

# ConNER: Consistency Training for Cross-lingual Named Entity Recognition

Ran Zhou<sup>\*1,2</sup> Xin Li<sup>†1</sup> Lidong Bing<sup>1</sup> Erik Cambria<sup>2</sup> Luo Si<sup>1</sup> Chunyan Miao<sup>2</sup>

<sup>1</sup>DAMO Academy, Alibaba Group <sup>2</sup>Nanyang Technological University, Singapore

{ran.zhou, xinting.lx, l.bing, luo.si}@alibaba-inc.com

{cambria, ascymiao}@ntu.edu.sg

## Abstract

Cross-lingual named entity recognition (NER) suffers from data scarcity in the target languages, especially under zero-shot settings. Existing translate-train or knowledge distillation methods attempt to bridge the language gap, but often introduce a high level of noise. To solve this problem, consistency training methods regularize the model to be robust towards perturbations on data or hidden states. However, such methods are likely to violate the consistency hypothesis, or mainly focus on coarse-grain consistency. We propose ConNER as a novel consistency training framework for cross-lingual NER, which comprises of: (1) translation-based consistency training on unlabeled target-language data, and (2) dropout-based consistency training on labeled source-language data. ConNER effectively leverages unlabeled target-language data and alleviates overfitting on the source language to enhance the cross-lingual adaptability. Experimental results show our ConNER achieves consistent improvement over various baseline methods.<sup>1</sup>

## 1 Introduction

With the emergence of large scale multilingual pretrained language models (Devlin et al., 2019; Conneau et al., 2020), cross-lingual named entity recognition (NER) performance has seen substantial improvement, especially under zero-shot setting where no labeled target-language training data is available (Tsai et al., 2016; Xie et al., 2018; Jain et al., 2019). Nevertheless, due to linguistic gaps between languages, NER models trained solely on the source-language data are likely to overfit the source language’s characteristics, and still suffer from sub-optimal performance when tested on target languages directly.

\* Ran Zhou is under the Joint Ph.D. Program between Alibaba and Nanyang Technological University.

<sup>†</sup> Corresponding author

<sup>1</sup>Our code is available at <https://github.com/RandyZhouRan/ConNER>.

To further enhance cross-lingual adaptability, various methods have been proposed to explicitly incorporate target-language information during training. These approaches can be roughly categorized into: (1) translate-train and (2) knowledge distillation. Translate-train produces parallel target-language training data by translating the source-language training data and mapping the labels (Xie et al., 2018; Bari et al., 2020; Li et al., 2020). However, the translated sentences are less natural and the label mapping is also error-prone. Moreover, without utilizing the abundant unlabeled target-language data, their performance is usually sub-optimal (Wu et al., 2020b). On the other hand, knowledge distillation trains the student model on target-language data with soft labels from a teacher model (Wu et al., 2020a,b; Chen et al., 2021; Liang et al., 2021; Zhong et al., 2022). Given the relatively low performance of the teacher model on target languages, the soft labels are noisy and limit the efficacy of unlabeled target-language data when using these labels as supervision signals.

In this paper, we primarily explore how to better utilize unlabeled target-language data. Instead of adopting knowledge distillation, which is vulnerable to the noise in soft labels, we adopt consistency training as a more fault-tolerant strategy. In general, consistency training works by enhancing the smoothness of output distributions and improving the model’s robustness and generalization (Miyato et al., 2018). Its training process is less sensitive towards noisy labels by regulating multiple predictions on different views of the same data point. Several works already attempted to apply consistency training on NER, including token-level and sequence-level consistency methods. Token-level consistency works on the same granularity as NER by regularizing the model to be invariant towards Gaussian noise (Zheng et al., 2021) or word replacement (Lowell et al., 2020).

However, such naïve noise or augmentation might violate the assumption that noised tokens share the same labels as the original ones. On the other hand, a sequence-level consistency method (Wang and Henaio, 2021) attempted to adapt back-translation-based (Edunov et al., 2018) consistency to NER. Due to difficulties in word alignment, they only use a coarse-grain consistency on the appearance of a certain type of entity in the original and back-translated sentences, which is plausibly a sub-optimal design for token-level tasks like NER. To this end, other works used constituent-based tagging schemes (Zhong et al., 2020).

We propose ConNER, a cross-lingual NER framework that primarily leverages consistency training on unlabeled target-language data using translation. Concretely, ConNER encourages the model to output consistent predictions between a span of tokens in the original sentence and their projection in the translated sentence. We tackle the problem of word alignment with alignment-free translation, and propose span-level consistency constraints to overcome the problem of token number changes during translation. Another advantage of translating unlabeled target-language data into the source language is that we can obtain more reliable training signals from the predictions on these translated sentences, because the model is trained with supervision from labeled source-language data and has better performance on the source language. Such reliable signals can be propagated back to the unlabeled target-language sentences through regularizing the prediction consistency between aligned spans. Furthermore, consistency training on the parallel unlabeled data helps to align different languages in the same representation space for better cross-lingual adaptability.

To mitigate overfitting on the source language, we also introduce a dropout-based consistency training on labeled source-language data, where we train our model to be robust towards noise induced by different dropout processes. In summary, ConNER enhances the model’s robustness towards both translation-based and dropout-based perturbations, which alleviates overfitting and achieves better cross-lingual generalization. To illustrate the effectiveness of ConNER, we conduct experiments on three transfer pairs and benchmark against various baseline methods. Experimental results show that ConNER achieves substantial improvement on cross-lingual NER over the baselines.

Our major contributions are as follows: (1) We propose a novel consistency training framework for cross-lingual NER and achieve consistent improvement over multiple baselines across various transfer pairs. (2) We present translation-based consistency training to effectively utilize unlabeled target-language data. It obviates using word-alignment tools and handles token number changes. (3) We introduce dropout-based consistency training to NER and reinforce the model’s robustness towards input noise and alleviate overfitting on labeled source-language data.

## 2 Methodology

In this section, we introduce ConNER, our consistency training framework for cross-lingual NER. As shown in Fig. 1, the overall framework consists of three components, namely (1) training with supervised cross-entropy loss, (2) dropout-based consistency training and (3) translation-based consistency training. Specifically, supervised cross-entropy loss is calculated on labeled source-language data, similar to a vanilla NER model. Secondly, dropout-out based consistency training (Section 2.4) feeds the same labeled source-language sample twice through the model, and enforces the model to output consistent probability distributions on the same token from two different dropout operations. Lastly, translation-based consistency training (Section 2.3) translates unlabeled target-language data into the source language via alignment-free translation method, and encourages the model to make consistent predictions on conjugate spans between the original and translated sentences.

### 2.1 Problem Definition

Following previous works, we formulate the task of cross-lingual NER as a sequence tagging problem. Given an input sentence comprising of  $n$  tokens  $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ , we aim to assign an NER tag  $y_i \in \mathcal{Y}$  to each token  $x_i \in \mathbf{X}$ , where  $\mathcal{Y}$  denotes the set of all possible NER tags. When there are a total number of  $N$  entity types  $\mathcal{C} = \{C_1, C_2, \dots, C_N\}$  with BIOES scheme  $\mathcal{P} = \{B-, I-, E-, S-\}$ , the token-level label space is  $\mathcal{Y} = \mathcal{P} \times \mathcal{C} \cup \{O\}$ . In our framework, apart from labeled source-language training data  $\mathcal{D}_l^{src}$ , unlabeled target-language data  $\mathcal{D}_u^{tgt}$  is also available during training.

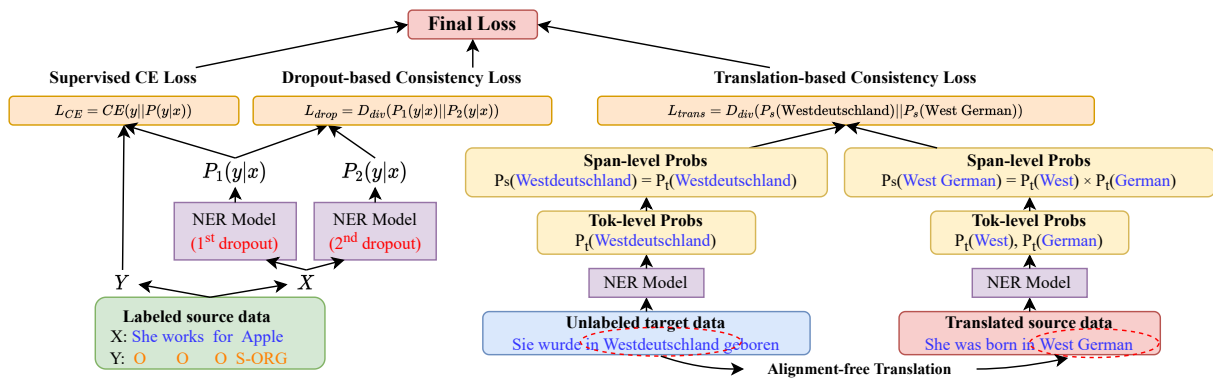


Figure 1: ConNER: consistency training for cross-lingual NER.

## 2.2 Consistency Training for NER

Consistency training (Miyato et al., 2018; Clark et al., 2018; Xie et al., 2020) aims to enhance the model’s robustness by regularizing model predictions to be invariant to label-preserving perturbations. Despite its success on CV and sentence-level NLP tasks (Miyato et al., 2018; Xie et al., 2020), consistency training for NER remains largely underexplored, where the model should be consistent on token-level or entity-level predictions.

In this work, we explore better consistency training methods for cross-lingual NER, which can be formulated as follows. Let  $\phi$  be a transformation function for generating small perturbation, such as text translation or hidden state noising. Given an input sequence of tokens (i.e., a sentence  $\mathbf{X}$ ) or hidden states, and a segment  $s$  in the sequence (e.g., a span of tokens), we apply  $\phi$  on the input to obtain a transformed sequence which contains the transformation of  $s$ , denoted as  $s'$ . For example, we translate  $\mathbf{X}$  to another language and  $s'$  is the translation of  $s$  in the translated sentence. We refer to  $(s, s')$  as a pair of conjugate segments. We encourage the model to output consistent probability distributions on  $s$  and  $s'$ , by minimizing the consistency loss as:

$$L_c = \frac{1}{m} \sum_{\substack{s \in \mathbf{X}, \\ s' \in \phi(\mathbf{X})}} D_{div}(P(y | s) || P(y | s')) \quad (1)$$

where  $D_{div}$  is a measure of divergence,  $m$  is the total number of conjugate segment pairs.

In this work, we introduce two variants of consistency training for NER which vary in the form of perturbation  $\phi$ : (1) translation-based perturbation on input text, and (2) dropout-based perturbation on token representations.

Depending on the form of perturbation,  $s$  can be a single token or a span of tokens, and correspondingly  $P(y|s)$  will be token-level or span-level probability. We explain in details in the following Sec. 2.3 and 2.4.

## 2.3 Translation-based Consistency Training

In this section, we present translation-based consistency training as the first variant, where we introduce a textual-level perturbation  $\phi$  by translating the original sentence into a different language. The conjugate segments  $(s, s')$  are now a span of tokens in the original sentence and its translation in the translated sentence, respectively. By enforcing the model to output consistent predictions on  $s$  and  $s'$ , the model tends to capture language-invariant features and thus has better cross-lingual adaptability.

To enable the above translation-based consistency training for NER, there are still two challenging issues to be resolved: (1) the word alignments between the original and translated sentences are not available and the existing alignment tools for inducing word alignments are often error-prone; (2) the perturbed sentence (i.e., the translation) may have different length and word order, and it is likely that the token-level one-to-one correspondence for calculating consistency loss will not be applicable. To tackle these challenges, we propose a novel method which leverages alignment-free translation to automatically construct span-level correspondence, and caters for token number changes as well. Note that our approach is not sensitive to the change of sequence length after perturbation because it no longer requires the consistency on token-level predictions.

### Input sentence with candidate span marked in blue:

Bruce Willis wurde in Westdeutschland geboren.

#### 1. Replace span with placeholder and translate:

Bruce Willis wurde in SPAN<sub>44</sub> geboren.

⇒ Bruce Willis was born in SPAN<sub>44</sub>.

#### 2. Obtain translation of span:

Westdeutschland ⇒ West German

#### 3. Replace placeholder with translated span

##### and extract conjugate pair:

Bruce Willis was born in West German.

Conjugate pair: (Westdeutschland, West German)

#### 4. Converting token-level probs to span-level probs:

$$\begin{aligned} P_s(\text{PER}|\text{West German}) &= P_t(\text{B-PER}|\text{West}) \cdot P_t(\text{E-PER}|\text{German}) \\ P_s(\text{LOC}|\text{West German}) &= P_t(\text{B-LOC}|\text{West}) \cdot P_t(\text{E-LOC}|\text{German}) \\ P_s(\text{ORG}|\text{West German}) &= P_t(\text{B-ORG}|\text{West}) \cdot P_t(\text{E-ORG}|\text{German}) \\ P_s(\text{MISC}|\text{West German}) &= P_t(\text{B-MISC}|\text{West}) \cdot P_t(\text{E-MISC}|\text{German}) \\ P_s(\text{O}|\text{West German}) &= P_t(\text{O}|\text{West}) \cdot P_t(\text{O}|\text{German}) \\ P_s(\text{illegal}|\text{West German}) &= 1 - P_s(\text{PER}) - P_s(\text{LOC}) \\ &\quad - P_s(\text{ORG}) - P_s(\text{MISC}) - P_s(\text{O}) \end{aligned}$$

Figure 2: Example illustrating the key steps of translation-based consistency training.

### Alignment-free Translation

Given an input sentence  $\mathbf{X} = \{x_1, \dots, x_n\}$  which contains a candidate span  $s_{ij} = \{x_i, \dots, x_j\}$ , we aim to translate it into a different language as  $\mathbf{X}' = \{x'_1, \dots, x'_{n'}\}$  and locate the span  $s'_{kl} = \{x'_k, \dots, x'_l\}$  that corresponds to  $s_{ij}$ . Adapting the method proposed in Liu et al. (2021), we conduct alignment-free translation to automatically locate  $s'_{kl}$  in the translated sentence. Specifically, as shown in Fig. 2, we first replace the candidate span  $s_{ij}$  with a placeholder token SPAN<sub>ij</sub> to obtain  $\hat{\mathbf{X}} = \{x_1, \dots, x_{i-1}, \text{SPAN}_{ij}, x_{j+1}, \dots, x_n\}$ . Then, we feed  $\hat{\mathbf{X}}$  through a translation engine to obtain its translation  $\hat{\mathbf{X}}'$ . We notice the placeholder SPAN<sub>ij</sub> remains mostly unchanged in the translated sentence, and it indicates where the candidate span should appear in the translation. Next, we aim to revert the placeholder back to real text, by replacing it with the translation of  $s_{ij}$ , denoted as  $s'_{kl}$ . Lastly, we substitute the placeholder in  $\hat{\mathbf{X}}'$  with  $s'_{kl}$ , to obtain the final translation  $\mathbf{X}'$ , and  $(s_{ij}, s'_{kl})$  is recorded as a pair of conjugate segments.

### Token-level to Span-level Probability

As mentioned above,  $s_{ij}$  and  $s'_{kl}$  might have different number of tokens, and thus it is intractable to calculate one-to-one consistency loss between

### Algorithm 1 Token to span probability conversion

---

Given candidate span  $s_{ij} = \{x_i, \dots, x_j\}$  and token-level probabilities  $\{P_t(y_u|x_u) \mid u \in [i, j]\}$   
 $P_s(\text{O}|s_{ij}) \leftarrow 1, P_s(\text{illegal}|s_{ij}) \leftarrow 1$   
**for** CLS  $\in \mathcal{C}$  **do**  
  **if**  $i = j$  **then**  
     $P_s(\text{CLS}|s_{ij}) \leftarrow P_t(\text{S-CLS}|x_i)$   
  **else if**  $j - i = 1$  **then**  
     $P_s(\text{CLS}|s_{ij}) \leftarrow P_t(\text{B-CLS}|x_i) \times P_t(\text{E-CLS}|x_i)$   
  **else**  
     $P_s(\text{CLS}|s_{ij}) \leftarrow P_t(\text{B-CLS}|x_i)$   
    **for**  $u \in (i, j)$  **do**  
       $P_s(\text{CLS}|s_{ij}) \leftarrow P_s(\text{CLS}|s_{ij}) \times P_t(\text{I-CLS}|x_u)$   
    **end for**  
     $P_s(\text{CLS}|s_{ij}) \leftarrow P_s(\text{CLS}|s_{ij}) \times P_t(\text{E-CLS}|x_j)$   
  **end if**  
   $P_s(\text{illegal}|s_{ij}) \leftarrow P_s(\text{illegal}|s_{ij}) - P_s(\text{CLS}|s_{ij})$   
**end for**  
**for**  $u \in [i, j]$  **do**  
   $P_s(\text{O}|s_{ij}) \leftarrow P_s(\text{O}|s_{ij}) \times P_t(\text{O}|x_u)$   
**end for**  
 $P_s(\text{illegal}|s_{ij}) \leftarrow P_s(\text{illegal}|s_{ij}) - P_s(\text{O}|s_{ij})$   
 $P_s(z_{ij}|s_{ij}) = \{P_s(\text{CLS}_1|s_{ij}), \dots, P_s(\text{CLS}_N|s_{ij}),$   
   $P_s(\text{O}|s_{ij}), P_s(\text{illegal}|s_{ij})\}$   
**return**  $P_s(z_{ij}|s_{ij})$ 

---

pairs of single tokens. We resolve this issue by converting token-level probability distributions (over label space  $\mathcal{Y}$  defined in Sec. 2.1) to span-level probability distributions over the set of possible entity types, i.e.,  $\mathcal{C} \cup \{\text{O}\}$ . Specifically, given the last hidden state  $h_i$  of token  $x_i$ , the NER model feeds it through a linear layer to obtain token-level probability distribution  $P_t(y_i|x_i, h_i)$ . As shown in Algorithm 1, to obtain span-level probability distribution  $P_s(z_{ij}|s_{ij})$ , we multiply probabilities of token-level labels which form a legal label sequence under BIOES scheme. Step 4 in Fig. 2 also gives an example calculation as illustration.

Note that we introduce an extra span-level class illegal to include all label sequences of  $s_{ij}$  that violate the BIOES rules (e.g., {S-PER, I-PER}), such that all entries of  $P_s(z_{ij}|s_{ij})$  add up to one to form a valid probability distribution. In other words, the label space of span-level label  $z_{ij}$  is  $\mathcal{C} \cup \{\text{O}, \text{illegal}\}$ . We apply the same transformation on  $s'_{kl}$  to obtain its span-level distribution  $P_s(s'_{kl})$ .

Finally, we calculate  $D_{div}$  between span-level probabilities of each conjugate segment pair  $(s_{ij}, s'_{kl})$  using bidirectional Kullback-Leibler (KL) divergence, and Eqn. 1 becomes:

$$\begin{aligned} L_{\text{trans}} &= \frac{1}{m} \sum_{\substack{s_{ij} \in \mathbf{X} \\ s'_{kl} \in \mathbf{X}'}} \frac{1}{2} [\text{KL}(P_s(z_{ij}|s_{ij}) \parallel P_s(z_{kl}|s'_{kl})) \\ &\quad + \text{KL}(P_s(z_{kl}|s'_{kl}) \parallel P_s(z_{ij}|s_{ij}))] \end{aligned} \quad (2)$$

In our experiments, we apply translation-based consistency training on unlabeled target-language sentences by translating them into the source language. It is noteworthy that, although our translation-based consistency training is applicable to arbitrary span pairs from the parallel sentences, it would be more valuable for NER task if the selected spans are likely to be entities. Therefore, we use a weak NER tagger trained with only labeled source-language data to assign NER labels on unlabeled target-language sentences, and use the predicted entities as our candidate spans for consistency training. Note that we only use predictions from the weak tagger to locate the boundary of candidate spans, which to some extent reduces the impact of predictions with wrong entity types.

## 2.4 Dropout-based Consistency Training

We introduce another variant of consistency training that is based on dropout (Srivastava et al., 2014). Inspired by Wu et al. (2021); Gao et al. (2021), we consider dropout as a form of perturbation  $\phi$  at representational level. By regularizing the model to be invariant to different random dropouts, we encourage the model to make predictions based on more diverse features rather than overfitting on certain spurious features. As word order does not change after applying dropout, the conjugate segments now become the representations of the same token undergoing different dropout processes.

Concretely, we pass the same sentence  $\mathbf{X}$  through the encoder twice. As a result of different stochastic dropout in each pass, we obtain two different sets of token representations for  $\mathbf{X}$ . Subsequently, the model outputs two different token-level probability distributions  $P_1(y_i|x_i)$  and  $P_2(y_i|x_i)$  over the label space  $\mathcal{Y}$ . We also adopt bidirectional KL divergence as  $D_{div}$ , and calculate dropout-based consistency loss as:

$$L_{\text{drop}} = \frac{1}{n} \sum_{x_i \in \mathbf{X}} \frac{1}{2} [\text{KL}(P_1(y_i|x_i) || P_2(y_i|x_i)) + \text{KL}(P_2(y_i|x_i) || P_1(y_i|x_i))] \quad (3)$$

We apply dropout-based consistency training on labeled source-language data in our experiments, but NER labels are not used.

## 2.5 Training Objective

As mentioned above, we apply translation-based consistency training on unlabeled target-language data, and dropout-based consistency training on labeled source-language data.

These consistency losses are combined with the supervised cross-entropy loss (i.e.,  $L_{\text{CE}}$ ) on labeled source-language data, which gives the total training objective as:

$$L_{\text{total}} = \sum_{X \in \mathcal{D}_l^{\text{src}}} (L_{\text{CE}} + \alpha L_{\text{drop}}) + \sum_{X \in \mathcal{D}_u^{\text{tgt}}} \beta L_{\text{trans}} \quad (4)$$

where  $\alpha$  and  $\beta$  are weight coefficients.

## 3 Experiments

In this section, we evaluate our consistency training framework ConNER on cross-lingual NER, and compare with various state-of-the-art models.

### 3.1 Dataset

We conduct experiments on CoNLL02 and CoNLL03 datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), which involve four languages: English (En), German (De), Spanish (Es) and Dutch (Nl). To evaluate on more distant transfer pairs, we also experiment on WikiAnn dataset (Pan et al., 2017) of English (En), Chinese (Zh), Arabic (Ar) and Hindi (Hi). We adopt the BIOES entity annotation scheme in our experiments. Following previous zero-shot cross-lingual NER works, we use the original English training set as our training data  $\mathcal{D}_l^{\text{src}}$ , while treating all other languages as target languages and evaluate on their test sets. We also use the original English development set for early-stopping and model selection. The NER labels from target language training sets are removed, and they are used as our unlabeled target-language data  $\mathcal{D}_u^{\text{tgt}}$ .

### 3.2 Implementation Details

We implement our vanilla NER model using XLM-RoBERTa-large (Conneau et al., 2020) with CRF head (Lample et al., 2016). The emission probabilities of the CRF layer, which are obtained by feeding the last hidden states of XLM-R through a feed-forward layer, are used as the token-level probabilities for calculating consistency losses in Section 2.3 and 2.4. We use AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate  $l_r = 2e-5$  and set batch sizes for both labeled and unlabeled data as 16. We train the NER model for 10 epochs and select the best checkpoint using English dev set. The model is evaluated on target-language test sets and we report the averaged micro-F1 score over 3 runs. We use Google Translate as our translation engine for the main experiments.

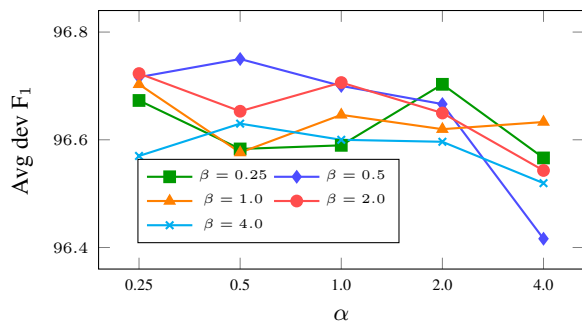


Figure 3: Averaged dev F1 for hyperparameter tuning.

We conduct a grid search on both hyperparameters  $\alpha$  and  $\beta$  over the range of  $\{0.25, 0.5, 1.0, 2.0, 4.0\}$ . The English dev set F1 averaged across three transfer pairs are shown in Fig. 3. We set  $\alpha = 0.5, \beta = 0.5$  based on the best dev set performance.

### 3.3 Main Results

As shown in Table 1, ConNER achieves significant performance improvement (avg. 2.61) over the zero-shot vanilla baseline trained with only labeled source-language data. This demonstrates the effectiveness of our method, achieved by enhancing the model’s robustness towards perturbations on both labeled and unlabeled data.

There is a considerable performance gap between ConNER and the translate-train methods (i.e., the upper section of Table 1). As translate-train suffers from unnatural translation and inaccurate label mapping, their performance is sub-optimal under full dataset settings where the vanilla baselines already present decent zero-shot performance. This also shows the importance of exploiting unlabeled target-language data, apart from leveraging the available source training data and its parallel translation. Our ConNER also outperforms various state-of-the-art knowledge distillation methods (i.e., the middle section of Table 1). Soft-labels in knowledge distillation are often noisy, and such noise is likely to be amplified by forcing the student model’s predictions to match the soft-labels. In comparison, our training method is more fault-tolerant, and induces better alignment between languages by training on parallel unlabeled data. Furthermore, we compare ConNER with two consistency training baselines, which use word replacement (Lowell et al., 2020) or Gaussian noise (Zheng et al., 2021) as perturbations (i.e., the lower section of Table 1).

As we can see, word replacement might change the NER label of the original word and violate the consistency hypothesis, and thus lead to performance drops on some languages. On the other hand, Gaussian noise only applies perturbations with small magnitudes, and brings in limited improvement. In contrast, our ConNER leverages translation and dropout as high-quality perturbations. As a result, ConNER shows significant improvements over the consistency training baselines. Note that although TOF (Zhang et al., 2021b) uses extra labeled MRC data during training, our ConNER still surpasses it on all of the target languages.

In general, our translation-based consistency training effectively utilizes unlabeled target data and encourages the model to extract language-invariant features, while our dropout-based consistency further alleviates overfitting on source-language data for better cross-lingual adaptability.

### 3.4 Low-resource NER

We also investigate the effectiveness of ConNER on low-resource cross-lingual NER, where source-language training data is also limited. Specifically, we randomly sample 5%, 10% and 25% examples from the original English training set as our low-resource training sets. We apply ConNER and evaluate on the original target-language test sets. As shown in Table 2, ConNER is also effective under low-resource settings, achieving average performance improvements of 4.32, 2.20, and 1.77 F1 score compared with vanilla baseline respectively. We observe larger gains as the low-resource training set becomes smaller, which demonstrates the importance of leveraging unlabeled data and preventing overfitting under low-resource settings. Another interesting finding is that, using only 25% of the original training data, our ConNER is already comparable to a vanilla NER model trained on the full training set.

### 3.5 Distant Languages

To evaluate the robustness of ConNER on a wider range of languages, we conduct experiments on three non-western target languages: Chinese (Zh), Arabic (Ar) and Hindi (Hi), from the WikiAnn dataset (Pan et al., 2017).

As shown in Table 3, we observe the vanilla baseline has relatively low performance on these non-western languages, due to larger linguistic gaps. This might also introduce some noise in the choice of candidate spans for our method.

Method	De	Es	Nl	Avg
vanilla baseline	72.10	79.64	81.29	77.68
<i>Translate-train</i>				
BWET (Xie et al., 2018) <sup>†</sup>	73.84	79.03	80.93	77.93
MTL-WS (Li et al., 2020)	74.92	76.73	78.62	76.76
MulDA (Liu et al., 2021) <sup>‡</sup>	74.55	78.14	80.22	77.64
<i>Knowledge-distillation</i>				
TSL (Wu et al., 2020a)	73.16	76.75	80.44	76.78
Unitrans (Wu et al., 2020b)	74.82	79.31	82.90	79.01
AdvPicker (Chen et al., 2021)	75.01	79.00	82.90	78.97
<i>Consistency training</i>				
MLM-replace (Lowell et al., 2020) <sup>‡</sup>	71.65	79.70	80.83	77.39
xTune (Zheng et al., 2021) <sup>‡</sup>	74.78	80.03	81.76	78.85
<b>ConNER (Ours)</b>	<b>77.14</b>	<b>80.50</b>	<b>83.23</b>	<b>80.29</b>
<i>Multi-task</i>				
TOF (Zhang et al., 2021b) <sup>*</sup>	76.57	80.35	82.79	79.90

Table 1: Experimental results of cross-lingual NER methods. <sup>†</sup> denotes reproduced results with XLM-R-large, which are higher than the original paper based on LSTM. <sup>‡</sup> denotes reimplementations on our datasets. Results without markers are cited from the original papers. <sup>\*</sup>Zhang et al. (2021b) use extra labeled MRC data for multi-task training, thus it is not directly comparable with our method.

#Train	Method	De	Es	Nl	Avg
5%	vanilla	64.91	65.76	71.22	67.30
	ConNER	69.94	70.15	74.76	71.62
10%	vanilla	67.58	69.92	77.70	71.73
	ConNER	70.88	71.23	79.70	73.94
25%	vanilla	70.72	77.39	79.24	75.78
	ConNER	73.47	77.92	80.66	77.35

Table 2: Low-resource NER results.

Method	Zh	Ar	Hi
vanilla	33.10	53.00	73.00
Wu and Dredze (2020)	43.90	45.50	66.60
Wu et al. (2020a)	31.14	50.91	72.48
ConNER (Ours)	39.17	59.62	74.49

Table 3: Results on distant transfer pairs.

Nonetheless, ConNER is still effective and demonstrates substantial improvements over the baseline. Such results verify that ConNER is general and robust for not only transfer pairs of the same language but also distant transfer pairs.

## 4 Analysis

### 4.1 Ablation Study

To analyze the contribution of each component and justify the framework design of ConNER, we conduct the following ablation studies: (1) **trans-**

Method	De	Es	Nl	Avg
vanilla	72.10	79.64	81.29	77.68
ConNER	77.14	80.50	83.23	80.29
trans-unlabel	76.87	80.76	81.66	79.77
dropout-label	74.46	80.90	81.63	78.99
trans-label	71.16	79.51	81.03	77.23
dropout-unlabel	61.72	77.80	80.80	73.44

Table 4: Ablation study.

**unlabel**, where we only keep the translation-based consistency on unlabeled target data; (2) **dropout-label**, where we only keep the dropout-based consistency on labeled source data; (3) **trans-label**, where we translate labeled source language data to the target language, and calculate consistency loss between entities and their translations in the translated sentences. (4) **dropout-unlabel**, where we feed the same unlabeled target data twice through the encoder, and calculate consistency loss between the two output distributions of the same token.

As shown in Table 4, both trans-unlabel and dropout-label achieve improved performance compared to the vanilla baseline. Nevertheless, the average improvement from trans-unlabel is more significant, demonstrating the importance of leveraging unlabeled target-language data for cross-lingual tasks. Combining trans-unlabel and dropout-label, our ConNER achieves further performance gains.

Method	De	Es	Nl	Avg
KL-unlabel	76.08	80.55	80.06	78.90
KL-trans	76.65	80.31	81.22	79.39
bi-KL	76.87	80.76	81.66	79.77

Table 5: Choice of divergence measure.

However, when applying translation-based consistency to labeled data (trans-label) instead, we observe some performance drop. This is possibly because label information is already present, such that model trained on labeled source data already learns to make accurate predictions on its parallel translated sentence, and translation-based consistency becomes redundant.

Moreover, if we apply the dropout-based consistency on unlabeled target-language data (dropout-unlabel), we observe an unreasonably low performance (e.g., German). We attribute this to catastrophic forgetting of multilinguality. Without the guidance from the labels on target-language data, the model is likely to enter a failure mode that trivially minimizes the divergence for target-language tokens. As a result, target-language features are diverted away from the source-language in the representation space, causing the model to fail on the target language. In contrast, the translation-based consistency of ConNER makes use of parallel unlabeled data such that the source and target language are always aligned through the conjugate spans.

## 4.2 Choice of Divergence Measure

In this section, we investigate the effect of different divergence measures on the translation-based consistency. We consider three configurations: (1) KL-unlabel, where the predictions on the unlabeled target data are taken as ground-truth probabilities (Eqn. 5a); (2) KL-trans, where the predictions on the translated source data as ground-truth probabilities (Eqn. 5b); (3) bi-KL, which adopts a bidirectional KL divergence (Eqn. 5c).

$$D_{\text{KL-unlabel}} = \text{KL}(P_s(z_{ij}|s_{ij}) || P_s(z_{kl}|s'_{kl})) \quad (5a)$$

$$D_{\text{KL-trans}} = \text{KL}(P_s(z_{kl}|s'_{kl}) || P_s(z_{ij}|s_{ij})) \quad (5b)$$

$$D_{\text{bi-KL}} = \frac{1}{2} [\text{KL}(P_s(z_{ij}|s_{ij}) || P_s(z_{kl}|s'_{kl})) + \text{KL}(P_s(z_{kl}|s'_{kl}) || P_s(z_{ij}|s_{ij}))] \quad (5c)$$

From the experimental results in Table 5, we observe the overall performance of KL-trans is higher than KL-unlabel. This is probably due to the higher performance of the model on English data.

By aligning predictions on unlabeled target-language data to the predictions on translated data in English, we propagate more accurate training signals from English to the target language and achieve better cross-lingual adaptation. Meanwhile, KL-unlabel also helps to reduce overfitting to English by taking target-language data as a reference. Bi-KL takes advantage of both KL-trans and KL-unlabel and achieves the best averaged performance.

## 4.3 Case Study

We conduct a case study to qualitatively show how translation-based consistency training benefits cross-lingual NER. As depicted in Fig. 4, the vanilla model predicts German word "*Augustäpfel*" as an organization, but gives an inconsistent prediction for its English translation "*August apples*" (i.e., O). In contrast, our model, which is enhanced by translation-based consistency training, attempts to behave similarly on the aligned spans and therefore is capable to calibrate the prediction on "*Augustäpfel*". Similarly, while the vanilla model predicts German "*Wicker*" as a location, our method leverages the context information to make the correct prediction as an organization.

## 4.4 Robustness to Translation Systems

In above experiments, we used Google Translate as our translation engine. To demonstrate the robustness of ConNER towards different machine translation systems, we rerun the experiments using Opus-MT (Tiedemann and Thottingal, 2020), a popular open-source machine translation system. As shown in Table 6, there exists a minor performance drop when changing Google Translate to Opus-MT, possibly due to its lower translation quality. Nevertheless, it is observed that ConNER with Opus-MT is still superior to all baseline methods.

Method	De	Es	Nl	Avg
Google Translate	77.14	80.50	83.23	80.29
Opus-MT	75.86	80.60	82.47	79.64

Table 6: Results using Google Translate and Opus-MT.

## 5 Related Work

NER is the task to locate and classify named entities into pre-defined categories (Xu et al., 2021; Zhou et al., 2022b; Xu et al., 2022). Similar to other tasks under cross-lingual settings (Zhang et al.,



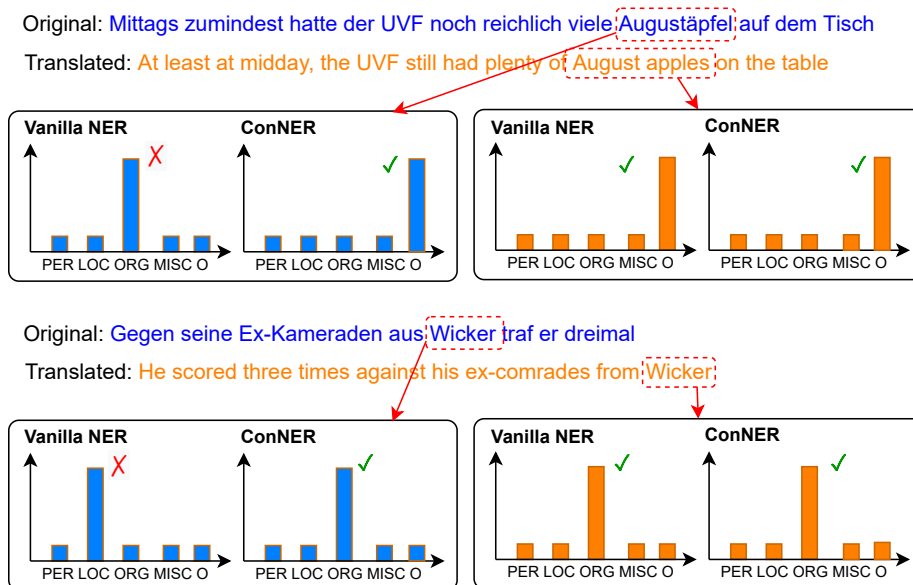


Figure 4: Case study on translation-based consistency.

2021a; Liu et al., 2022a,b; Zhou et al., 2022a), cross-lingual NER suffers from linguistic gaps. Recent works on cross-lingual NER can be roughly categorized into three categories: (1) translate-train (2) knowledge distillation (3) consistency training.

**Translate-train** creates target-language pseudo-labeled data by translating the source-language training set and projecting the NER labels of source-language texts to the translated texts. Xie et al. (2018) use bilingual embeddings to translate training data word-by-word and leverage self-attention for word order robustness. Bari et al. (2020) propose unsupervised translation using word-level adversarial learning and augmented fine-tuning with parameter sharing and feature augmentation. Li et al. (2020) propose a warm-up mechanism to distill multilingual task-specific knowledge from the translated data in each language. Translate-train focuses on translating the available labeled source-language data, but it ignores the abundant unlabeled target-language data.

**Knowledge distillation** trains the student model using the soft-labels obtained from a teacher model. Wu et al. (2020a) apply knowledge distillation and propose a similarity measuring method to better weight the supervision from different teacher models. Wu et al. (2020b) leverage both soft-labels and hard-labels from teacher models trained on the source data and translated training data. Chen et al. (2021) use a discriminator to select less language-dependent target-language data via similarity to

the source language. Zhang et al. (2021b) use extra machine reading comprehension data to enable multi-task adaptation and achieve improved NER performance. In general, knowledge distillation is sensitive to the noise in pseudo-label, especially when transferring to the target language.

**Consistency training** applies perturbation on data or hidden states and regularizes the model to output consistent predictions. Existing works apply word replacement (Lowell et al., 2020) or Gaussian noise (Zheng et al., 2021) as perturbations. However, the validity and diversity of these noise or augmentations largely limit their performance (Xie et al., 2020). Wang and Henao (2021) use back-translation as a high-quality augmentation. However, due to word alignment issues, they focus on the entity-appearance consistency in the whole sentence, but ignore the location of entities.

## 6 Conclusions

In this work, we proposed a novel consistency framework for cross-lingual NER. Our method ConNER enhances the model’s robustness towards perturbations on labeled and unlabeled data, via translation-based and dropout-based consistency. With translation-based consistency, we tackle the challenges of word alignment and token number changes via alignment-free projection and token-level to span-level conversion. Compared with multiple baseline methods, ConNER achieves consistent performance improvements.

## 7 Limitations

The method proposed requires the use of machine translation models or systems, which in some cases might not be easily accessible. Also, our translation-based consistency works the best when the selected candidate spans are likely to be entities and do not cross entity boundaries. Therefore, the strategy for candidate span selection could be a direction for future improvement.

## Acknowledgements

This research is partly supported by the Alibaba-NYU Singapore Joint Research Institute, Nanyang Technological University.

## References

- 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*, volume 34, pages 7415–7423.
- 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)*, pages 743–753.
- Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. **Semi-supervised sequence modeling with cross-view training**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. **Unsupervised cross-lingual representation learning at scale**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. **SimCSE: Simple contrastive learning of sentence embeddings**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.
- 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)*, pages 1083–1092.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. **Neural architectures for named entity recognition**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.
- Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020. Unsupervised cross-lingual adaptation for sequence tagging and beyond. *arXiv preprint arXiv:2010.12405*.
- 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 & Data Mining*, pages 3231–3239.
- 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)*, pages 5834–5846.
- Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq Joty, and Luo Si. 2022a. Enhancing multilingual language model with massive multilingual knowledge triples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Linlin Liu, Thien Hai Nguyen, Shafiq Joty, Lidong Bing, and Luo Si. 2022b. **Towards multi-sense cross-lingual alignment of contextual embeddings**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4381–4396, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. **Decoupled weight decay regularization**. In *International Conference on Learning Representations*.
- David Lowell, Brian E Howard, Zachary C Lipton, and Byron C Wallace. 2020. Unsupervised data augmentation with naive augmentation and without unlabeled data. *arXiv preprint arXiv:2010.11966*.

- Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.
- 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)*, pages 1946–1958.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.
- Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*.
- 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)*.
- 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 Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- 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*, pages 219–228.
- Rui Wang and Ricardo Henao. 2021. Unsupervised paraphrasing consistency training for low resource named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5308.
- Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34.
- Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020a. [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*, pages 6505–6514.
- Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. 2020b. [Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data](#). In *Proceedings of IJCAI*, pages 3926–3932.
- Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482.
- Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime G Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379.
- Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.
- Lu Xu, Zhanming Jie, Wei Lu, and Lidong Bing. 2021. Better feature integration for named entity recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3457–3469.
- Weiwen Xu, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. Peerda: Data augmentation via modeling peer relation for span identification tasks. *arXiv preprint arXiv:2210.08855*.
- Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021a. [Cross-lingual aspect-based sentiment analysis with aspect term code-switching](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230.
- Ying Zhang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021b. Target-oriented fine-tuning for zero-resource named entity recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1603–1615.
- Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. 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)*, pages 3403–3417.
- Xiaoshi Zhong, Erik Cambria, and Amir Hussain. 2020. Extracting time expressions and named entities with constituent-based tagging schemes. *Cognitive Computation*, 12(4):844–862.
- Xiaoshi Zhong, Erik Cambria, and Amir Hussain. 2022. Does semantics aid syntax? An empirical study on named entity recognition and classification. *Neural Computing and Applications*, 34:8373–8384.
- Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2022a. Enhancing cross-lingual prompting with mask token augmentation. *arXiv preprint arXiv:2202.07255*.

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