CROP: Zero-shot Cross-lingual Named Entity Recognition with Multilingual Labeled Sequence Translation

Jian Yang†, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, Furu Wei

1State Key Lab of Software Development Environment, Beihang University; 2Microsoft Research Asia; 3The University of Hong Kong

{jiaya, hongchengguo, lizj}@buaa.edu.cn; {shaohan, shumma, lidong1, dozhang, fuwei}@microsoft.com; yuweiyin@hku.hk

Abstract

Named entity recognition (NER) suffers from the scarcity of annotated training data, especially for low-resource languages without labeled data. Cross-lingual NER has been proposed to alleviate this issue by transferring knowledge from high-resource languages to low-resource languages via aligned cross-lingual representations or machine translation results. However, the performance of cross-lingual NER methods is severely affected by the unsatisfactory quality of translation or label projection. To address these problems, we propose a Cross-lingual Entity Projection framework (CROP) to enable zero-shot cross-lingual NER with the help of a multilingual labeled sequence translation model. Specifically, the target sequence is first translated into the source language and then tagged by a source NER model. We further adopt a labeled sequence translation model to project the tagged sequence back to the target language and label the target raw sentence. Ultimately, the whole pipeline is integrated into an end-to-end model by the way of self-training. Experimental results on two benchmarks demonstrate that our method substantially outperforms the previous strong baseline by a large margin of +3~7 F1 scores and achieves state-of-the-art performance.

1 Introduction

Named entity recognition (NER) focuses on recognizing entities from raw text into predefined types (Sang, 2002; Sang and Meulder, 2003; Yadav and Bethard, 2018; Fang et al., 2021; Lin et al., 2021; Wang and Henao, 2021), which is an essential component for downstream natural language processing (NLP) tasks, such as information retrieval (Banerjee et al., 2019) and question answering (Przybyla, 2016; Aliod et al., 2006). However, most of the existing approaches are highly dependent on the annotated training data and do not perform well in low-resource languages.

Zero-shot cross-lingual NER aims to address this challenging problem by transferring knowledge from the high-resource source language with lots of annotated corpora to those languages without any labeled data (Xie et al., 2018). Some methods leverage the cross-lingual representations (Ni et al., 2017), where the NER model is trained on the labeled corpus of the source language and then directly evaluated on target languages. Due to the success of multilingual pretrained language models (Devlin et al., 2019; Conneau et al., 2020), these model-based transfer methods have shown a significant improvement in cross-lingual NER. Another line of research is the data-based transfer (Wu et al., 2020a), which adopts word-to-word translation to project the cross-lingual NER labels. For example, Liu et al. (2021) employ a multilingual translation model with placeholders for label projection. Nevertheless, these methods are still limited by weak entity projection and do not leverage the unlabeled corpora in target languages.

Along the line of using the multilingual model to encourage knowledge transfer among different languages, we propose a Cross-lingual Entity Projection (CROP) framework to enable zero-shot cross-lingual NER with the help of a multilingual labeled sequence translation model. Specifically, the target sequence is first translated into the source language and then tagged by a source NER model. We further adopt a labeled sequence translation model to project the tagged sequence back to the target language and label the target raw sentence. Ultimately, the whole pipeline is integrated into an end-to-end model by the way of self-training. Experimental results on two benchmarks demonstrate that our method substantially outperforms the previous strong baseline by a large margin of +3~7 F1 scores and achieves state-of-the-art performance.
Projection (CROF) framework to leverage the unlabeled corpora of target languages, which is supported by a strong multilingual labeled sequence translation model guided by multiple bilingual corpora and the corresponding phrase-level alignment information. In Figure 1, the unlabeled target sentence is forward translated to the source language and tagged by the source NER model. Then, we use the labeled sequence translation model to back-translate the annotated sentence to the target language. Given the target annotated sentence, we project the entity labels of “Gothenburg” to the target raw sentence through lexical matching. Finally, we use self-training to integrate the pipeline into an end-to-end NER model.

Specifically, we construct multilingual corpora to train the labeled sequence translation model, where the aligned spans of the sentence pair are both surrounded by the boundary symbols. We conduct experiments on two benchmarks, including XTREME-40 of 40 languages and CoNLL-5 of 5 languages. Experimental results show that our method reaches new state-of-the-art results. Furthermore, we also evaluate the performance of the multilingual labeled sequence translation model and visualize multilingual sentence representations. Analytic results demonstrate that our method can transfer knowledge among even distant languages.

2 Zero-shot Cross-lingual NER

Given the source NER model Θsrc \(_{ner}\) only trained on the source NER dataset and the target raw sentence \(x = (x_1, \ldots, x_m)\) with \(m\) words, the zero-shot cross-lingual NER aims to identify each word of target language to predefined types and then obtains the labels \(t = (t_1, \ldots, t_m)\). The problem definition of zero-shot cross-lingual NER is described as:

\[
P(t|x) = \prod_{i=1}^{m} P(t_i|x; \Theta_{src}^{ner})
\]

where the target raw sentence \(x\) and labels \(t\) have the same length \(m\). \(t_i\) is the \(i\)-th label. The source language has annotated labels but the target corpora have no accessible handcrafted labels. \(P(t|x)\) represents the predicted distributions of labels. The source NER model \(\Theta_{src}^{ner}\) trained on the source annotated corpus is expected to be evaluated on the target language without any labeled dataset. The previous work (Wu et al., 2020a) propose to unify the model-based transfer and data-based transfer with machine translation to transfer knowledge from the source language to the target language.

3 CROF

3.1 Framework

In Figure 2, given \(K\) target languages \(L_{tgt} = \{L_k\}_{k=1}^{K}\), the NER model is first trained on the NER dataset \(D_{x,t}^{Lsrc} = \{(x(i), t(i))\}_{i=1}^{N}\) of the source language \(L_{src}\) with \(N\) samples, where \(x(i)\) is the input sentence and \(t(i)\) contains labels. The raw sentences in \(D_{x,t}^{Lsrc}\) are translated to the source language and tagged by the source NER model. Then, the source annotated corpora are back-translated to the target sentences via a labeled sequence translation model. The labels of the target translated sentences are projected to the target raw corpora to construct the annotated corpora \(D_{x,f(x)}^{L_k}\), where \(f(x)\) is projected label of the sentence \(x\) by a simple lexical matching between translated entities and original words. The source corpus \(D_{x,f(x)}^{Lsrc}\) and target annotated corpora \(D_{x,f(x)}^{L_k}\) are further utilized by self-training.

3.2 Backbone Model for NER

Our backbone model for NER is comprised of an encoder and a linear classifier to identify entities to predefined types. Given the input sentence \(x = (x_1, \ldots, x_m)\) with \(m\) words, we use the encoder \(\Theta_e\) to extract top-layer features:

\[
H = Encoder(x; \Theta_e)
\]

where \(H = (h_1, \ldots, h_m)\) are the representations of the last encoder layer, where \(h_i\) is the \(i\)-th word representation of the input sentence \(x\). \(\Theta_e\) are parameters of the feature extractor.

Then, a sequence of representations \(H = (h_1, \ldots, h_m)\) are fed into a linear classifier with the softmax function to generate the probability distribution of each input word:

\[
P(t|x) = \text{Softmax}(W_cH + b_c)
\]

where \(t = (t_1, \ldots, t_m)\) are corresponding labels of the input sentence, and \(\Theta_{ner} = \{W_c, b_c\}\) represent model parameters of the NER backbone model. \(P(t|x) \in R^{m \times T}\) is the predicted probabilities and \(T\) is the number of the predefined types. In this work, we set \(T = 7\) on the XTREME benchmark and \(T = 9\) on the CoNLL benchmark.

3.3 Labeled Sequence Translation

We adopt the multilingual labeled sequence translation (LST) to transfer knowledge from high-resource to low-resource languages. The
bilingual pair \( x = (x_1, \ldots, x_m) \) with \( m \) words and \( y = (y_1, \ldots, y_n) \) with \( n \) words are used to construct the pseudo labeled pair \( x_p = (x_1, \ldots, b_1, y_{u_1}, b_1, \ldots, x_m) \) and \( y_p = (y_1, \ldots, b_1, y_{u_1}, b_1, \ldots, y_n) \), where \( y_{u_1} \) is the target translation of source piece \( x_{u_1} \). \( x_{u_2} \) denotes the source phrase from the \( u_1 \)-th token to the \( u_2 \)-th token and \( y_{v_1} \) denotes the target phrase from the \( v_1 \)-th token to the \( v_2 \)-th token. \( b_i \) is the boundary symbol to indicate the \( i \)-th entity. We use the alignment tool eflomal\(^1\) to extract the aligned phrases of the sentence pair. \( x_p \) and \( y_p \) are used to help the model tackle labeled sequence translation, where \( x_p \) and \( y_p \) have multiple aligned spans surrounded by boundary tokens. For each sentence pair, we randomly sample 0~10 aligned spans from the pair and use the boundary symbols \( \{b_1, \ldots, b_{10}\} \) to construct the labeled sequence \( x_p \) and \( y_p \) in the labeled sequence translation training.

Given bilingual corpora \( D_b = \{D_b^k\}_{k=1}^K \) of \( K \) languages, where one side is the source language \( L_{src} \) and the other side is the language \( L_{tgt} \), the multilingual model is trained on corpora \( D_b \):

\[
\mathcal{L}_t = -\sum_{k=1}^K \mathbb{E}_{x,y \in D_b^k} \log \mathbb{P}(y|x; \Theta_{mt})
\]

(4)

where \( \Theta_{mt} \) are parameters of translation model.

To support labeled sequence translation (LST), we use the sentence pair to construct training samples, where the aligned spans in the sentence pair are surrounded by the boundary symbols using phrase-level alignment pairs. In Figure 3, \( x \) and \( y \) are sentence pair. The aligned fragments of the source sentence and target sentence are both annotated by the boundary symbols. These samples are used for the training of labeled sequence translation:

\[
\mathcal{L}_{lst} = -\sum_{k=1}^K \mathbb{E}_{x,y \in D_b^k} \log \mathbb{P}(y|x; \Theta_{mt})
\]

(5)

where \( (x_p,y_p) \) is the sentence pair constructed by the original sentence pair and the phrase-level alignment pairs.

Our model is optimized by jointly minimizing the translation objective and labeled sequence translation objective:

\[
\mathcal{L}_{mt} = \alpha \mathcal{L}_t + (1 - \alpha) \mathcal{L}_{lst}
\]

(6)

where \( \mathcal{L}_t \) is the objective of multilingual translation and \( \mathcal{L}_{lst} \) is the objective of the multilingual labeled sequence translation. We alternate two training objectives by setting \( \alpha = 0.5 \). Our multilingual model supports (i) multilingual translation and (ii) labeled sequence translation. After alternately training on two objectives, we obtain the final multilingual translation model \( \Theta_{mt} \). Once the multilingual training is done, our model serves as the off-the-shelf multilingual labeled translation model and does not require alignments.

During the inference stage, the source sentence \( x \) with labels is switched to labeled sequence \( x_p \), where all entities are surrounded by indicators. Then, the model translates the source labeled sentence \( x_p \) to the target labeled sentence \( y_p \). The boundary symbol indicates the entities in the translation sentence. For example, the translation phrase \( y_{v_1} \) have the same NER labels with the source phrase \( x_{u_1} \), where both phrases are surrounded by the boundary token \( b_i \).

### 3.4 Cross-lingual Entity Projection

Given the labeled corpus \( D^{L_{src}}_x \) of the source language \( L_{src} \) and the unlabeled corpora \( \{D^k_L\}_{k=1}^K \) of \( K \) languages, the source NER

---

\(^1\)https://github.com/robertostling/eflomal
model $\Theta^{\text{src}}_{\text{ner}}$ is used to tag the unlabeled training corpora of target languages, aided by the labeled translation model $\Theta_{\text{mt}}$.

**Forward Translation** The target raw corpora $\{D^x_{\text{src}}\}_{k=1}^K$ of $K$ languages are translated into the source language and tagged by the source NER model $\Theta^{\text{src}}_{\text{ner}}$. We obtain the source labeled translated corpora $\{D^x_{\text{mt}}\}_{k=1}^K$, where $f(\cdot)$ is the predictor of the source NER model $\Theta^{\text{src}}_{\text{ner}}$.

**Labeled Sequence Back-translation** The source annotated corpora are back-translated to the target languages with entity labels. In Figure 3, the source sentence $x_p = (b_1, e_1, b_1, x_2, x_3, b_2, e_2, b_2)$ is translated into the target sentence $y_p = (b_2, e_2, b_2, y_3, b_1, e_1, b_1)$. The boundary symbols $b_1$ and $b_2$ are used to locate the translated entities $e_1$ and $e_2$ in $y_p$. We obtain the back-translated data $\{D^y_{\text{mt}}\}_{k=1}^K$ by the translation model $\Theta_{\text{mt}}$.

**Entity Matching** Given the target translated entities with labels $D_{\text{tgt}}$, we search the matched entities in the unlabeled target sentence by lexical matching (string matching word by word). In Figure 1, “哥德堡” in the unlabeled sentence matches “哥德堡” in the translated sentence, so “哥德堡” is labeled with the same entity type LOC (Location). The labels of translated entities are projected into the raw sentence to construct target labeled corpora $\{D^y_{\text{mt}}\}_{k=1}^K$. Finally, the target annotated corpora and the original corpus $D^x_{\text{src}}$ are used for multilingual NER model training.

### 3.5 Self-training

Given a labeled corpus $D^x_{\text{src}}$ of the source language and target unlabeled corpora $\{D^x_{\text{src}}\}_{k=1}^K$ of target languages, the training objective based on $D^x_{\text{src}}$ is formulated as below:

$$L_{\text{src}} = \mathbb{E}_{x,t \in D^x_{\text{src}}} \left[ - \log P(t|x; \Theta^{\text{all}}_{\text{ner}}) \right]$$

where $\Theta^{\text{all}}_{\text{ner}}$ are NER model parameters.

Then, we leverage the source NER model $\Theta^{\text{src}}_{\text{ner}}$ trained on the labeled corpus to project the entity labels to the target raw corpora described in Section 3.4 and get labeled corpora $\{D^x_{\text{mt}}\}_{k=1}^K$. The multilingual corpora of target languages with predicted labels are adopted to train a neural network with the combined loss function $L_{\text{tgt}}$ as below:

$$L_{\text{tgt}} = \sum_{L_k \in L_{\text{tgt}}} \mathbb{E}_{x,y \in D^y_{\text{mt}}} \left[ - \log P(f(x)|y; \Theta^{\text{all}}_{\text{ner}}) \right]$$

where $x$ is the golden target data and $f(x)$ is the pseudo label generated by the cross-lingual entity projection.

The multilingual NER model is jointly trained on the original dataset and target corpora with labels:

$$L_{\text{all}} = L_{\text{src}} + L_{\text{tgt}}$$

where $L_{\text{src}}$ and $L_{\text{tgt}}$ are training objectives on the original and distilled dataset.

### 4 Experiments

#### 4.1 Dataset

**CCaligned** Our labeled multilingual model continues to be tuned on the same training data called CCaligned (El-Kishky et al., 2020) as the previous work (Fan et al., 2020; Goyal et al., 2021). We use a collection of parallel data in different languages from the CCaligned dataset, where the parallel data is paired with English and other 39 languages. The valid and test sets are from the FLORES-101 dataset (Goyal et al., 2021).

**CoNLL-5** Following the previous work (Wu et al., 2020a), we construct a cross-lingual dataset from CoNLL-2002 (Sang, 2002) for Spanish (es) and Dutch (nl) NER, CoNLL-2003 (Sang and Meulder, 2003) for English (en) and German (de) NER, and NoDaLiDa-2019 (Johansen, 2019) for Norwegian (no) NER. All entities are classified into 4 entity types in BOI-2 format, including LOC, MISC, ORG, and PER. Each dataset is split into training, dev, and test set. Detailed statistics can be found in Table 1.

**XTREME-40** The proposed method is further evaluated on the cross-lingual NER dataset from the XTREME benchmark (Hu et al., 2020). Named entities in Wikipedia are annotated with LOC, PER, and ORG tags in BOI-2 format. Following the previous work (Hu et al., 2020), we use the same split for the training, dev, and test set.
Post-Processing The synthetic data is post-processed to train the multilingual NER model. (i) We use the language detection toolkit to filter the translated sentence with the incorrect language. (ii) We delete sequences, which exceed the maximum length (128 words) and only contain O (other) tags. (iii) The NER model trained on the multilingual corpora is directly employed to tag the unlabeled corpora. The discarded sentence is re-labeled by the multilingual NER model. Finally, we combine the labels predicted by the source NER model $\Theta_{\text{src}}$ trained on the original dataset and the multilingual NER model $\Theta_{\text{all}}$ trained by self-training to improve the accuracy of label projection.

4.2 Baselines and Evaluation

Our method is compared with the different baselines initialized by cross-lingual pretrained models including mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) for model-based transfer. We also conduct experiments without any pretrained model on the Transformer (Vaswani et al., 2017) architecture. UniTrans (Wu et al., 2020a) unifies both model transfer and data transfer for cross-lingual NER. Following this line of research, MulDA (Liu et al., 2021) proposes a sequence translation method to translate labeled training data of the source languages to other languages and avoids the word order change caused by word-to-word or phrase-to-phrase translation. Besides, we also produce the results of Translate-Train, where the labeled source corpus is translated into the other labeled corpora of multiple languages using our multilingual model. Following the previous work (Sang, 2002; Wu et al., 2020a), the metrics are the entity-level precision, recall, and F1 scores. For simplicity, we report the F1 scores of different methods in all tables.

4.3 Training Details

Multilingual Labeled Translation The pre-trained multilingual model M2M$\_\text{large}$ is adopted as the translation model, which has 12 layers with an embedding size of 1024 and 16 attention heads. We continue fine-tuning the model with Adam ($\beta_1 = 0.9, \beta_2 = 0.98$) on the labeled corpora constructed by the multilingual corpora and the alignment pairs from CCAligned, where the parallel data is paired with English and other 39 languages and the phrase-level alignment pairs are extracted by the alignment tool $\text{floital}$. The learning rate is set as $1e-4$ with a warm-up step of 4,000. The batch size is set as 1536 tokens on 32 A100 GPUs.

Cross-lingual NER For a fair comparison, we implement all methods using the same architecture and model size. We separately adopt the base architecture of Transformer, mBERT, and XLM-R as the backbone model, which all have 12 layers with an embedding size of 768, a feed-forward network size of 3072, and 12 attention heads. We set the batch size as 24 for CoNLL-5 and 32 for XTREME-40. The NER model is trained on CoNLL-5 for 15 epochs and XTREME-40 for 10 epochs, where the warm-up step is the 10% steps of the whole training steps. The synthetic data is post-processed to train the multilingual NER model. We delete sequences, which exceed the maximum length (128 words) and only contain O (other) tags. The NER model trained on the multilingual corpora is directly employed to tag the unlabeled corpora. The discarded sentence is re-labeled by the multilingual NER model. Finally, we combine the labels predicted by the source NER model $\Theta_{\text{src}}$ trained on the original dataset and the multilingual NER model $\Theta_{\text{all}}$ trained by self-training to improve the accuracy of label projection.

4.4 Main Results

CoNLL-5 Table 2 presents the results of our method and previous baselines on transferring knowledge from English to other four languages, including es, nl, de, no. We can observe that the XLM-R gains strong improvement compared to previous baselines due to the effective cross-lingual transfer. Based on the cross-lingual pretrained model, our method can leverage cross-lingual entity projection to further encourage transferability from the NER model of the source language to the multilingual NER model of all languages. Our method significantly outperforms the previous strong baseline UniTrans on average, especially on German by a large margin +5.3 points. It can be attributed to our multilingual model, which has better translation quality on German and Norwegian than Spanish and Dutch.
### Table 1: Statistics of the CoNLL-5 (Sang, 2002; Sang and Meulder, 2003) and NoDaLiDa (Johansen, 2019) NER benchmarks.

<table>
<thead>
<tr>
<th>Language</th>
<th>Type</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (en)</td>
<td>#Sentences</td>
<td>15.0K</td>
<td>3.5K</td>
<td>3.7K</td>
</tr>
<tr>
<td></td>
<td>#Entities</td>
<td>23.5K</td>
<td>6.0K</td>
<td>5.7K</td>
</tr>
<tr>
<td>German (de)</td>
<td>#Sentences</td>
<td>12.7K</td>
<td>3.1K</td>
<td>3.2K</td>
</tr>
<tr>
<td></td>
<td>#Entities</td>
<td>11.9K</td>
<td>4.8K</td>
<td>3.7K</td>
</tr>
<tr>
<td>Spanish [es]</td>
<td>#Sentences</td>
<td>8.3K</td>
<td>1.9K</td>
<td>1.5K</td>
</tr>
<tr>
<td></td>
<td>#Entities</td>
<td>18.8K</td>
<td>4.3K</td>
<td>3.6K</td>
</tr>
<tr>
<td>Dutch [nl]</td>
<td>#Sentences</td>
<td>15.8K</td>
<td>2.9K</td>
<td>5.2K</td>
</tr>
<tr>
<td></td>
<td>#Entities</td>
<td>13.3K</td>
<td>2.6K</td>
<td>3.9K</td>
</tr>
<tr>
<td>Norwegian [no]</td>
<td>#Sentences</td>
<td>15.7K</td>
<td>2.4K</td>
<td>1.9K</td>
</tr>
<tr>
<td></td>
<td>#Entities</td>
<td>10.9K</td>
<td>1.6K</td>
<td>1.4K</td>
</tr>
</tbody>
</table>

Table 2: Results of our proposed method CROP and prior state-of-the-art methods for zero-resource cross-lingual NER. The dag symbol represents that the score is directly reported from the previous work.

<table>
<thead>
<tr>
<th>Language</th>
<th>es</th>
<th>nl</th>
<th>de</th>
<th>no</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Täckstrom et al. (2012)†</td>
<td>59.3</td>
<td>58.4</td>
<td>40.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tsai et al. (2016)†</td>
<td>60.6</td>
<td>61.6</td>
<td>48.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Smith et al. (2017)†</td>
<td>65.1</td>
<td>65.4</td>
<td>58.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mayhew et al. (2017)†</td>
<td>64.1</td>
<td>63.4</td>
<td>57.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Xie et al. (2018)†</td>
<td>72.4</td>
<td>71.3</td>
<td>57.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bari et al. (2019)†</td>
<td>73.5</td>
<td>69.9</td>
<td>61.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jain et al. (2019)†</td>
<td>75.9</td>
<td>74.6</td>
<td>65.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wu and Dredze (2019)†</td>
<td>74.5</td>
<td>79.5</td>
<td>71.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wu et al. (2020b)†</td>
<td>76.8</td>
<td>80.4</td>
<td>73.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>mBERT (Devlin et al., 2019)</td>
<td>74.6</td>
<td>77.9</td>
<td>75.0</td>
<td>77.4</td>
<td>76.2</td>
</tr>
<tr>
<td>XLM-R (Conneau et al., 2020)</td>
<td>77.4</td>
<td>78.9</td>
<td>73.4</td>
<td>80.9</td>
<td>77.7</td>
</tr>
<tr>
<td>+Translate-Train</td>
<td>77.8</td>
<td>79.2</td>
<td>74.2</td>
<td>81.3</td>
<td>78.1</td>
</tr>
<tr>
<td>UniTrans† (Wu et al., 2020a)</td>
<td>79.3</td>
<td>82.9</td>
<td>74.8</td>
<td>81.2</td>
<td>79.6</td>
</tr>
<tr>
<td>MuIDA (Liu et al., 2021)</td>
<td>77.5</td>
<td>78.4</td>
<td>78.2</td>
<td>82.1</td>
<td>79.1</td>
</tr>
<tr>
<td>CROP (Our Method)</td>
<td>78.1</td>
<td>79.5</td>
<td>80.1</td>
<td>83.1</td>
<td>80.2</td>
</tr>
</tbody>
</table>

### XTREME-40
Table 3 compares the performance of our method with previous relevant methods initiated by different cross-lingual pretrained language models including mBERT and XLM-R. Given our translation model, the multilingual translated annotated corpora (Translate-Train) from the data of source languages can be used to improve the model performance compared to the XLM-R. Particularly, our proposed method gains significant improvement compared to other languages by a large margin (nearly +6 F1 points), due to the effectiveness of cross-lingual entity projection. All experimental results demonstrate that our proposed framework strengthens transferability from the source language to nearly 39 target languages.

### Ablation Study
To verify the effectiveness of our method, we separately study the effects of the model-based transfer by cross-lingual pretrained model and the data-based transfer by cross-lingual entity projection. Our method has two advantages: (1) the model is trained on the original multilingual corpora with pseudo labels, which avoids the extra translation error, (2) our method uses the multilingual model trained on 41 languages to improve the entity projection of low-resource languages. In Table 4, Transformer 3 without any transfer methods gets the worst performance (only 15.1 F1 scores). Our method 2 without any pretrained model outperforms Transformer 3 by +43.0 F1 points, which has the similar transferability to the cross-lingual pretrained language models. Combining the merits of the cross-lingual pretrained model and self-training for multiple languages, we obtain the best performance on the XTREME-40 benchmark.

### Distribution of Multilingual Corpora
An important difference between our method and the previous baselines is that we provide an effective way to leverage the unlabeled corpora of target languages. The raw data is first translated to the source language data and annotated by the NER model trained on the original dataset. Then, the translated source sentences are back-translated to target languages, where the entity labels are projected to the target raw words. Our cross-lingual entity projection avoids the extra translation errors instead of direct utilization for translated labeled corpora. In Figure 4(a), we visualize the distribution of the encoder representations by randomly sampling 1K sentences of each language from the target golden corpora. Figure 4(b) shows the distribution of the round-trip translated target corpora. We observe that the distribution of translated corpora has changed a lot since there are incorrectly translated words highly affected by translation quality, especially for low-resource languages.

### Performance of Multilingual Translation
To ensure the effectiveness of our method, we evaluate the translation performance of 40 languages between M2M (Goyal et al., 2021) and our labeled sequence multilingual translation model on the FLORES-101 benchmark. Compared to M2M, our model supports the additional language eu by extending the fine-tuning data. Therefore, we report the SentencePiece-based BLEU using SacreBLEU5 of 39 translation directions except eu languages.

---

5https://github.com/ngoyal2707/sacrebleu
Table 3: Results of our proposed method CROP and other relevant baselines for cross-lingual NER. “Avgall” represents the average F1 scores of all 39 languages on the test set of the XTREME-40 benchmark.

Table 4: Ablation study of our proposed method. Avgall denotes the average F1 scores of 39 languages.

![Figure 4: t-SNE (Maaten and Hinton, 2008) visualization of the sentence representations for the golden data (a) and the translated data generated by our multilingual model (b). Each color denotes one language.](image)

Quality of Labeled Sequence Translation Section 3.3 introduces the multilingual labeled sequence translation, where the entities are surrounded by the boundary symbols and then translated to the target language. The multilingual model is trained with the bilingual corpus and the corresponding phrase-level alignment pairs to ensure the quality of labeled sequence translation. We calculate the precision of the baseline model and our model by randomly sampling 250 sentence pairs of each language from the whole training data.

Effect of Training Data Size To discuss the effect of the target labeled corpora, we plot F1 scores with different training data sizes in Figure 5. The performance is influenced by the ratio between the size of the original dataset (20K sentences) and the multilingual corpora (400K sentences after filtering). We randomly sample \( N = \{1K, 2K, \ldots, ALL\} \) sentences from the whole corpora to train the NER model. With the training data size increasing, the NER model gets better performance. Surprisingly, only 1K pseudo annotated
sentences bring large improvement to the zero-shot cross-lingual NER, which benefits from knowledge transfer of the multilingual self-training. When the size of target annotated corpora is greater than 10K, our method gets exceptional performance.

**Quality of Entity Projection** Given the target annotated translated sentence and the raw sentence, our method searches the matched entity and projects the labels to the raw sentence. After filtering the sentences, we utilize the labeled sentences with pseudo labels for multilingual NER model training. Figure 6 reports the F1 scores of the projected labels of the target corpora compared to the ground-truth labels, where each language has high F1 scores. The accurate cross-lingual label projection with an average of 82.1 F1 scores of 39 languages guarantees the positive influence of our method to avoid excessive noise interference.

**Example Study** Table 7 lists a concrete example to compare our multilingual model with the baseline. In practice, we set the special token \_SLOT\_ as the boundary symbol \_SLOT\_ for translation, where \_SLOT\_ is used as the boundary symbol. The positions of the boundary symbols \_SLOT\_ and \_SLOT\_ are misplaced during translation for the baseline model.

In contrast, the multilingual model trained with alignment pairs accurately translates sentences and maintains the correct position of boundary symbols owing to the phrase-level alignment information.

**Transfer for Distant Languages** Compared to the transferability inaugurated by cross-lingual pre-trained models, our method bridges the gap between the source language and distant target languages. The average F1 scores of similar and distant languages to English are denoted by \(\text{Avg}_{\text{sim}}\) and \(\text{Avg}_{\text{dis}}\). In Figure 7, \(\text{Avg}_{\text{sim}}\) gains +5.1 points improvement while \(\text{Avg}_{\text{dis}}\) outperforms XLM-R by a large margin +19.2 points. The NER model trained on the English corpus initialized by the pre-trained model is easier to be extended to similar languages but is hard to be transferred to distant languages (Leng et al., 2019). Through cross-lingual entity projection, our method productively encourages knowledge transfer from the source language to distant languages contrasted with the baseline.

**Explanation for Entity Matching** In Figure 8, we list two detailed examples of entity matching (a) mismatched entity and (b) matched entity. For the first example, “哥的堡” in the back-translated target is not mismatched to “哥德堡” in the raw target by the lexical matching, so the labels of “哥的堡” (LOC) can not be projected to the “哥德堡”. For the second example, “哥德堡” in the back-translated target is the same as “哥德堡” in the raw sentence word by word, so we can obtain the labeled entity “哥德堡” (LOC) in the target.
sentence. The target sentences with missing entities are discarded, where the labels can not be projected to the entity like in the first example. Finally, we only need to select the 10% sentences of all raw target sentences for the multilingual NER training to avoid extra noise and get state-of-the-art performance compared to previous baselines.

5 Related Work

Cross-lingual NER Named entity recognition (NER) identifying the named entities into the pre-defined types has achieved huge progress in recent years (Sang and Meulder, 2003; Yadav and Bethard, 2018; Li et al., 2020, 2021; Wang et al., 2021; Aly et al., 2021; Shaffer, 2021). Cross-lingual NER model supporting multiple languages is a key component for various downstream natural language processing (NLP) tasks, including information retrieval (Banerjee et al., 2019), question answering (Aliod et al., 2006), and co-reference resolution (Hajishirzi et al., 2013). The previous works can be classified into two categories including model-based transfer (Xie et al., 2018; Mueller et al., 2020) and data-based transfer methods (Lison et al., 2020; Ding et al., 2020; Liu et al., 2021; Zhou et al., 2021b). The model-based transfer methods benefit from the state-of-the-art cross-lingual pretrained model (Devlin et al., 2019; Conneau et al., 2020) and the aligned cross-lingual word embeddings (Xie et al., 2018). Wu et al. (2020a) emphasizes that the model-based transfer and data-based transfer methods are complementary to each other.

Multilingual Translation Inspired by the success of the neural machine translation (Bahdanau et al., 2015; Vaswani et al., 2017), multilingual translation has attracted considerable attention due to its capability to handle multiple languages in a shared single model (Pan et al., 2021; Xie et al., 2021; Zhou et al., 2021a; Zhang et al., 2021). Previous works explicitly leverage the word-level or phrase-level extracted alignment information to improve performance. (Song et al., 2019; Yang et al., 2020, 2021). Recently, massively multilingual models (Fan et al., 2020; Goyal et al., 2021) are proposed, which all are trained on large sources of training data. Motivated by previous works, we combine the phrase-level alignment pairs and the many-to-many multilingual model covering 40 languages to construct a labeled sequence translation system for the cross-lingual NER task in this work.

6 Conclusion

In this work, we propose a novel zero-shot cross-lingual NER framework with a multilingual labeled sequence translation model advised by multilingual corpora and phrase-level alignment pairs. The knowledge of the source NER model is effectively transferred to target languages by a round-trip translation and label projection. In this way, the multilingual translation model plays the role of the bridge to transfer knowledge from source languages to low-resource target languages. Experimental results evaluated on the CoNLL-5 and XTREME-40 benchmarks demonstrate the effectiveness of our method compared to the strong baselines.

7 Limitations

The total number of languages in our multilingual labeled sequence translation was limited owing to the data availability of cross-lingual NER. Once NER datasets of more languages are available, we can train a stronger multilingual translation model to further enhance the overall performance. In future work, our method can be scaled up to hundreds of languages to meet the needs of practical industrial scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62276017, U1636211, 61672081), the 2022 Tencent Big Travel Rhino-Bird Special Research Program, and the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2021ZX-18).
References


Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In ACL 2020, pages 5849–5859.


Ran Zhou, Ruidan He, Xin Li, Lidong Bing, Erik Cambria, Lao Si, and Chunya Miao. 2021b. MELM: data augmentation with masked entity language modeling for cross-lingual NER. CoRR, abs/2108.13655.