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

Jian Yang[1] [*], **Shaohan Huang**[2], **Shuming Ma**[2], **Yuwei Yin**[3],
Li Dong[2], **Dongdong Zhang**[2], **Hongcheng Guo**[1], **Zhoujun Li**[1] [†], **Furu Wei**[2]
[1]State Key Lab of Software Development Environment, Beihang University
[2]Microsoft Research Asia; [3]The University of Hong Kong
{jiaya, hongchengguo, lizj}@buaa.edu.cn;
{shaohanh, 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 **Cro**ss-lingual Entity **P**rojection 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.

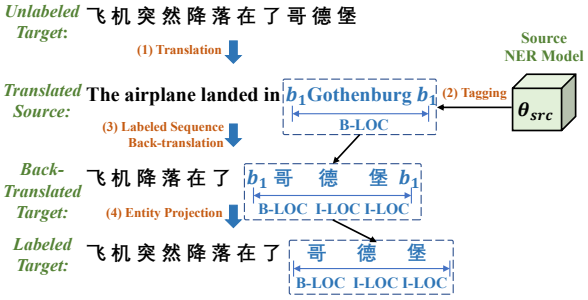


Figure 1: Illustration of our method. It enables cross-lingual zero-shot transfer from source (English) to target (Chinese) language via labeled sequence translation and then entity projection.

Zero-shot cross-lingual NER aims to address this challenging problem by transferring knowledge from the high-resource source language with lots amounts 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 **Cro**ss-lingual Entity

[*]Contribution during internship at Microsoft Research Asia.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 486–496
December 7-11, 2022

Projection (CROP) 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 Θ_{ner}^{src} 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:

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

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 Θ_{ner}^{src} 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 CROP

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 $D_{x,t}^{L_{src}} = \{(x^{(i)}, t^{(i)})\}_{i=1}^{N}$ 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 $\{D_x^{L_k}\}_{k=1}^{K}$ 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 $\{D_{x,f(x)}^{L_k}\}_{k=1}^{K}$, where $f(x)$ is projected label of the sentence x by a simple lexical matching between translated entities and original words. The source corpus $D_{x,t}^{L_{src}}$ and target annotated corpora $\{D_{x,f(x)}^{L_k}\}_{k=1}^{K}$ 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 Θ_e to extract top-layer features:

$$H = \text{Encoder}(x; \Theta_e) \quad (2)$$

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. Θ_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:

$$P(t|x) = \text{Softmax}(W_c H + b_c) \quad (3)$$

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

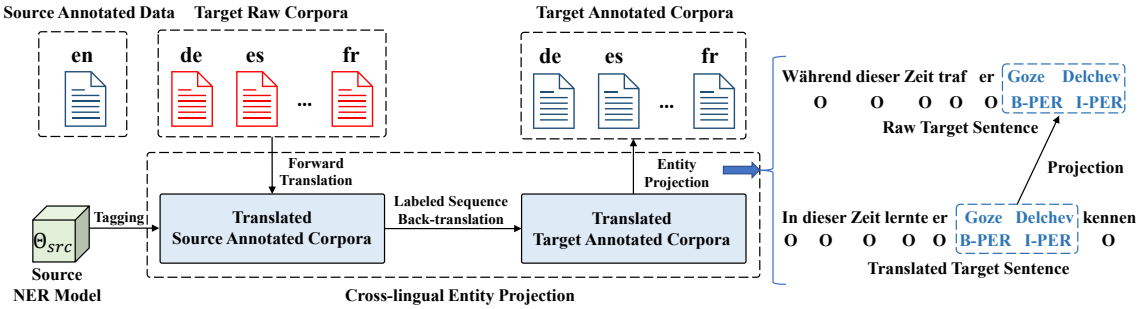


Figure 2: Framework of our proposed method CROP, which projects the labels of the translated entities into the target raw data by the multilingual forward translation and labeled sequence back-translation.

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 $x_p = (x_1, \ldots, b_i, x_{u_1:u_2}, b_i, \ldots, x_m)$ and $y_p = (y_1, \ldots, b_i, y_{v_1:v_2}, b_i, \ldots, y_n)$, where $y_{v_1:v_2}$ is the target translation of source piece $x_{u_1:u_2}$. $x_{u_1:u_2}$ denotes the source phrase from the u_1-th token to the u_2-th token and $y_{v_1:v_2}$ 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 `efloma`[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 $D_b = \{D_b^{L_k}\}_{k=1}^K$ of K languages, where one side is the source language L_{src} and the other side is the language $L_k \in L_{tgt}$, the multilingual model is trained on corpora D_b:

$$\mathcal{L}_t = -\sum_{k=1}^{K} \mathbb{E}_{x,y \in D_b^{L_k}} [\log P(y|x; \Theta_{mt})] \quad (4)$$

where Θ_{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:

$$\mathcal{L}_{lst} = -\sum_{k=1}^{K} \mathbb{E}_{x,y \in D_b^{L_i}} [\log P(y_p|x_p; \Theta_{mt})] \quad (5)$$

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:

$$\mathcal{L}_{mt} = \alpha \mathcal{L}_t + (1 - \alpha)\mathcal{L}_{lst} \quad (6)$$

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 Θ_{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 $y_{v_1:v_2}$ have the same NER labels with the source phrase $x_{u_1:u_2}$, where both phrases are surrounded by the boundary token b_i.

3.4 Cross-lingual Entity Projection

Given the labeled corpus $D_{x,t}^{L_{src}} = \{(x^{(i)}, t^{(i)})\}_{i=1}^N$ of the source language L_{src} and the unlabeled corpora $\{D_x^{L_k}\}_{k=1}^K$ of K languages, the source NER

[1] https://github.com/robertostling/eflomal

model Θ_{ner}^{src} is used to tag the unlabeled training corpora of target languages, aided by the labeled translation model Θ_{mt}.

Forward Translation The target raw corpora $\{D_x^{L_k}\}_{k=1}^K$ of K languages are translated into the source language and tagged by the source NER model Θ_{ner}^{src}. We obtain the source labeled translated corpora $\{D_{x,f(x)}^{L_k^{src}}\}_{k=1}^K$, where $f(\cdot)$ is the predictor of the source NER model Θ_{ner}^{src}.

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 $\{D_{x,f(x)}^{L_k^{bt}}\}_{k=1}^K$ by the translation model Θ_{mt}.

Entity Matching Given the target translated entities with labels D_{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 $\{D_{x,f(x)}^{L_k}\}_{k=1}^K$. Finally, the target annotated corpora and the original corpus $D_{x,t}^{L_{src}}$ are used for multilingual NER model training.

3.5 Self-training

Given a labeled corpus $D_{x,t}^{L_{src}}$ of the source language and target unlabeled corpora $\{D_x^{L_k}\}_{k=1}^K$ of target languages, the training objective based on $D_{x,t}^{L_{src}}$ is formulated as below:

$$\mathcal{L}_{src} = \mathbb{E}_{x,t \in D_{x,t}^{L_{src}}} \left[-\log P(t|x; \Theta_{ner}^{all}) \right] \quad (7)$$

where Θ_{ner}^{all} are NER model parameters.

Then, we leverage the source NER model Θ_{src}^{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 $\{D_{x,f(x)}^{L_k}\}_{k=1}^K$. The multilingual corpora of target languages with predicted labels are adopted to train a neural network with the combined loss function $\mathcal{L}_{tgt}$ as below:

$$\mathcal{L}_{tgt} = \sum_{L_k \in L_{tgt}} \mathbb{E}_{x,y \in D_{x,f(x)}^{L_k}} \left[-\log P(f(x)|x; \Theta_{ner}^{all}) \right] \quad (8)$$

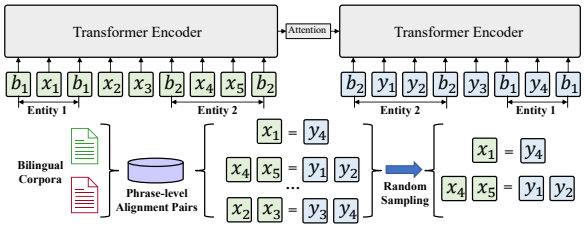


Figure 3: Multilingual labeled sequence translation.

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:

$$\mathcal{L}_{all} = \mathcal{L}_{src} + \mathcal{L}_{tgt} \quad (9)$$

where $\mathcal{L}_{src}$ and $\mathcal{L}_{tgt}$ 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[2] 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 Θ_{ner}^{src} trained on the original dataset and the multilingual NER model Θ_{ner}^{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 pretrained multilingual model M2M$_{large}$[3] 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[4], where the parallel data is paired with English and other 39 languages and the phrase-level alignment pairs are extracted by the alignment tool`eflomal`. 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 Θ_{ner}^{src} trained on the original dataset and the multilingual NER model Θ_{ner}^{all} trained by self-training in Equation 9 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.

[2] https://github.com/saffsd/langid.py

[3] https://dl.fbaipublicfiles.com/flores101/pretrained_models/flores101_mm100_615M.tar.gz

[4] https://opus.nlpl.eu/CCAligned.php

Language	Type	Train	Dev	Test
English (en)	#Sentences	15.0K	3.5K	3.7K
(CoNLL-2003)	#Entities	23.5K	6.0K	5.7K
German (de)	#Sentences	12,7K	3.1K	3.2K
(CoNLL-2003)	#Entities	11.9K	4.8K	3.7K
Spanish [es]	#Sentences	8.3K	1.9K	1.5K
(CoNLL-2002)	#Entities	18.8K	4.3K	3.6K
Dutch [nl]	#Sentences	15.8K	2.9K	5,2K
(CoNLL-2002)	#Entities	13.3K	2.6K	3,9K
Norwegian [no]	#Sentences	15.7K	2.4K	1.9K
(NoDaLiDa-2019)	#Entities	10.9K	1.6K	1.4K

Table 1: Statistics of the CoNLL-5 (Sang, 2002; Sang and Meulder, 2003) and NoDaLiDa (Johansen, 2019) NER benchmarks.

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

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.

XTREME-40 Table 3 compares the performance of our method with previous relevant methods initialed 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 SacreBLEU[5] of 39 translation directions except eu languages.

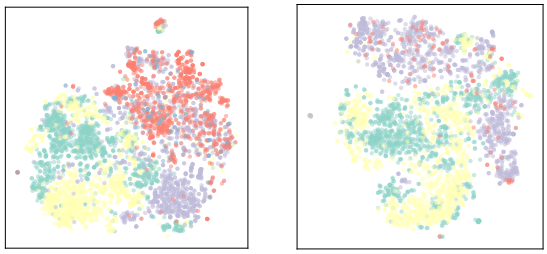
[5] https://github.com/ngoyal2707/sacrebleu

Initialized By Pretrained Cross-lingual Language Model mBERT																					
Method	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	ka
mBERT (Devlin et al., 2019)	76.9	44.5	77.1	68.8	78.8	71.6	74.0	76.3	68.0	48.2	77.2	79.7	56.5	66.9	76.0	46.3	**81.1**	28.9	**66.4**	67.7
+Translate Train	74.5	37.6	77.8	73.2	77.2	74.9	69.4	74.1	63.2	43.1	75.9	76.1	55.4	68.1	77.2	**48.2**	77.2	36.6	55.1	64.4
UniTrans (Wu et al., 2020a)	78.2	47.0	79.5	74.6	79.8	75.6	75.2	76.5	67.2	49.3	75.6	80.1	58.4	72.1	77.9	44.6	78.3	37.6	56.2	69.9
CROP (Our Method)	**81.0**	**48.0**	**80.8**	**74.9**	**80.3**	**78.7**	**84.2**	**78.3**	**70.6**	**63.2**	**79.1**	**83.5**	**64.7**	**77.1**	**82.5**	46.4	79.9	**45.3**	57.7	**74.1**
Method	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh	**Avg**$_{all}$
mBERT (Devlin et al., 2019)	50.4	60.2	53.7	56.2	61.9	47.6	82.1	79.6	65.2	72.8	50.8	46.8	0.4	71.2	75.5	36.9	69.7	51.7	44.1	61.7
+Translate Train	48.2	61.2	61.0	58.7	67.5	57.3	79.6	78.4	61.2	**69.2**	62.7	51.2	2.4	72.7	72.6	58.9	69.5	51.1	45.3	62.3
UniTrans (Wu et al., 2020a)	52.5	61.4	63.5	62.3	65.8	59.2	82.4	80.3	64.8	65.2	63.2	56.1	3.1	73.4	77.9	64.1	69.7	50.1	47.4	64.5
CROP (Our Method)	**54.9**	**62.6**	**72.7**	**70.6**	**71.1**	**61.3**	**84.6**	**81.7**	**69.7**	68.3	**64.9**	**61.6**	**3.9**	**76.9**	**80.4**	**78.0**	**70.0**	**51.8**	**54.4**	**68.4**
Initialized By Pretrained Cross-lingual Language Model XLM-R																					
Method	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	ka
XLM-R (Conneau et al., 2020)	74.6	46.0	78.0	68.3	75.2	75.7	70.2	72.2	59.9	52.0	75.8	76.6	52.4	69.6	78.2	47.4	77.7	21.0	61.8	66.5
+Translate Train	76.2	47.8	79.2	74.3	75.8	67.7	68.4	75.8	61.2	41.0	76.8	76.4	55.0	71.9	76.0	**50.6**	78.1	35.4	54.7	68.4
UniTrans (Wu et al., 2020a)	78.1	**48.1**	79.3	74.6	75.2	74.9	73.8	76.9	62.7	49.2	74.6	76.5	53.4	70.4	76.9	48.6	77.3	21.6	62.2	66.8
CROP (Our Method)	**80.3**	45.2	**80.4**	**75.7**	**79.6**	**78.5**	**83.1**	**77.2**	**66.8**	**65.5**	**77.9**	**82.9**	**63.5**	**77.4**	**81.6**	46.1	**78.8**	**45.4**	**63.2**	**74.0**
Method	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh	**Avg**$_{all}$
XLM-R (Conneau et al., 2020)	43.2	49.9	62.3	59.6	67.3	53.5	80.2	78.1	64.3	**70.3**	55.0	50.1	3.0	69.4	78.1	63.6	68.2	47.5	27.7	61.3
+Translate Train	40.1	55.5	60.0	59.8	69.8	61.6	79.6	76.4	60.9	70.0	63.7	50.7	3.4	74.7	72.3	62.7	69.6	46.8	41.2	62.3
UniTrans (Wu et al., 2020a)	46.5	57.2	65.5	64.5	70.2	62.6	81.8	79.4	68.8	68.9	65.1	56.1	4.8	74.8	76.4	71.0	69.8	**55.1**	44.4	64.2
CROP (Our Method)	**50.2**	**59.8**	**73.8**	**71.6**	**71.8**	**69.0**	**83.5**	**81.5**	**70.2**	69.0	**65.6**	**59.9**	3.1	**75.5**	**80.5**	**80.4**	**70.1**	52.6	**50.3**	**68.2**

Table 3: Results of our proposed method CROP and other relevant baselines for cross-lingual NER. "Avg$_{all}$" represents the average F1 scores of all 39 languages on the test set of the XTREME-40 benchmark.

ID	**Method**	es	eu	ta	tl	zh	Avg$_{all}$
①	CROP	**83.1**	**66.8**	**65.6**	**75.5**	**50.3**	**68.2**
②	① - XLM-R	72.6	62.6	55.9	72.2	38.9	58.1
③	② - Transfer	20.0	14.7	6.0	33.7	1.6	15.1

Table 4: Ablation study of our proposed method. Avg$_{all}$ denotes the average F1 scores of 39 languages.



(a) Gold (b) Translation

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.

	Avg$_{X\rightarrow En}$	Avg$_{En\rightarrow X}$	Avg$_{all}$
M2M (Goyal et al., 2021)	24.50	22.08	23.61
Our Multilingual Model	**32.70**	**30.31**	**31.51**

Table 5: Comparison of BLEU points between M2M (Goyal et al., 2021) and our multilingual model on the FLORES-101 benchmark of 39 languages.

	Alignment Pairs	En→Zh	En→De	En→Fr
Our Multilingual Model		84.4%	84.8%	86.8%
	✓	**97.2%**	**95.6%**	**94.8%**

Table 6: Comparison of labeled translation quality between our multilingual model with alignment pairs and the counterpart without alignment information.

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 and human evaluation. More specifically, we check whether the boundary symbol surrounds the equivalent entity in both source and target sentences. The baseline model encounters the problem of boundary symbol missing and incorrect alignment. In Table 6, our model guided by the phrase-level alignment information outperforms the baseline model showing the strength of our method.

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

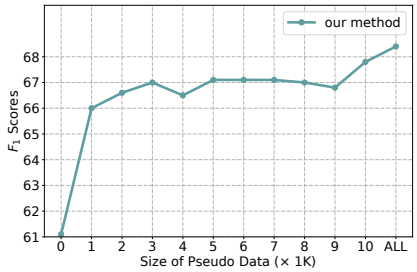


Figure 5: Evaluation results on the original source annotated corpus and pseudo corpora with different training sizes by randomly down-sampling.

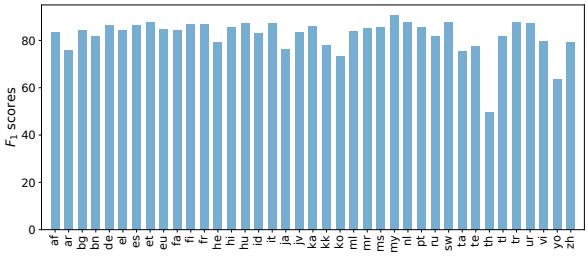


Figure 6: F1 scores of cross-lingual entity projection based on the golden labels. The languages are ordered by alphabet order.

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{i}__` as the boundary symbol b_i. The entities are surrounded by "`__SLOT{i}__`" for translation, where "`__SLOT{i}__`" is used as the boundary symbol. The positions of the boundary symbols "`__SLOT0__`" and "`__SLOT1__`" are misplaced during translation for the baseline model.

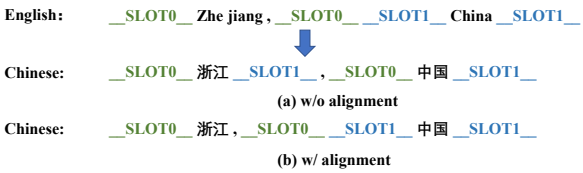


Figure 7: Comparison between the multilingual model w/o the alignment information and the counterpart w/ the alignment information in the training.

	de	fr	et	Avg$_{sim}$	ja	ta	zh	Avg$_{dis}$
XLM-R	75.2	76.6	72.2	74.7	21.0	55.0	27.7	34.6
+10K	77.7	82.2	76.1	78.7	42.4	64.5	48.9	51.3
+50K	78.8	82.4	76.8	79.3	44.6	65.3	49.4	53.1
+100K	78.4	82.5	77.0	79.3	45.0	65.4	49.3	53.2
+ALL	**79.6**	**82.6**	**77.2**	**79.8**	**45.5**	**65.6**	**50.3**	**53.8**

Table 7: Evaluation results for similar and distant languages of the source language with different sizes of pseudo data. Avg$_{sim}$ and Avg$_{dis}$ represent the average F1 scores of similar languages and distant languages.

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 Avg$_{sim}$ and Avg$_{dis}$. In Figure 7, Avg$_{sim}$ gains +5.1 points improvement while Avg$_{dis}$ 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

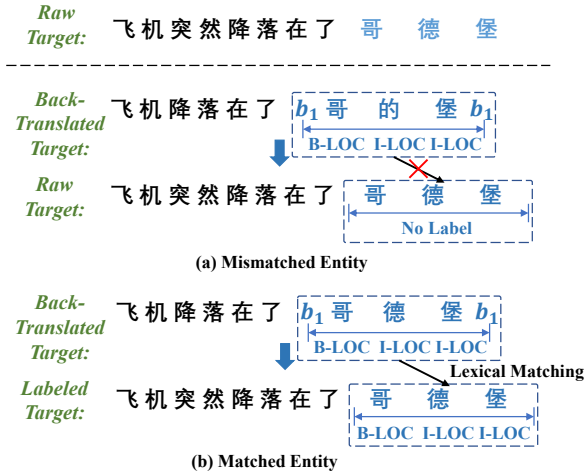


Figure 8: Entity matching includes (a) mismatched entity and (b) matched entity.

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 predefined 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 pre-trained 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

Diego Mollá Aliod, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *ALTA 2006*, pages 51–58.

Rami Aly, Andreas Vlachos, and Ryan McDonald. 2021. Leveraging type descriptions for zero-shot named entity recognition and classification. In *ACL 2021*, pages 1516–1528.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Partha Sarathy Banerjee, Baisakhi Chakraborty, Deepak Tripathi, Hardik Gupta, and Sourabh S. Kumar. 2019. A information retrieval based on question and answering and NER for unstructured information without using SQL. *WPC 2019*, 108(3):1909–1931.

M. Saiful Bari, Shafiq R. Joty, and Prathyusha Jwalapuram. 2019. Zero-resource cross-lingual named entity recognition. *CoRR*, abs/1911.09812.

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 *ACL 2020*, 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 2019*, pages 4171–4186.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2020. DAGA: data augmentation with a generation approach for low-resource tagging tasks. *CoRR*, abs/2011.01549.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. Ccaligned: A massive collection of cross-lingual web-document pairs. In *EMNLP 2020*, pages 5960–5969.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Zheng Fang, Yanan Cao, Tai Li, Ruipeng Jia, Fang Fang, Yanmin Shang, and Yuhai Lu. 2021. TEBNER: Domain specific named entity recognition with type expanded boundary-aware network. In *EMNLP 2021*, pages 198–207.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke S. Zettlemoyer. 2013. Joint coreference resolution and named-entity linking with multi-pass sieves. In *EMNLP 2013*, pages 289–299.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. In *EMNLP 2019*, pages 1083–1092.

Bjarte Johansen. 2019. Named-entity recognition for norwegian. In *NODALIDA 2019*, pages 222–231.

Yichong Leng, Xu Tan, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2019. Unsupervised pivot translation for distant languages. In *ACL 2019*, pages 175–183.

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 *ACL 2021*, pages 200–209.

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.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. RockNER: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *EMNLP 2021*, pages 3728–3737.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *ACL 2020*, pages 1518–1533.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2021. Mulda: A multilingual data augmentation framework for low-resource cross-lingual NER. In *ACL 2021*, pages 5834–5846.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *JMLR*, 9:2579–2605.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *EMNLP 2017*, pages 2536–2545.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *ACL 2020*, pages 8093–8104.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *ACL 2017*, pages 1470–1480.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *ACL 2021*, pages 244–258.

Piotr Przybyla. 2016. Boosting question answering by deep entity recognition. *CoRR*, abs/1605.08675.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *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 *CoNLL 2003*, pages 142–147.

Kyle Shaffer. 2021. Language clustering for multilingual named entity recognition. In *EMNLP 2021*, pages 40–45.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *ICLR 2017*.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *NAACL 2019*, pages 449–459.

Oscar Täckström, Ryan T. McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL 2012*, pages 477–487.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *CoNLL 2016*, pages 219–228.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS 2017*, pages 5998–6008.

Rui Wang and Ricardo Henao. 2021. Unsupervised paraphrasing consistency training for low resource named entity recognition. In *EMNLP 2021*, pages 5303–5308.

Yucheng Wang, Bowen Yu, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun. 2021. Discontinuous named entity recognition as maximal clique discovery. In *ACL 2021*, pages 764–774.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jianguang Lou. 2020a. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *IJCAI 2020*, pages 3926–3932.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. 2020b. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *AAAI 2020*, pages 9274–9281.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *EMNLP 2019*, pages 833–844.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime G. Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *EMNLP 2018*, pages 369–379.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. In *ACL 2021*, pages 5725–5737.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *COLING 2018*, pages 2145–2158.

Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2021. Multilingual agreement for multilingual neural machine translation. In *ACL 2021*, pages 233–239.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. CSP: code-switching pre-training for neural machine translation. In *EMNLP 2020*, pages 2624–2636.

Mingliang Zhang, Fandong Meng, Yunhai Tong, and Jie Zhou. 2021. Competence-based curriculum learning for multilingual machine translation. In *EMNLP 2021*, pages 2481–2493.

Chunting Zhou, Daniel Levy, Xian Li, Marjan Ghazvininejad, and Graham Neubig. 2021a. Distributionally robust multilingual machine translation. In *EMNLP 2021*, pages 5664–5674.

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