline). Prior hybrid transfer methods do not take into account these discrepant transfer barriers, which leads to coarse and insufficient language transfer.

This paper develops the **d**ual-**l**evel gap **bri**dging framework **DLBri** from the viewpoints of representation and labeling. To address the first issue, from the **representation point of view**, we propose a progressive contrastive bridging method to pre-finetune mBERT for learning cross-lingual entity-aware representations. This method blends the context and entities of two languages to construct two intermediate language text sequences, one containing the source language context and the target language entities, and the other containing the target language context and the source language entities. By performing contrastive learning on entity words from the source language to the intermediate and then to the target, language discrepancy can be gradually bridged, especially for distant languages. To address the second issue, we decompose the NER task into span extraction and type prediction subtask for respective cross-lingual transfer, and then perform the subtask combination for modeling the relevance of two subtasks from the **labeling point of view**. This decomposition-then-combination procedure reduces the transfer complexity and encourages capturing more fine-grained information for contributing to entity knowledge transfer to a higher degree. Comprehensively, representation-level gap bridging is first conducted for effectively narrowing the language discrepancy, and then labeling-level gap bridging is expected to transfer the entity span and type knowledge separately from labeled source-language data to unlabeled target-language data. These two steps are performed under the semi-supervised learning framework, following the common practice in cross-lingual NER [21, 48, 61].

The major contributions of this paper are summarized as follows:

- We find that existing cross-lingual NER models rely on multilingual language models that only learn general language knowledge, lacking the ability to understand entities across languages. Therefore, we propose a progressive contrastive bridging method to pre-finetune the model, and gradually encourage entity alignment between the source and target languages, thus **learning better cross-lingual entity-aware representations**.
- We first observe and analyze that in cross-lingual NER, entity spans and types have different transfer difficulties, so we propose a decomposition-then-combination procedure, which separately transfers the entity span and type information, **reducing the difficulty of cross-lingual entity labeling**.
- We evaluate our method on 13 diverse language pairs, and experimental results and analyses demonstrate the rationality and effectiveness of our DLBri framework especially on non-western languages (about average 8.82% absolute F1 score increase).

2 RELATED WORK

Cross-lingual NER aims to handle the labeled data scarcity in low-resource languages and can be grouped into *data transfer based* and *model transfer based* methods. *Data transfer based* methods [2, 3, 8, 11, 30, 51] convert high- and low-resource languages into the unified text feature space by mapping [55, 62] or translator [14, 27], reducing the language discrepancy. For example, Ni et al. [30] transformed target-language word embeddings into source-language space based on paired bilingual dictionary. Jain et al. [14] used machine translation system to translate sentences and entities for improving annotation-projection methods. Liu et al. [24] introduced entity placeholders and boundary identifiers to avoid problems such as word order change and entity span determination after translating. *Model transfer based* methods [1, 13, 15, 49] aim to exploit the knowledge from the source-language model by knowledge distillation [4, 22], parameter sharing [16, 60] and so on. Wu et al. [47] proposed a teacher-student learning method where the trained models on the source language serve as teacher to train a student model on unlabeled data in target language. Li et al. [21] introduced the similarity metric model as an auxiliary task based on knowledge distillation and multi-task learning to effectively transfer knowledge across languages. As it should be, some methods [37, 48, 61] are both data and model transfer based.

Decomposing task into multiple subtasks has been widely studied in various fields, such as question answering [32], object detection [33, 36, 52] and information extraction [44, 54, 58]. Ma et al. [25] proposed a decomposed meta-learning method for few-shot NER. To deal with nested entities, some studies [18, 38] firstly detected entity spans and then performed span classification. Liu et al. [23] decomposed the joint extraction into relation extraction and head-tail entity extraction to solve overlapping problems. **Contrastive learning** [17, 39, 46] is commonly used and has achieved promising performance in natural language processing (NLP). Das et al. [5] proposed contrastive learning technique to optimize the inter-token distribution distance for few-shot NER.

Different from prior studies, this paper proposes to bridge cross-lingual gap from representation and labeling level, where we devise progressive contrastive bridging for learning better cross-lingual entity representations, and decompose NER task into *span extraction* and *type prediction* subtask for considering discrepant transferability of entity span and type, studied for the first time in this paper.

3 PRELIMINARIES

3.1 Task Definition

Given a sentence X=<$x_1, x_2, ..., x_n$>, x_i is a word (token) and n is length of the sentence. An entity is a span of X with a category:

$e = \{(x_{start}, x_{start+1}, ..., x_{end}), l^e\}$, where $l^e \in C$ is an entity type (category), e.g., person, location. C is a set of entity types. NER is to detect entity spans and tag corresponding entity types. Following previous cross-lingual NER studies [4, 21, 22, 47, 48, 61], we focus on transferring from high-resource language ($\mathcal{S}$) to low-resource language ($\mathcal{T}$), where there are $N_{\mathcal{S}}$ labeled sentences in source language $\mathcal{S}$ denoted as $\mathcal{D}_{\mathcal{S}}$, and the target language $\mathcal{T}$ only has $N^u_{\mathcal{T}}$ unlabeled sentences denoted as $\mathcal{D}^u_{\mathcal{T}}$. The entity type set in $\mathcal{S}$ and $\mathcal{T}$ is identical. Formally, cross-lingual NER is to learn a model by using labeled $\mathcal{D}_{\mathcal{S}}$ and unlabeled $\mathcal{D}^u_{\mathcal{T}}$ to achieve good performance on target-language test data, which is typically ***a semi-supervised learning problem with cross-lingual gap***. For example, mainstream methods [4, 21, 47] use semi-supervised knowledge distillation [12] to train a student model on unlabeled $\mathcal{D}^u_{\mathcal{T}}$ with pseudo labels provided by the teacher model trained on labeled $\mathcal{D}_{\mathcal{S}}$.

3.2 Semi-supervised Learning (SSL) Framework

As described above, cross-lingual NER is a semi-supervised learning problem in essence. Inspired by [22, 29, 34, 57], we introduce the confidence-based semi-supervised learning framework in our method to exploit labeled source data $\mathcal{D}_{\mathcal{S}}$ and unlabeled target data $\mathcal{D}^u_{\mathcal{T}}$. It consists of training teacher model on $\mathcal{D}_{\mathcal{S}}$, then teaching student models and conducting confidence-based learning on $\mathcal{D}^u_{\mathcal{T}}$.

Source Teacher Distillation. We first train the teacher network on labeled source-language data $\mathcal{D}_{\mathcal{S}}$ via cross entropy loss:

$$\mathcal{L}_{\mathcal{S}} = -\frac{1}{|\mathcal{D}_{\mathcal{S}}|} \sum_{X_i \in \mathcal{D}_{\mathcal{S}}} \sum_{x_j \in X_i} \sum_{k=1}^{|C_{tag}|} y_{j,k} \log(p(tag_k|x_j)) \quad (1)$$

where x_j is a word in sentence X_i. $y_{j,k}$ is the k-th element in y_j, and y_j is the one-hot label of x_j. C_{tag} is a label set of the task.

Based on the trained teacher model, we perform knowledge distillation on unlabeled target data $\mathcal{D}^u_{\mathcal{T}}$. Specifically, the student model is respectively taught by distilling from teacher network using Kullback-Leibler divergence loss:

$$\mathcal{L}_{Distil} = \frac{1}{|\mathcal{D}^u_{\mathcal{T}}|} \sum_{X_i \in \mathcal{D}^u_{\mathcal{T}}} \sum_{x_j \in X_i} D_{KL}(p(tag|x_j)||p_{tea}(tag|x_j)) \quad (2)$$

where $p(tag|x_j)$ and $p_{tea}(tag|x_j)$ are label probability distributions respectively predicted by the student and teacher model on $\mathcal{D}^u_{\mathcal{T}}$.

Confidence-based Learning. To effectively explore the target language data $\mathcal{D}^u_{\mathcal{T}}$, we then propose co-matching process using two student models on $\mathcal{D}^u_{\mathcal{T}}$ together with the above source-language knowledge distillation. Concretely, one student model makes its predictions match the *high-confidence* predictions of its peer model to mutually train the model via cross entropy loss. Thus, the training objective of co-matching procedure for one model is defined as:

$$\mathcal{L}_{CoM} = -\frac{1}{|\mathcal{D}^u_{\mathcal{T}}|} \sum_{X_i \in \mathcal{D}^u_{\mathcal{T}}} \sum_{x_j \in X_i} \mathbb{I}_{>\delta} \sum_{k=1}^{|C_{tag}|} \hat{y}_{j,k} \log(p(tag_k|x_j)) \quad (3)$$

where $\hat{y}_{j,k}$ is the k-th element in $\hat{y}_j$, $\hat{y}_j$ is the pseudo one-hot label predicted by the other model. $\mathbb{I}_{>\delta} \in \{0, 1\}$ and $\mathbb{I}_{>\delta}$ equals 1 when $max(\hat{p}(tag|x_j)) > \delta$. δ is confidence threshold and $\hat{p}(tag|x_j)$ is the probability distributions for word x_j predicted by the other model.

Conversely, we can also utilize the *low-confidence* predictions for training two student models themselves. As high-confidence predictions are more reliable, low-confidence predictions are unlikely to be correct. Therefore, each model learns by penalizing itself using the low-confidence predictions. That is the most confident label predicted by the model with probability value lower than δ should be suppressed:

$$\mathcal{L}_{P} = -\frac{1}{|\mathcal{D}^u_{\mathcal{T}}|} \sum_{X_i \in \mathcal{D}^u_{\mathcal{T}}} \sum_{x_j \in X_i} \mathbb{I}_{<\delta} \sum_{k=1}^{|C_{tag}|} \bar{y}_{j,k} \log(1 - p(tag_k|x_j)) \quad (4)$$

where $\mathbb{I}_{<\delta} = 1$ when $max(p(tag|x_j)) < \delta$, otherwise is 0. $\bar{y}_{j,k}$ is predicted by the student model itself.

4 METHODOLOGY

Figure 1 depicts overview of dual-level gap bridging framework **DLBri**. First, (*Step* 1) progressive contrastive bridging is proposed to learn cross-lingual entity-aware representations for reducing language discrepancy. Then, (*Step* 2) we perform NER task decomposition-then-combination procedure to transfer knowledge more effectively in subtasks and then obtain the relevance between subtasks.

4.1 Progressive Contrastive Bridging

To better understand and align cross-lingual entities instead of directly applying multilingual BERT (mBERT) for transfer, we build the *source-to-target oriented sentence pairs* to pre-finetune mBERT with *contrastive bridging loss* for gradually narrowing the language discrepancy and providing strong support for subsequent transfer.

4.1.1 Source-to-Target Oriented Sentence Pair. To effectively align cross-lingual entities and make up the language difference, we construct the intermediate languages for forming the source-to-target oriented sentence pairs. Concretely, we assemble two kinds of intermediate language text, including *source oriented intermediate* and *target oriented intermediate* sentence. Source oriented intermediate sentence is equipped with source-language (e.g., English) context and target-language (e.g., Chinese) entities, while target oriented intermediate one has the inverse composition. These two intermediate-language texts are all transformed from source-language training data using Google translation system (https://cloud.google.com/translate). However, NER is a token-level task which faces the matching challenge between the word and label after translation. Following Liu et al. [24], we mark each entity in the sentence with a special symbol (e.g., "[]") before feeding it into the translation system for locating the entity in the translated text.

As shown in top right corner of Figure 1, the sentence "*The* 2022 *World Cup was held in Qatar*" in source language is converted into source oriented intermediate text "*The* 2022 *World Cup was held in* 卡塔尔" and target oriented intermediate text "2022 年世界杯在 *Qatar* 举行". The source-language sentence is also translated into target-language text "2022 年世界杯在卡塔尔举行". Then we get three source-to-target oriented pairs with these four sentences: source language and source oriented intermediate sentences, source oriented intermediate and target oriented intermediate sentences, and target oriented intermediate and target language sentences, which gradually incline from source to target language style. For simplicity, we assume above sentence contains one entity (*Qatar*).

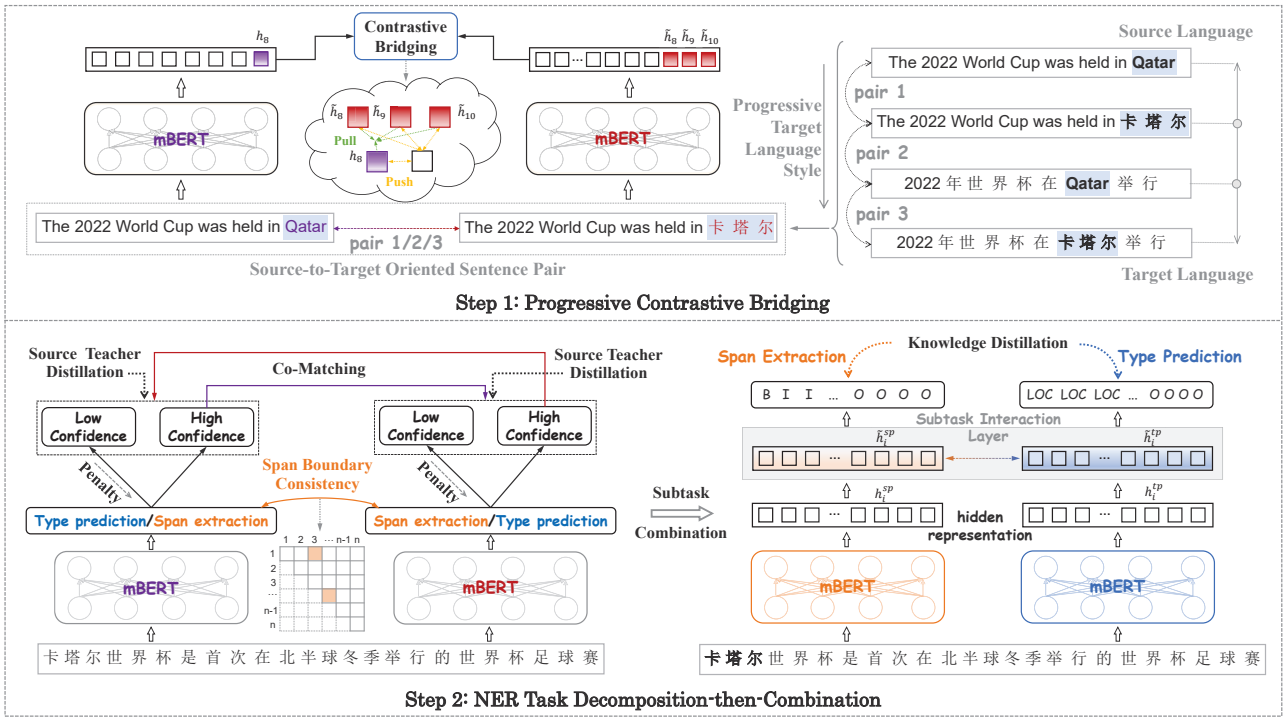

Figure 1: Overview of DLBri which comprises of: *Progressive Contrastive Bridging* for better cross-lingual entity-aware representations and reducing language discrepancy, *NER Task Decomposition-then-Combination* for more effective transfer.

4.1.2 Contrastive Bridging. Figure 1 upper part depicts the contrastive learning procedure. The two sentences in each source-to-target oriented pair are respectively encoded by multilingual BERT [6]. Specifically, two sentences $X = <x_1, x_2, ..., x_n>$ and $\tilde{X} = <\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_m>$ are separately input into encoder to extract the hidden representations of all words $\mathbf{H} = <\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n> \in \mathbb{R}^{n\times d}$ and $\tilde{\mathbf{H}} = <\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, ..., \tilde{\mathbf{h}}_m> \in \mathbb{R}^{m\times d}$ as:

$$\mathbf{H} = \text{mBERT}(X), \quad \tilde{\mathbf{H}} = \text{mBERT}(\tilde{X}) \tag{5}$$

where d is the last hidden layer dimension.

For each entity e in one sentence of source-to-target oriented pair, we can get the corresponding entity $\tilde{e}$ in another sentence, e.g., "*Qatar*" and "卡塔尔" in top left corner of Figure 1. Then the positive examples of each token x_i in entity e are tokens $\{\tilde{x}_o, ..., \tilde{x}_l\}$ of its corresponding entity $\tilde{e}$, while the other words in another sentence can form negative pairs with x_i. Let $sim(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$ denote the dot product between ℓ_2 normalized $\boldsymbol{u}$ and $\boldsymbol{v}$, which is the cosine similarity for measuring the positive and negative pairs. The contrastive loss function for token x_i in entity e is defined as:

$$\mathcal{L}_i^{cts} = -\frac{1}{|\tilde{X}|}\sum_{j=1}^{m} \mathbb{I}_{\tilde{x}_j\in\tilde{e}} \log \frac{\exp(sim(\mathbf{h}_i, \tilde{\mathbf{h}}_j)/\tau)}{\sum_{k=1}^{m}\exp(sim(\mathbf{h}_i, \tilde{\mathbf{h}}_k)/\tau)} \tag{6}$$

where τ is a scalar temperature hyperparameter. $\mathbb{I} \in \{0, 1\}$ is indicator function, $\mathbb{I}_{\tilde{x}_j\in\tilde{e}} = 1$ when word $\tilde{x}_j$ is a token of entity $\tilde{e}$. In this way, same entities in different languages stay close to each other while dissimilar ones are far apart. It is worth noting that positive and negative samples of each token x_i in one sentence are all from another sentence of the source-to-target oriented pair. For instance, in top left of Figure 1, token *Qatar* forms positive pairs with tokens {卡, 塔, 尔} in another sentence, while the remainder tokens are negative samples. For three sentence pairs in Figure 1, we continually train the model on these pairs with contrastive loss.

4.2 Task Decomposition-then-Combination

With *progressive contrastive bridging* above, we can get the bridged model which effectively adapts to the language discrepancy between source and target language. To perceive different transferability of entity span and type information, we decompose NER task into *span extraction* and *type prediction* subtask, and perform respective cross-lingual transfer under semi-supervised learning framework (Sec. 3.2) in each subtask. Afterwards, *subtask combination* is devised to build correlation of subtasks for further improvement.

4.2.1 Decomposing NER Task into Subtasks for Transfer. We decompose NER task into span extraction and type prediction subtask for respective cross-lingual transfer in each subtask.

Span Extraction Subtask. This subtask aims to extract entity spans in text with label set $C_{sp} = \{\text{B, I, O}\}$. Given a sentence $X = <x_1, x_2, ..., x_n>$, its hidden sequence representations are $\mathbf{H} = <\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n>$ and entity span distribution for each word x_i can be got:

$$\mathbf{H} = \text{Encoder}_{sp}(X)$$
$$p(C_{sp}^k|x_i) = \frac{\exp\{\mathbf{w}_k^\top \mathbf{h}_i + b_k\}}{\sum_{j=1}^{|C_{sp}|}\exp\{\mathbf{w}_j^\top \mathbf{h}_i + b_j\}} \tag{7}$$

where $[\mathbf{w}_k; b_k]$ are classification head corresponding to k-th entity span class C^k_{sp}. $p(C^k_{sp}|x_i)$ is the probability that x_i belongs to k-th class. B, I, O indicate the **b**eginning, **i**nside of entity and non entity.

To transfer knowledge from source language to the target, we introduce the semi-supervised learning framework described in section preliminaries (Sec. 3.2), following the common practice in cross-lingual NER [21, 22, 48, 61]. From Eq. 1 to 4, the *tag* is span label and C_{tag} is C_{sp}, and all label probability distributions are obtained using Eq. 7. Encoder$_{sp}$ of two student models and their corresponding teachers are all respectively initialized by the two bridged models trained in Sec. 4.1. Based on the semi-supervised learning, entity span knowledge is transferred from labeled source-language data to the unlabeled target-language data.

Span Boundary Consistency. In comparison with type prediction subtask, span extraction needs more label dependency and explicit boundary modeling, and its cross-lingual difficulty is greater as shown in Table 1. Besides extracting entity spans based on token-wise sequence labeling, we expect that the span extraction model captures the explicit boundary information and more bridging features for stimulating cross-lingual transfer. Specifically, we calculate the attentive correlation on each student model of semi-supervised learning framework for a token pair (x_i, x_j) following the idea of self-attention [43]:

$$\mathbf{q}_i = \mathbf{W}_q\mathbf{h}_i + b_q, \quad \mathbf{k}_i = \mathbf{W}_k\mathbf{h}_i + b_k \tag{8}$$

$$Score_{bd}(i,j) = \mathbf{q}_i^\top \mathbf{k}_j + \mathbf{W}_{bd}(\mathbf{h}_i + \mathbf{h}_j) \tag{9}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_{bd}$, b_q and b_k are trainable parameters. $\mathbf{h}_i$ is hidden representation for token x_i. Then we compute boundary scores for two student models of Sec. 3.2 as $Score^1_{bd}(i,j)$ and $Score^2_{bd}(i,j)$. The span boundary consistency loss is finally defined as:

$$\mathcal{L}^{\mathcal{T}}_{SBC} = \frac{1}{|\mathcal{D}^u_{\mathcal{T}}|} \sum_{X_l \in \mathcal{D}^u_{\mathcal{T}}} \sum_{i=1}^{|X_l|} \sum_{j=i}^{|X_l|} (Score^1_{bd}(i,j) - Score^2_{bd}(i,j)) \tag{10}$$

where X_l is a sentence in $\mathcal{D}^u_{\mathcal{T}}$ and we only calculate the consistency on the upper triangular matrix as shown in the bottom left of Figure 1. To enhance the language transfer, we also compute the span boundary consistency loss on source data $\mathcal{D}_{\mathcal{S}}$ by sharing all consistency parameters with target data $\mathcal{D}^u_{\mathcal{T}}$, notated as $\mathcal{L}^{\mathcal{S}}_{SBC}$.

Type Prediction Subtask. The aim of this subtask is to predict the entity category in the label set C_{tp} = {PER, LOC, ..., O} for each word of the sentence. Similarly, the hidden representations and entity type distribution for word x_i can be obtained:

$$\mathbf{H} = \text{Encoder}_{tp}(X)$$

$$p(C^t_{tp}|x_i) = \frac{\exp\{\mathbf{w}_t^\top \mathbf{h}_i + b_t\}}{\sum_{j=1}^{|C_{tp}|} \exp\{\mathbf{w}_j^\top \mathbf{h}_i + b_j\}} \tag{11}$$

where $[\mathbf{w}_t; b_t]$ are classification head parameters corresponding to the t-th entity type class C^t_{tp}. $p(C^t_{tp}|x_i)$ is the probability that x_i belongs to the t-th class.

Similar to span extraction subtask, we transfer entity type knowledge from labeled source-language data to unlabeled target-language data under the semi-supervised framework (Sec. 3.2), where the *tag* is type label and C_{tag} is C_{tp} in Eq. 1 to 4, and all label probability distributions are obtained using Eq. 11. The parameter initialization of Encoder$_{tp}$ is also from Sec. 4.1, like in span extraction subtask.

4.2.2 Subtask Combination. Despite of more effective transfer in subtask, two subtasks can actually be enhanced by each other for further improvement. For instance, all words of an entity span in span extraction should have the identical entity type in type prediction subtask. Therefore, as shown in Figure 1, we devise the *subtask interaction layer* together with *distilling the knowledge* from span extraction and type prediction models trained in subtask transfer (Sec. 4.2.1).

Subtask Interaction Layer. The subtask interaction aims to exchange label information between two subtasks. Inspired by [19, 35, 53], we conduct label attention over entity span and type label to obtain the span label and type label related representations for each word, and then these two related features are interacted by co-interactive attention for the exchange of subtask information. Concretely, we utilize parameters $\mathbf{w}$ of the fully-connected span label and type label classification layer in Eq.7, 11 as label representations, then the span label embedding matrix is notated as $\mathbf{W}_{sp} \in \mathbb{R}^{|C_{sp}|\times d}$ and type label matrix is $\mathbf{W}_{tp} \in \mathbb{R}^{|C_{tp}|\times d}$.

To get label related representations, we use an attention mechanism over each subtask labels to obtain explicit label features for each word. The label related representations are computed by:

$$\mathbf{A} = \text{softmax}(\mathbf{H}\mathbf{W}_c^\top), \quad \tilde{\mathbf{H}}_c = \mathbf{H} + \mathbf{A}\mathbf{W}_c \tag{12}$$

where $\mathbf{H} \in \mathbb{R}^{n\times d}$ is hidden sequence representations in each subtask. $\mathbf{W}_c \in \{\mathbf{W}_{sp}, \mathbf{W}_{tp}\}$, and then we can calculate the corresponding $\tilde{\mathbf{H}}_{sp} = <\tilde{\mathbf{h}}^{sp}_1, \tilde{\mathbf{h}}^{sp}_2, ..., \tilde{\mathbf{h}}^{sp}_n>$, $\tilde{\mathbf{H}}_{tp} = <\tilde{\mathbf{h}}^{tp}_1, \tilde{\mathbf{h}}^{tp}_2, ..., \tilde{\mathbf{h}}^{tp}_n>$. Thus, for the span label related feature $\tilde{\mathbf{h}}^{sp}_i$ of word x_i, we use the biaffine mechanism [7] as co-interactive attention to obtain the span correlation score with each type label related feature $\tilde{\mathbf{h}}^{tp}_j$:

$$\begin{aligned} r^{sp}_{i,j} &= \tilde{\mathbf{h}}^{sp\top}_i \mathbf{U}_{bia} \tilde{\mathbf{h}}^{tp}_j + \mathbf{W}_{bia}(\tilde{\mathbf{h}}^{sp}_i \oplus \tilde{\mathbf{h}}^{tp}_j) + b_{bia} \\ \alpha^{sp}_{i,j} &= \text{softmax}(r^{sp}_{i,j}), \quad j \in [1, n] \end{aligned} \tag{13}$$

where $\tilde{\mathbf{h}}^{sp\top}_i$ is the transposition of $\tilde{\mathbf{h}}^{sp}_i$, $\mathbf{U}_{bia}$ is the weight matrix of bi-linear term and $\mathbf{W}_{bia}$ is the weight matrix of linear term. $\oplus$ is the vector concatenation. Similarly, the type correlation score with each span label related feature is denoted as $\alpha^{tp}_{i,j}$. Ultimately, based on the co-interactive correlation score , we can attain the hidden representations of word x_i after interaction in each subtask:

$$\begin{aligned} \hat{\mathbf{h}}^{st}_i &= \sum_{j=1}^{n} \alpha^{sp}_{i,j} \tilde{\mathbf{h}}^{tp}_j, \quad \hat{\mathbf{h}}^{ts}_i = \sum_{j=1}^{n} \alpha^{tp}_{i,j} \tilde{\mathbf{h}}^{sp}_j \\ \bar{\mathbf{h}}^{sp}_i &= \tilde{\mathbf{h}}^{sp}_i \oplus \hat{\mathbf{h}}^{st}_i, \quad \bar{\mathbf{h}}^{tp}_i = \tilde{\mathbf{h}}^{tp}_i \oplus \hat{\mathbf{h}}^{ts}_i \end{aligned} \tag{14}$$

Knowledge Distillation. Based on the interacted representation $\bar{\mathbf{h}}^{sp}_i$ and $\bar{\mathbf{h}}^{tp}_i$ in two subtasks, we can get the span label and type label probability distributions $\bar{p}^{sp}_i$ and $\bar{p}^{tp}_i$ via fully-connected classifier as in Eq.7 and Eq.11. And then we distil knowledge from subtask transfer (Sec. 4.2.1) by calculating with Kullback-Leibler divergence:

$$\mathcal{L}^i_{KD} = \sum_{c\in\{sp,tp\}} D_{KL}(\bar{p}^c_i || p^{c,1}_i) + D_{KL}(\bar{p}^c_i || p^{c,2}_i) \tag{15}$$

where $p^{c,1}_i$ and $p^{c,2}_i$ ($c \in \{sp, tp\}$) are label probability distributions respectively predicted by two student models trained in semi-supervised framework (introduced in Sec. 3.2, used in Sec. 4.2.1) in each subtask. $\bar{p}^c_i$ is $\bar{p}^{sp}_i$ or $\bar{p}^{tp}_i$.

Algorithm 1 Training procedure of our method

Input: labeled source language $\mathcal{D}_{\mathcal{S}}$ and unlabeled target data $\mathcal{D}^{u}_{\mathcal{T}}$.
Output: span extraction model Θ_{sp} and type prediction Θ_{tp}.

▷ **Step 1: Progressive Contrastive Bridging**
Optimize two bridging models Θ^{1}_{bri}, Θ^{2}_{bri} using $\mathcal{L}^{cts}_{i}$ (Eq. 6).
▷ **Step 2: NER Task Decomposition-then-Combination**
For ***span extraction*** subtask:
  Train $\Theta^{sp,1}_{\mathcal{S}}$, $\Theta^{sp,2}_{\mathcal{S}}$ by respectively fine-tuning Θ^{1}_{bri}, Θ^{2}_{bri} on $\mathcal{D}_{\mathcal{S}}$ using $\mathcal{L}_{\mathcal{S}}$ (Eq. 1).
  Train $\Theta^{sp,1}_{\mathcal{T}}$, $\Theta^{sp,2}_{\mathcal{T}}$ by respectively fine-tuning Θ^{1}_{bri}, Θ^{2}_{bri} using $\mathcal{L}_{Distil}$+$\mathcal{L}_{CoM}$+$\mathcal{L}_{P}$ on $\mathcal{D}^{u}_{\mathcal{T}}$ and $\mathcal{L}^{\mathcal{S}}_{SBC}$+$\mathcal{L}^{\mathcal{T}}_{SBC}$ on $\mathcal{D}_{\mathcal{S}}$, $\mathcal{D}^{u}_{\mathcal{T}}$.
For ***type prediction*** subtask:
  Train $\Theta^{tp,1}_{\mathcal{S}}$, $\Theta^{tp,2}_{\mathcal{S}}$ by respectively fine-tuning Θ^{1}_{bri}, Θ^{2}_{bri} on $\mathcal{D}_{\mathcal{S}}$ using $\mathcal{L}_{\mathcal{S}}$.
  Train $\Theta^{tp,1}_{\mathcal{T}}$, $\Theta^{tp,2}_{\mathcal{T}}$ by respectively fine-tuning Θ^{1}_{bri}, Θ^{2}_{bri} using $\mathcal{L}_{Distil}$+$\mathcal{L}_{CoM}$+$\mathcal{L}_{P}$ on $\mathcal{D}^{u}_{\mathcal{T}}$.
For ***subtask combination***:
  Jointly optimize two models Θ_{sp}, Θ_{tp} on $\mathcal{D}^{u}_{\mathcal{T}}$ using $\mathcal{L}^{i}_{KD}$.
return Θ_{sp}, Θ_{tp}.

4.3 Training and Inference

4.3.1 Training Process. As shown in Figure 1, DLBri is divided into *Progressive Contrastive Bridging* and *NER Task Decomposition-then-Combination*, and *Algorithm* 1 gives the detailed training procedure.

4.3.2 Inference. We finally use the trained models Θ_{sp}, Θ_{tp} in subtask combination (Sec. 4.2.2) for inference. For each word x_i in the sentence X, span extraction model Θ_{sp} predicts a label in C_{sp} = {B, I, O}, and then we can get a set of entities $\{e_1, e_2, ...\}$ of X. As for the type label l^{e}_{i} of each entity e_i, we use the label (belonging to C_{tp} = {PER, LOC, ..., O}) of rightmost word in e_i predicted by type prediction model Θ_{tp}. So far, we can get the entity span and its type. Specifically, e_i is not regarded as an entity if its type label is predicted as O.

5 EXPERIMENTS

In this section, we verify the following research questions:

- **RQ1**: Does our DLBri achieve the significant performances?
- **RQ2**: Does DLBri effectively narrow the language gap?
- **RQ3**: How is the effect of intermediate language text?
- **RQ4**: Is the cross-lingual transfer more advantageous in two subtasks than in NER task with a combined tagging (e.g., B-LOC)?

5.1 Datasets and Evaluation

Following previous cross-lingual NER works, we evaluate our DLBri on CoNLL (i.e., CoNLL2002 [40], CoNLL2003 [41]) and WikiAnn [31] benchmark datasets. CoNLL is comprised of English (EN), German, Spanish and Dutch languages with four entity types {PER, LOC, ORG, MISC}, and WikiAnn contains English and three non-western languages: Arabic, Hindi and Chinese (ZH). For fair comparison with previous methods, we also use English as source language and the other languages as the target language in CoNLL and WikiAnn. In cross-lingual NER, the labeled source-language training set and unlabeled target-language training data are utilized, and we evaluate the model on target-language test set in each language pair and use the exact span matching based F1 score as evaluation metric following prior studies. To further explore the effectiveness under scenarios of large language discrepancy, we use LOWNER [26] dataset including English and seven non-western languages, annotated with six entity types {PER, LOC, CORP, GRP, PROD, CW}. English serves as source language and the other languages are target. Other cross-lingual settings are identical to CoNLL and WikiAnn. Detailed statistics of datasets and explanation of entity types (e.g., CW) can be found in corresponding references.

5.2 Experimental Settings

5.2.1 Baselines. We compare our method with the following strong cross-lingual NER baselines: (1) **Wiki** [42] introduced a language independent method by building on cross-lingual wikification. (2) **WS** [30] presented weakly supervised methods for cross-lingual NER. (3) **TMP** [14] improved annotation projection by leveraging machine translation in cross-lingual NER. (4) **BERT-f** [50] directly applied mBERT to cross-lingual NER. (5) **AdvCE** [16] improved upon mBERT's cross-lingual performance with language-adversarial training. (6) **TSL** [47] proposed a teacher-student network for cross-lingual NER. (7) **Unitrans** [48] unified both model and data transfer for cross-lingual NER by leveraging unlabeled target data. (8) **AdvPicker** [4] designed an adversarial learning framework to select less language-dependent target-language data for training. (9) **RIKD** [22] developed the reinforced iterative knowledge distillation method to explore unlabeled target-language data. (10) **TOF** [59] proposed to transfer knowledge from three aspects: domain, language and task in cross-lingual NER. (11) **MTMT** [21] introduced the similarity metric model as an auxiliary task to improve the cross-lingual NER performance. (12) **ConNER** [61] proposed a consistency training framework which contained translation-based and dropout-based consistency training respectively on unlabeled target-language and labeled source-language data.

5.2.2 Implementation Details. Following prior competitive baselines, we use multilingual BERT-base [6] (mBERT) as encoders. All hyper-parameters are tuned according to the results on dev set with grid-search. For *Step* 1, batch size is 32 and maximum training epoch is 3. For *Step* 2, batch size is 16 and maximum training epoch is 50. The learning rate is 1e-5 and random number seed is 0 in all steps of training procedure. The temperature τ is set to 0.1 by tuning from {10, 1, 0.1, 0.15}. For confidence threshold δ, we use an adaptive threshold instead of a fixed one by dynamically calculating the mean and standard deviation of predicted maximum probability value within a mini-batch, because the model confidence is changed continuously during training. The hidden dimension d is 768. Our method is implemented with Pytorch based on huggingface Transformers [45], and is conducted on NVIDIA Tesla T4 GPU. The baseline results on CoNLL and WikiAnn datasets are all reported by [21, 61]. For baseline results on LOWNER dataset, we run the open-source official codes of baselines.

5.3 Experimental Results (RQ1)

The main results on test sets are reported in Table 2 and Table 3, which highlight the best F1 score in bold and second highest in

Table 2: F1 scores on CoNLL, transferring from English (EN) to three western languages respectively.

Method	German	Spanish	Dutch	Average
Wiki [42]	48.12	60.55	61.56	56.74
WS [30]	58.50	65.10	65.40	63.00
TMP [14]	61.50	73.50	69.90	68.30
BERT-f [50]	69.56	74.96	77.57	74.03
AdvCE [16]	71.90	74.30	77.60	74.60
TSL [47]	73.16	76.75	80.44	76.78
Unitrans [48]	74.82	79.31	82.90	79.01
AdvPicker [4]	75.01	79.00	82.90	78.97
RIKD [22]	76.08	79.78	82.96	79.61
TOF [59]	76.57	80.35	82.79	79.90
MTMT [21]	76.80	81.82	83.41	80.68
ConNER [61]	77.14	80.50	83.23	80.29
DLBri (Ours)	**78.43**	**83.19**	**84.43**	**82.02**

Table 3: F1 scores on WikiAnn, transferring from English (EN) to three non-western languages respectively. † notes we run the official code for producing the results.

Method	Arabic	Hindi	Chinese (ZH)	Average
BERT-f [50]	42.30	67.60	52.90	54.27
TSL [47]	43.12	69.54	48.12	53.59
Unitrans [48]†	44.57	70.84	53.10	56.17
AdvPicker [4]†	49.16	73.26	53.32	58.58
RIKD [22]	45.96	70.28	50.40	55.55
MTMT [21]	52.77	70.76	52.26	58.60
ConNER [61]	59.62	74.49	39.17	57.76
DLBri (Ours)	**65.07**	**77.51**	**59.67**	**67.42**

underline. Table 2 shows the performances of transferring from English to three western languages (German, Spanish, Dutch), while Table 3 shows F1 scores on three non-western languages (Arabic, Hindi, Chinese). We also report the average F1 score over three western and non-western target languages respectively.

For western languages in Table 2, we observe that our method consistently improves the previous state-of-the-art (SOTA) methods, because DLBri learns better cross-lingual entity-aware representations and performs more effective transfer in two subtasks, significantly bridging the cross-lingual gap from the representation and labeling level. Compared with remarkable baselines (e.g., Unitrans [48], ConNER [61]) which also use the translation but ignore the cross-lingual entity alignment and distinct transferability of entity span and type, DLBri obtains the obvious superiority on all language pairs. In Table 3, our method achieves an average absolute increase of 8.82% compared to prior SOTA when transferring from western English to non-western languages, which shows our significant advantages of cross-lingual transfer between distant languages because of more effective gap bridging methods in DLBri. Table 4 reaches the same conclusion on seven distant transfer pairs, which powerfully shows the language transfer ability of DLBri. Comparing Table 2 with Table 3, 4, we see that improvements of DLBri on

Table 4: F1 scores on LOWNER, transferring from English (EN) to seven non-western languages respectively. We report results of open-source baselines by running the official codes.

Language \ Method	BERT-f	AdvPicker	ConNER	DLBri (Ours)
Bangla	41.69	47.53	45.20	**52.06**
Farsi	48.11	51.60	51.22	**55.07**
Hindi	45.29	48.76	50.26	**53.25**
Korean	54.89	57.91	58.58	**60.75**
Russian	50.97	54.70	55.98	**60.89**
Turkish	62.28	64.05	65.43	**69.15**
Chinese (ZH)	57.15	63.19	65.04	**69.14**

western languages are less than non-western ones. The reason is that language gap between English and other western languages is not particularly large, and previous baselines have achieved promising performance on the transfer of similar languages, while DLBri still further improves performances on all language pairs and has greater advantages on distant language transfer.

5.3.1 Ablation Study. To verify the effect of each component, we perform ablation study in Table 5. We observe that: (1) Without *Progressive Contrastive Bridging*, we directly apply original mBERT in *Step* 2 and F1 score separately drops by 1.84%, 3.38% on western and non-western languages. The reason is that the progressive contrastive bridging effectively lessens the language discrepancy especially on distant language pairs and facilitates cross-lingual transfer. (2) Without high confidence (i.e., threshold $\delta = 0$) and low confidence (i.e., removing $\mathcal{L}_P$), the performance respectively reduces by 1.56%, 2.25% and 0.50%, 1.00%. That shows both high-confidence and low-confidence predictions can benefit model learning by collaborative-matching and self-penalization. (3) Without boundary consistency, we do not compute $\mathcal{L}^{\mathcal{T}}_{SBC}$, $\mathcal{L}^{\mathcal{S}}_{SBC}$, which leads to 0.96%, 1.96% declines and confirms the effectiveness of explicit boundary information in span extraction subtask. (4) Without *Subtask Combination*, we directly select the best of two student models in SSL framework for final results. The F1 scores drop by 1.66% and 1.84% on western and non-western languages. Meanwhile, we only remove subtask interaction layer, F1 scores decrease by 0.95% and 1.12%. That shows both knowledge distillation and subtask interaction are essential and effective in decomposed task.

5.3.2 Complexity Analysis. To conduct the analyses of complexity in Table 6, we run official codes of baselines under the same batch size and GPU for reporting results. MTMT [21] has not released codes, so we only estimate the number of parameters according to the paper. (1) For *#Param*, more model parameters may promote performance, and we calculate the number of parameters directly used for NER task prediction. Table 6 shows the parameter quantity of DLBri is close to or even less than some competitive baselines, which demonstrates our DLBri does not mainly gain from more parameters but more effective cross-lingual gap bridging strategies. (2) For *#Training Time*, we compute the spent time of processing per mini-batch data during training. Some methods (e.g., Unitrans, AdvPicker and DLBri) include multiple training steps, we accumulate the time of each step. (3) For *#Inference Speed*, we calculate the

Table 5: Ablation study on dev sets. Scores are averaged over 3 western and non-western languages in CoNLL, WikiAnn.

Method	Dev F1	
	western	non-western
DLBri (Ours)	**84.19**	**68.77**
w/o Progressive Contrastive Bridging	82.35	65.39
w/o high confidence	82.63	66.52
w/o low confidence	83.69	67.77
w/o boundary consistency	83.23	66.81
w/o Subtask Combination	82.53	66.93
w/o subtask interaction layer	83.24	67.65

Table 6: Analysis on the number of parameters and efficiency. s/B is the spent seconds per batch, while B/s is contrary.

Method	Unitrans	AdvPicker	MTMT	ConNER	DLBri (Ours)
#Param	177.86M	**177.27M**	~356M	559.91M	359.26M
#Training Time	5.56 s/B	3.91 s/B	–	**1.26** s/B	5.42 s/B
#Inference Speed	3.06 B/s	**3.47 B/s**	–	2.71 B/s	2.35 B/s

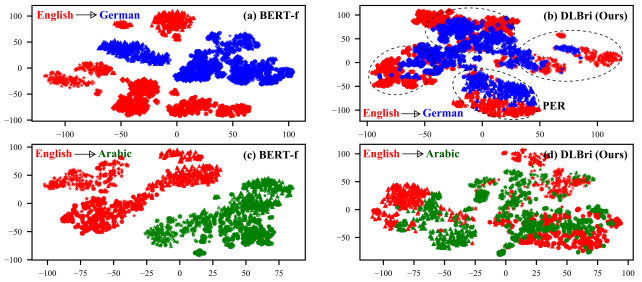

Figure 2: t-SNE visualization on two language pairs with BERT-f and DLBri. Different colors mark distinct languages and various entity types are marked with different shapes.

number of processed batches per second during test. The running efficiency of DLBri is not the best among compared strong baselines, but their efficiencies are still of the same order of magnitude. The reason may be that span extraction and type prediction subtask are trained and tested in serial on a GPU server. The efficiency actually can be improved by parallel processing. Comprehensively, considering the significant performance advantage of our DLBri, the running efficiency is reasonable and acceptable.

5.4 Experimental Analyses

*5.4.1 Effectiveness of Gap Bridging (***RQ2***).* To check if our DLBri effectively narrows the language gap and improves cross-lingual NER performance, Figure 2 visualizes the last hidden layer features of three languages on dev sets in BERT-f and DLBri when respectively transferring from English to German and Arabic. Each point in figure indicates an entity, which is an average features of all tokens in the entity and further averaged over two subtasks in DLBri. Red, blue and green color separately mark English, German and Arabic. Different shapes of points mark distinct entity types. Compared with BERT-f that fine-tunes mBERT on source-language training data and directly tests on target language, our DLBri pulls the source and target languages closer, and entities with the same type are clustered together more effectively despite of languages in Figure 2 (b), (d). As DLBri also uses mBERT as backbone, this shows that our method significantly bridges the language gap with better cross-lingual entity-aware representations and more effective transfer in subtasks, compared to BERT-f only based on mBERT.

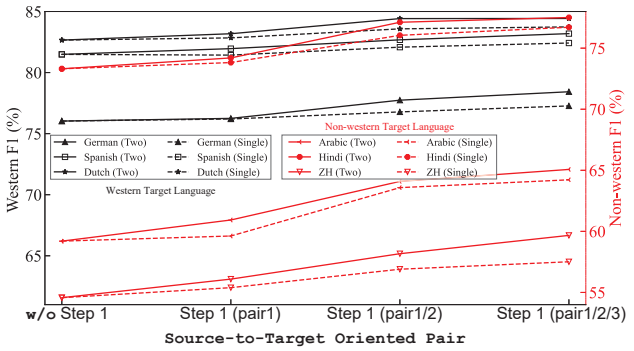

Figure 3: F1 score with the gradual addition of sentence pairs.

*5.4.2 Analysis on Source-to-Target Oriented Sentence Pair (***RQ3***).* To show the effect of intermediate language text, we gradually perform contrastive bridging on different pairs and report the influence on final F1 score of DLBri on six target-language test data in Figure 3. Our method adopts two mBERT encoders in *Step* 1 (progressive contrastive bridging) and we also conduct experiments with single mBERT for encoding the sentence pair. We can observe that two are better than one, because two encoders respectively preserve source and target oriented information, and provide better bridging effect and heterogeneous information for semi-supervised learning in subtask transfer. Meanwhile, performance increases roughly as we progressively perform the contrastive bridging on new sentence $pair_i$ ($i \in \{1, 2, 3\}$), which confirms the effectiveness and necessity of source-to-target oriented sentence pairs. Especially, our DLBri gains more from $pair_2$ with source-oriented and target-oriented intermediate sentences, as two intermediate texts have distinct language styles and would better serve as the intermediate languages. We also perform contrastive bridging only on the pair comprised of source and target language sentences, which respectively leads to 1.63% and 3.86% declines on CoNLL and WikiAnn test sets, and shows the effect of source-to-target oriented sentence pairs.

*5.4.3 Effectiveness of Decomposing NER Task (***RQ4***).* To verify the significant superiority of transferring in subtasks, we respectively conduct the cross-lingual transfer based on the combined tagging (e.g., B-LOC) and the decomposed subtasks in Table 8. BERT-f$_{decom}$ means BERT-f is decomposed into two subtasks, and DLBri$_{ctag}$ means we perform *Step* 2 with the combined tagging instead of decomposing NER task in DLBri. We see that BERT-f is obviously improved by decomposition (BERT-f$_{decom}$), and our DLBri also outperforms its variant DLBri$_{ctag}$ significantly. The reason is that entity span and type have different transfer difficulties, and transfer in subtasks can model the respective transferability, promoting

Table 7: F1 scores of two subtasks on 13 language pairs. Source$_{ft}$ (Target$_{ft}$) means fine-tuning mBERT on source (target) language training set and both test on target language. Comparing Δ in two subtasks, higher is marked in red and lower is blue.

Δ=Target$_{ft}$ − Source$_{ft}$		**CoNLL** (western)			**WikiAnn** (non-western)			**LOWNER** (non-western)						
Subtask		German	Spanish	Dutch	Arabic	Hindi	ZH	Bangla	Farsi	Hindi	Korean	Russian	Turkish	ZH
Span Extraction	Source$_{ft}$	84.59	86.47	93.57	63.03	81.09	59.54	62.41	56.89	63.15	62.48	65.39	72.38	70.78
	Target$_{ft}$	89.05	95.54	97.42	95.01	95.32	87.57	83.71	79.32	80.18	85.44	79.05	85.45	90.15
	Δ	4.46	9.07	3.85	31.98	14.23	28.03	21.30	22.43	17.03	22.96	13.66	13.07	19.37
Type Prediction	Source$_{ft}$	82.50	84.69	83.36	82.91	87.29	86.33	66.37	80.63	68.50	81.32	82.33	82.47	80.04
	Target$_{ft}$	89.10	92.19	93.58	93.62	95.28	91.05	81.24	89.10	80.87	90.96	86.52	90.65	91.01
	Δ	6.60	7.50	10.22	10.71	7.99	4.72	14.87	8.47	12.37	9.64	4.19	8.18	10.97

Table 8: F1 scores of variants to validate task decomposition, *decom* is decomposed, *ctag* is combined tagging, *sha* is shared.

Method / Language	BERT-f	BERT-f$_{decom}$	DLBri$_{ctag}$	DLBri$_{sha}$	DLBri (Ours)
German	69.56	71.29	76.78	77.41	**78.43**
Spanish	74.96	76.18	82.08	81.92	**83.19**
Dutch	77.57	78.92	83.53	84.10	**84.43**
Arabic	42.30	48.76	57.92	60.75	**65.07**
Hindi	67.60	69.84	74.96	75.40	**77.51**
ZH	52.90	53.76	57.51	57.90	**59.67**

the sufficient transfer. We also share the mBERT encoder but with respective classification head (DLBri$_{sha}$) between two subtasks, causing performance decline, which shows the necessity of separate encoding in two subtasks. This is also why we perform completely independent transfer in subtasks and then make subtask interaction.

5.4.4 Different Transfer Difficulties of Subtasks. As shown in Table 7 (detailed version of Table 1), to confirm the discrepant transfer difficulties of span extraction and type prediction subtask, we fine-tune mBERT for each subtask on labeled source-language (Source$_{ft}$) and target-language (Target$_{ft}$) training data respectively, and both predict on target-language test data. *Span extraction* subtask evaluates the predicted entity spans and *Type prediction* evaluates the predicted types of ground-truth entity spans. As the performance degradation of Source$_{ft}$ compared with Target$_{ft}$ results from the language discrepancy, the larger difference Δ implies the greater cross-lingual gap and more difficult transfer. Table 7 shows that span extraction subtask has the higher transfer difficulty than type prediction when transferring from English to non-western languages. The reason may be that there are significant differences in syntax and grammar between distant languages, which leads to the poor boundary detection in span extraction. Meanwhile, we see that type information transfer is less affected by language discrepancy, because type prediction relies more on token semantics which may benefit from multilingual BERT (mBERT). Therefore, decomposing NER task is reasonable, which fully considers the transfer characteristics of entity span and type information in corresponding subtasks and effectively reduces the cross-lingual gap. Overall, Table 7 shows that decomposing the task for cross-lingual NER is largely limited by span extraction, especially on distant language pairs (English to non-western).

Table 9: Average F1 scores on CoNLL, WikiAnn test set under different SSL frameworks. DLBri* marks using SSL of MTMT.

Method / Dataset	RIKD [22]	MTMT [21]	DLBri*	DLBri (Ours)
CoNLL (Western)	79.61	80.68	81.05	**82.02**
WikiAnn (Non-western)	55.55	58.60	65.33	**67.42**

5.4.5 Impact of Semi-supervised Learning (SSL) Framework. To explore the effect of different SSL frameworks in Table 9, we adopt the same SSL (i.e., semi-supervised knowledge distillation) in DLBri as MTMT [21], notated as DLBri*. We see that DLBri* still significantly outperforms MTMT and RIKD which exploits iterative knowledge distillation especially on distant language pairs. This demonstrates that the advantages of our DLBri do not mainly come from confidence-based SSL but representation-level and labeling-level gap bridging strategies. Comparing DLBri* with DLBri, we observe that devising better SSL framework can also further improve cross-lingual NER performance in the future.

6 CONCLUSION AND FUTURE WORK

This paper develops DLBri to bridge cross-lingual gap from the perspective of *representation* and *labeling*. Concretely, we devise the progressive contrastive bridging for learning better cross-lingual entity-aware *representations*. As for *labeling*, we decompose NER task into two subtasks to favour more sufficient transfer with the consideration of distinct transfer characteristics of entity span and type. These two perspectives are performed under the SSL framework. Extensive experiments confirm the rationality and effectiveness of our method.

ACKNOWLEDGMENTS

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] M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-resource cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 7415–7423.

[2] Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically Aware Neural Model for Named Entity Recognition in Low Resource Transfer Settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1462–1472.

[3] Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, and Jaime Carbonell. 2019. A Little Annotation does a Lot of Good: A Study in Bootstrapping Low-resource Named Entity Recognizers. 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, 5164–5174.

[4] Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 743–753.

[5] Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 6338–6353.

[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] Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net.

[8] 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 Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 4071–4077.

[9] 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.

[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] Ruohao Guo and Dan Roth. 2021. Constrained Labeled Data Generation for Low-Resource Named Entity Recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 4519–4533.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[13] Xiaolei Huang, Jonathan May, and Nanyun Peng. 2019. What Matters for Neural Cross-Lingual Named Entity Recognition: An Empirical Analysis. 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, 6395–6401.

[14] Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity Projection via Machine Translation for Cross-Lingual NER. 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, 1083–1092.

[15] Andrew Johnson, Penny Karanasou, Judith Gaspers, and Dietrich Klakow. 2019. Cross-lingual Transfer Learning for Japanese Named Entity Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*. Association for Computational Linguistics, 182–189.

[16] Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial Learning with Contextual Embeddings for Zero-resource Cross-lingual Classification and NER. 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, 1355–1360.

[17] Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-Guided Contrastive Learning for BERT Sentence Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2528–2540.

[18] 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.

[19] 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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 200–209.

[20] Yinghui Li, Yangning Li, Yuxin He, Tianyu Yu, Ying Shen, and Hai-Tao Zheng. 2022. Contrastive Learning with Hard Negative Entities for Entity Set Expansion. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 1077–1086.

[21] Zhuoran Li, Chunming Hu, Xiaohui Guo, Junfan Chen, Wenyi Qin, and Richong Zhang. 2022. An Unsupervised Multiple-Task and Multiple-Teacher Model for Cross-lingual Named Entity Recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 170–179.

[22] Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2021. Reinforced Iterative Knowledge Distillation for Cross-Lingual Named Entity Recognition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 3231–3239.

[23] Jie Liu, Shaowei Chen, Bingquan Wang, Jiaxin Zhang, Na Li, and Tong Xu. 2020. Attention as Relation: Learning Supervised Multi-head Self-Attention for Relation Extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, 3787–3793.

[24] Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 5834–5846.

[25] Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. 2022. Decomposed Meta-Learning for Few-Shot Named Entity Recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 1584–1596.

[26] Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A Large-scale Multilingual Dataset for Complex Named Entity Recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 3798–3809.

[27] Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap Translation for Cross-Lingual Named Entity Recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2536–2545.

[28] 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.

[29] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. 2021. FixBi: Bridging Domain Spaces for Unsupervised Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1094–1103.

[30] Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly Supervised Cross-Lingual Named Entity Recognition via Effective Annotation and Representation Projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1470–1480.

[31] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1946–1958.

[32] Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised Question Decomposition for Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 8864–8880.

[33] 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.

[34] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*. 135–152.

[35] Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021. A Co-Interactive Transformer for Joint Slot Filling and Intent Detection. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP). 8193–8197.

[36] Zengyi Qin, Jinglu Wang, and Yan Lu. 2022. MonoGRNet: A General Framework for Monocular 3D Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2022), 5170–5184. https://doi.org/10.1109/TPAMI.2021.3074363

[37] Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively Multilingual Transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 151–164.

[38] 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.

[39] Varsha Suresh and Desmond Ong. 2021. Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 4381–4394.

[40] Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

[41] 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*. 142–147.

[42] Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-Lingual Named Entity Recognition via Wikification. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 219–228.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

[44] 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.

[45] 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.

[46] Bohong Wu, Zhuosheng Zhang, Jinyuan Wang, and Hai Zhao. 2022. Sentence-aware Contrastive Learning for Open-Domain Passage Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1062–1074.

[47] Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 6505–6514.

[48] Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. 2020. UniTrans: Unifying Model Transfer and Data Transfer for Cross-Lingual Named Entity Recognition with Unlabeled Data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, 3926–3932.

[49] Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9274–9281.

[50] Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. 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, 833–844.

[51] Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural Cross-Lingual Named Entity Recognition with Minimal Resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 369–379.

[52] 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.

[53] Bowen Yu, Zhenyu Zhang, Jiawei Sheng, Tingwen Liu, Yubin Wang, Yucheng Wang, and Bin Wang. 2021. Semi-Open Information Extraction. In *Proceedings of the Web Conference 2021*. Association for Computing Machinery, 1661–1672.

[54] 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)*.

[55] Dongxu Zhang, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and Weiran Xu. 2016. Bitext Name Tagging for Cross-lingual Entity Annotation Projection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 461–470.

[56] 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.

[57] Xinghua Zhang, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Jiawei Sheng, Xue Mengge, and Hongbo Xu. 2021. Improving Distantly-Supervised Named Entity Recognition with Self-Collaborative Denoising Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 10746–10757.

[58] 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 International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 301–311.

[59] Ying Zhang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Target-oriented Fine-tuning for Zero-Resource Named Entity Recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 1603–1615.

[60] 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.

[61] Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. ConNER: Consistency Training for Cross-lingual Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[62] Ayah Zirikly and Masato Hagiwara. 2015. Cross-lingual Transfer of Named Entity Recognizers without Parallel Corpora. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 390–396.