MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER

Linlin Liu*[1][2] Bosheng Ding*[1][2] Lidong Bing[2] Shafiq Joty[1] Luo Si[2] Chunyan Miao[1]

[1]Nanyang Technological University, Singapore [2]DAMO Academy, Alibaba Group

{linlin.liu, bosheng.ding, l.bing, luo.si}@alibaba-inc.com

{srjoty, ascymiao}@ntu.edu.sg

Abstract

Named Entity Recognition (NER) for low-resource languages is a both practical and challenging research problem. This paper addresses zero-shot transfer for cross-lingual NER, especially when the amount of source-language training data is also limited. The paper first proposes a simple but effective labeled sequence translation method to translate source-language training data to target languages and avoids problems such as word order change and entity span determination. With the source-language data as well as the translated data, a generation-based multilingual data augmentation method is introduced to further increase diversity by generating synthetic labeled data in multiple languages. These augmented data enable the language model based NER models to generalize better with both the language-specific features from the target-language synthetic data and the language-independent features from multilingual synthetic data. An extensive set of experiments were conducted to demonstrate encouraging cross-lingual transfer performance of the new research on a wide variety of target languages.[1]

1 Introduction

Named entity recognition (NER) aims to identify and classify entities in a text into predefined types, which is an essential tool for information extraction. It has also been proven to be useful in various downstream natural language processing (NLP) tasks, including information retrieval (Banerjee et al., 2019), question answering (Fabbri et al., 2020) and text summarization (Nallapati et al., 2016). However, except for some resource-rich languages (e.g., English, German), training sets for most of the other languages are still very limited. Moreover, it is usually expensive and time-consuming to annotate such data, particularly for low-resource languages (Kruengkrai et al., 2020). Therefore, zero-shot cross-lingual NER has attracted growing interest recently, especially with the influx of deep learning methods (Mayhew et al., 2017; Joty et al., 2017; Jain et al., 2019; Bari et al., 2021).

Existing approaches to cross-lingual NER can be roughly grouped into two main categories: instance-based transfer via machine translation (MT) and label projection (Mayhew et al., 2017; Jain et al., 2019), and model-based transfer with aligned cross-lingual word representations or pretrained multilingual language models (Joty et al., 2017; Baumann, 2019; Wang et al., 2020; Conneau et al., 2020; Bari et al., 2021). Recently, Wu et al. (2020) unify instance-based and model-based transfer via knowledge distillation.

These recent methods have demonstrated promising zero-shot cross-lingual NER performance. However, most of them assume the availability of a considerable amount of training data in the source language. When we reduce the size of the training data, we observe significant performance decrease. For instance-based transfer, decreasing training set size also amplifies the negative impact of the noise introduced by MT and label projection. For model-based transfer, although the large-scale pretrained multilingual language models (LM) (Conneau et al., 2020; Liu et al., 2020) have achieved state-of-the-art performance on many cross-lingual transfer tasks, simply fine-tuning them on a small training set is prone to over-fitting (Wu et al., 2018; Si et al., 2020; Kou et al., 2020).

To address the above problems under the setting of low-resource cross-lingual NER, we propose a multilingual data augmentation (MulDA) framework to make better use of the cross-lingual

*Equal contribution, order decided by coin flip. Linlin Liu and Bosheng Ding are under the Joint PhD Program between Alibaba and Nanyang Technological University.

[1] Our code is available at https://ntunlpsg.github.io/project/mulda/.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5834–5846
August 1–6, 2021. ©2021 Association for Computational Linguistics

generalization ability of the pretrained multilingual LMs. Specifically, we consider a low-resource setting for cross-lingual NER, where there is very limited source-language training data and no target-language train/dev data. Such setting is practical and useful in many real scenarios.

Our proposed framework seeks the initial help from the instance-based transfer (i.e., translate train) paradigm (Li et al., 2020; Fang et al., 2020). We first introduce a novel labeled sequence translation method to translate the training data to the target language as well as to other languages. This allows us to finetune the LM based NER model on multilingual data rather than on the source-language data only, which helps prevent over-fitting on the language-specific features. One commonly used tool for translation is the off-the-shelf Google translate system[2], which supports more than 100 languages. Alternatively, there are also many pretrained MT models conveniently accessible, e.g., more than 1,000 MarianMT (Junczys-Dowmunt et al., 2018; Kim et al., 2019) models have been released on the Hugging Face model hub.[3]

Note that the instance-based transfer methods add limited semantic variety to the training set, since they only translate entities and the corresponding contexts to a different language. In contrast, data augmentation has been proven to be a successful method for tackling the data scarcity problem. Inspired by a recent monolingual data augmentation method (Ding et al., 2020), we propose a generation-based multilingual data augmentation method to increase the diversity, where LMs are trained on multilingual labeled data and then used to generate more synthetic training data.

We conduct extensive experiments and analysis to verify the effectiveness of our methods. Our main contributions can be summarized as follows:

- We propose a simple but effective labeled sequence translation method to translate the source training data to a desired language. Compared with exiting methods, our labeled sentence translation approach leverages placeholders for label projection, which effectively avoids many issues faced during word alignment, such as word order change, entity span determination, noise-sensitive similarity metrics and so on.

- We propose a generation-based multilingual data augmentation method for NER, which leverages the multilingual language models to add more diversity to the training data.

- Through empirical experiments, we observe that when fine-tuning pretrained multilingual LMs for low-resource cross-lingual NER, translations to more languages can also be used as an effective data augmentation method, which helps improve performance of both the source and the target languages.

2 MulDA: Our Multilingual Data Augmentation Framework

We propose a multilingual data augmentation framework that leverages the advantages of both instance-based and model-based transfer for cross-lingual NER. In our framework, a novel labeled sequence translation method is first introduced to translate the annotated training data from the source language S to a set of target languages $\mathcal{T} = \{T_1, \dots, T_n\}$. Then language models are trained on $\{\mathcal{D}^S, \mathcal{D}^{T_1}, ..., \mathcal{D}^{T_n}\}$ to generate multilingual synthetic data, where $\mathcal{D}^S$ is the source-language training data, and $\mathcal{D}^{T_i}$ is the translated data in language T_i. Finally, we post-process and filter the augmented data to train multilingual NER models for inference on target-language test sets.

2.1 Labeled Sequence Translation

We leverage labeled sequence translation for the training data of the source language to generate multilingual NER training data, which can also be viewed a method for data augmentation. Prior methods (Jain et al., 2019; Li et al., 2020) usually perform translation and label projection in two separate steps: 1) translate source-language training sentences to the target language; 2) propagate labels from the source training data to the translated sentences via word-to-word/phrase-to-phrase mapping with alignment models or algorithms. However, these methods suffer from a few label projection problems, such as word order change, word-span determination (Li et al., 2020), and so on. An alternative to avoid the label projection problems is word-by-word translation (Xie et al., 2018), but often at the sacrifice of the translation quality.

We address the problems identified above by first replacing named entities with contextual placeholders before sentence translation, and then after translation, we replace placeholders in translated

[2] https://cloud.google.com/translate

[3] https://huggingface.co/transformers/model_doc/marian.html

Labeled sentence in the source language:
[PER Jamie Valentine] was born in [LOC London].

1. Translate sentence with placeholders:
src: PER0 was born in LOC1.
tgt: PER0 nació en LOC1.

2. Translate entities with context:
PER0
src: [Jamie Valentine] was born in London.
tgt: [Jamie Valentine] nació en Londres.

LOC1
src: Jamie Valentine was born in [London].
tgt: Jamie Valentine nació en [Londres].

3. Replace placeholders with translated entities:
[PER Jamie Valentine] nació en [LOC Londres].

Figure 1: An example of labeled sentence translation, where **src** and **tgt** are the translation model inputs and outputs, respectively. For the example shown in this figure, Google translation system and the MarianMT model generate the same translations in step 1 and 2.

B-PER	E-PER	O O O	S-LOC	O
Jamie	Valentine	was born in	London	.

⇓ Linearization

B-PER Jamie E-PER Valentine was born in S-LOC London .

Figure 2: An example of labeled sequence linearization.

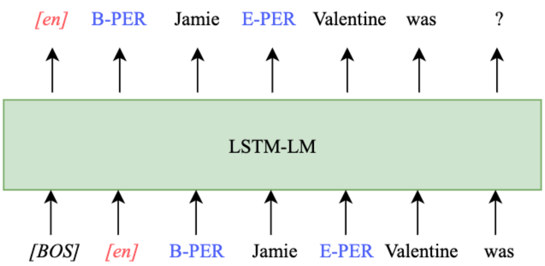

Figure 3: Training of multilingual LSTM-LM on the linearized sequences.

sentences with the corresponding translated entities. An illustration of the method is shown in Figure 1.

Assume a sentence $X^S = \{x_1, \dots, x_M\} \in \mathcal{D}^S$ and the corresponding NER tags $\{y_1, \dots, y_M\}$ are given, where x_i's are the sentence tokens and M is the sentence length. Let $\{E_1, \dots, E_n\}$ denote the predefined named entity types. Our method first replaces all entities in $\{x_1, \dots, x_M\}$ with placeholders (**src** of step 1 in Figure 1). Placeholders Ek are reconstructed tokens with the corresponding entity type E as prefix and the index of the entity k as suffix. Assume $\{x_i, \dots, x_j\}$ is the k_{th} entity in the source sentence, and the corresponding type is E_z, then we can replace the entity with the placeholder E_zk to get $\{\dots, x_{i-1}, E_zk, x_{j+1}, \dots\}$. We use X^S_* to denote the generated sentence after replacing all entities with placeholders. X^S_* is fed into an MT model to get the translation X^T_* in the target language T. With such design, the placeholder prefix E can provide the MT model[4] with relevant contextual information about the entities, so that the model can translate the sentence with reasonably good quality. Besides, we observe most of placeholders are unchanged after translation,[5] which can be used to help locate the position of entities.

In the second step, we translate each entity with the corresponding context. More specifically, we use brackets to mark the span of each entity and translate it to the target language successively, one at a time (**src** of step 2 in Figure 1). For example, to translate entity $\{x_i, \dots, x_j\}$, we feed $\{\dots, x_{i-1}, [x_i, \dots, x_j], x_{j+1}, \dots\}$ into the MT model. Then we can get entity translations by extracting the square bracket marked tokens from the translated sentences. We translate the entities directly if the square brackets are not found.

Finally, we can replace placeholders in X^T_* (obtained from the first step) with the corresponding entity translations (obtained from the second step) and copy placeholder prefix as entity labels to generate the synthetic training data in the target language (step 3 in Figure 1). We tested the proposed method with Google translate and the MarianMT (Junczys-Dowmunt et al., 2018; Kim et al., 2019) models, and we found that both produce high quality synthetic data as we had expected.

2.2 Synthetic Data Generation with Language Models

Although labeled sequence translation generates high quality multilingual NER training data, it adds limited variety since translation does not introduce new entities or contexts. Inspired by DAGA (Ding et al., 2020), we propose a generation-based multilingual data augmentation method to add more diversity to the training data. DAGA is a monolingual data augmentation method designed for sequence labeling tasks, which has been shown to

[4] When the MT model use subword vocabularies.

[5] See Appendix for more examples.

be able to add significant diversity to the training data. As the example shown in Figure 2, it first linearizes labeled sequences by adding the entity type before sentence tokens. Then an LSTM-based LM (LSTM-LM) is trained on the linearized sequences in an autoregressive way, after which the begin-of-sentence token *[BOS]* is fed into the LSTM-LM to generate synthetic training data autoregressively. The monolingual LSTM-LM of DAGA is trained in a similar way as the example shown in Figure 3, except that there is no language tag *[en]*.

To extend this method for multilingual data augmentation, we add special tokens at the beginning of each sentence to indicate the language that it belongs to. The source-language data and the multilingual data obtained via translation are concatenated to train/finetune multilingual LMs with a shared vocabulary (as shown in Figure 5). Given a labeled sequence $\{x_1, \ldots, x_M\}$ from the multilingual training data, the LMs are trained to maximize the probability $p(x_1, \ldots, x_M)$ in Eq. 1:

$$p(x_1, \ldots, x_M) = \prod_{t=1}^{M} p_\theta(x_t | x_{<t}) \qquad (1)$$

where θ is the parameter to optimize, and $p_\theta(x_t|x_{<t})$ is the probability of the next token given the previous tokens in the sequence, which is usually computed with the softmax function. Figure 3 shows an example of how the multilingual LSTM-LM is trained in the autoregressive way. After training the LSTM-LM, we can feed the *[BOS]* token and a language token to the model to generate synthetic training data for the specified language.

Besides, to leverage the cross-lingual generalization ability of large scale pretrained multilingual LMs, we also finetune a recent state-of-the-art seq2seq model mBART (Liu et al., 2020), which is pretrained with multilingual denoising tasks. Sentence permutation and word-span masking are the two noise injection methods used to add noise to original sentence $X = \{x_1, \ldots, x_M\}$ to output $g(X)$, where $g(.)$ is used to denote the noise injection function. After encoding $g(X)$ with the Transformer encoder, the Transformer decoder is trained to generate the original sequence X autoregressively by maximizing Eq. 1.

Denoising word-span masked sequences is the most relevant to our data augmentation method, since only small modifications are required to make our finetuning task as consistent to the pretraining task as possible. More specifically, we design our finetuning task with the following changes: 1) use the linearized labeled sequences (as shown in Figure 5) as input X; 2) modify $g(.)$ to mask random trailing sub-sequences such that $g(X) = \{x_1, \ldots, x_z, [mask]\}$, where $1 \leq z \leq |X|$ is a random integer. After finetuning with such task, we can conveniently feed a randomly masked sequence $\{x_1, \ldots, x_z, [mask]\}$ into mBART to generate synthetic data. Figure 4 shows a more concrete example to illustrate how mBART is finetuned with the linearized sequences in our work.

2.3 Semi-supervised Method

Unlabeled multilingual sentences are usually easy to get, for example, data from the Wikimedia[6]. To make better use of these unlabeled multilingual data, we propose a semi-supervised method to prepare more pseudo labeled data for finetuning multilingual LMs. Inspired by self-training (Zoph et al., 2020; Xie et al., 2020), we use the NER model trained on the multilingual translated data to annotate the unlabeled sentences. After that, we use two additional NER models trained with different random seeds to filter the annotated data by removing those with different tag predictions.

2.4 Post-Processing

We also design several straightforward methods to post-process and filter the augmented data generated by the LMs:

- Delete sequences that contain only O (other) tags.
- Convert the generated labeled sequences to the same format as gold data by separating sentence tokens and NER tags.
- Use the NER model trained on the multilingual translated data to label the generated sequences (after tag removal). Then compare the tags generated by the LM and NER model predictions, and remove the sentences with inconsistencies.

3 Experiments

We conduct experiments to evaluate the effectiveness of the proposed multilingual data augmentation framework. Firstly, we compare our labeled sequence translation method with the previous instance-based transfer (i.e., translate train) methods. Following that, we show the benefit of adding multilingual translations. Then we continue

[6] https://dumps.wikimedia.org/

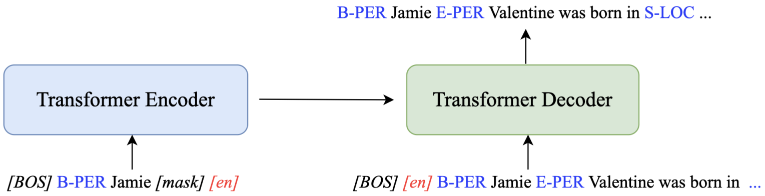
B-PER Jamie E-PER Valentine was born in S-LOC ...
Transformer Encoder
Transformer Decoder
[BOS] B-PER Jamie [mask] [en]
[BOS] [en] B-PER Jamie E-PER Valentine was born in ...

Figure 4: Finetune mBART with the linearized sequences. The transformer decoder is trained to generate labeled sequences autoregressively. Following the mBART pretraining tasks, we add language tokens at the end of masked sequences when feed them into encoder.

[BOS] [en] B-PER Jamie E-PER Valentine was born in S-LOC London.
[BOS] [de] B-PER Jamie E-PER Valentine wurde in S-LOC London geboren.
[BOS] [es] B-PER Jamie E-PER Valentine nació en S-LOC Londres.
[BOS] [nl] B-PER Jamie E-PER Valentine werd geboren in S-LOC Londen.
...

Figure 5: The source-language data and the multilingual data obtained via translation are concatenated to train/finetune multilingual LMs.

to evaluate the generation-based multilingual data augmentation method by comparing cross-lingual NER performance of the models trained on monolingual, bilingual, and multilingual augmented data, respectively. Finally, we further evaluate our methods on a wider range of distant languages.

We use the most typical Transformer-based NER model[7] in our experiments, which is implemented by adding a randomly initialized feed forward layer to the Transformer final layer for label classification. Specifically, to demonstrate that our framework can help achieve additional performance gain even on the top of the state-of-the-art multilingual LMs, the checkpoint of the pretrained XLM-R large (Conneau et al., 2020) model is used to initialize our NER models.

3.1 Labeled Sequence Translation

We finetune the NER model on the translated target-language data to compare our labeled sequence translation method (§2.1) with the existing instance-based transfer methods.

Experimental settings The CoNLL02/03 NER dataset (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) is used for evaluation, which contains data in four different languages: English, German, Dutch and Spanish. All of the data are annotated with the same set of NER tags. We follow the steps described in §2.1 to translate English train data to the other three languages. Following Jain et al. (2019) and Li et al. (2020), Google translation system is used in the experiments. Since our NER model is more powerful than those used by Jain et al. (2019) and Li et al. (2020), we reproduce their results with XLM-R large for a fair comparison. All of the NER models are finetuned on the translated target-language sentences only for 10 epochs with the best model selected using English dev data, and then evaluated on the target-language original test data.

Method	de	es	nl	avg
Mayhew et al. (2017)	60.1	65.0	67.6	64.23
Xie et al. (2018)	57.8	72.4	70.4	66.87
Jain et al. (2019)	61.5	73.5	69.9	68.30
Bari et al. (2020)	65.24	**75.93**	74.61	71.93
Li et al. (2020)†	66.90	70.49	73.46	70.28
Jain et al. (2019)†	70.99	74.64	76.63	74.09
ours	**73.89**	75.48	**79.60**	**76.32**

Table 1: Cross-lingual NER performance of the instance-based transfer methods. † denotes the reproduced results with XLM-R large.

Results We present the results in Table 1. As we can see, our method outperforms the best baseline method by 2.90 and 2.97 on German and Dutch respectively, and by 2.23 on average. Since our models are only finetuned with the data generated by the labeled sequence translation method, the results directly demonstrate the effectiveness of our method. Moreover, compared with the two recent baseline methods (Jain et al., 2019; Li et al., 2020), our method does not rely on complex label projection algorithms and is much easier to implement.

3.2 Multilingual Translation as Data Augmentation

After showing that our labeled sequence translation method can generate high quality labeled data in the target language, in this section, we run ex-

[7] Similar to the token classification model in https://github.com/huggingface/transformers.

periments to verify the hypothesis that multilingual translation may help improve the cross-lingual transfer performance of multilingual LMs in low resource scenarios.

Experimental settings We use the same NER dataset as above. In order to simulate low resource scenarios, we randomly sample 500, 1k and 2k sentences from the gold English train set. Our labeled sequence translation method is used to translate the sampled data to pseudo labeled data in the three target languages, German, Spanish and Dutch. To better demonstrate how the training data affects cross-lingual NER performance, we train the NER model on four different conditions: 1) **En**: train the models on English data only; 2) **Tgt-Tran**: train the models on the pseudo labeled data in a certain target language only; 3) **En + Tgt-Tran**: train the models on the combination of English data and pseudo labeled target-language data; 4) **En + Multi-Tran**: train one single model on the combination of English data and pseudo labeled data in all three target languages. We find filtering the translated sentences can further improve cross-lingual transfer performance, so we use an NER model trained on the sampled English data to label the translated sentences, count the number of entities in each sentence different from NER model predictions, and then remove the top 20% sentences with the most inconsistent entities. This is similar to the third step described in §2.4, except that we remove all the inconsistent sentences from the augmented data, since the LMs can be used to generate a large number of candidate sentences. We set max number of epochs to 10 and use 500 sentences randomly sampled from the English dev data to select the best models for each setting. Then the best models are evaluated on the original target language test sets.

Results Table 2 compares the cross-lingual NER performance of the models trained on the different training sets. Although the performances of **En** and **Tgt-Tran** are relatively bad in most of the cases, combining them can always boost the performance significantly, especially when the dataset size is small. Adding multilingual translated data further improves cross-lingual performance by more than 1% on average when English data size is 1k or less. Therefore, multilingual translation can be used as an effective data augmentation approach in the low resource scenarios of cross-lingual NER. Moreover, the trained single model with En + Multi-Tran can be applied to all target languages.

En Size	Method	de	es	nl	avg
500	En	60.18	55.68	66.09	60.65
	Tgt-Tran	59.97	53.53	60.39	57.96
	En + Tgt-Tran	69.16	64.57	71.40	68.38
	En + Multi-Tran	**70.40**	**65.70**	**72.20**	**69.43**
1k	En	68.95	67.3	73.43	69.89
	Tgt-Tran	70.3	67.22	73.98	70.50
	En + Tgt-Tran	**73.63**	69.81	75.83	73.09
	En + Multi-Tran	73.42	**72.71**	**76.74**	**74.29**
2k	En	69.47	75.2	77.64	74.10
	Tgt-Tran	71.93	72.94	77.95	74.27
	En + Tgt-Tran	74.45	75.88	**78.40**	76.24
	En + Multi-Tran	**75.91**	**76.04**	77.85	**76.60**

Table 2: Cross-lingual NER performance of the models trained on different combinations of training sets.

Method	500	1k	2k
En	78.62	87.00	89.56
Tgt-Tran (avg)	70.07	83.27	87.10
En + Tgt-Tran (avg)	84.62	88.62	90.51
En + Multi-Tran	**85.35**	**88.99**	**90.98**

Table 3: NER Model performance on English test data.

Besides, we also observe that multilingual translated data can even help improve NER performance of the source language. Table 3 summarizes English test data results for the above settings. Tgt-Tran (avg) is the average English results of the models trained on three different Tgt-Tran of German, Spanish and Dutch respectively. En + Tgt-Tran (avg) is the average for combining En with each of the three different Tgt-Tran. As we can see, adding additional translated data consistently improves English NER performance. Particularly, En + Multi-Tran achieves the best performance. Therefore, we can also use multilingual translated data to improve low-resource monolingual NER performance.

3.3 Generation-based Multilingual Data Augmentation

In this section, we run experiments to verify whether applying generation-based data augmentation methods to the multilingual translated data can further improve cross-lingual performance in the low resource scenarios.

Experimental settings We follow the steps described in §2.2 to implement the proposed data augmentation framework on top of LSTM-LM (Kruengkrai, 2019) and mBART (Liu et al., 2020) sep-

	500				1k				2k			
Method	**de**	**es**	**nl**	**avg**	**de**	**es**	**nl**	**avg**	**de**	**es**	**nl**	**avg**
En + Multi-Tran	70.40	65.70	72.20	69.43	73.42	72.71	76.74	74.29	75.91	76.04	77.85	76.60
MulDA-LSTM	70.04	67.38	72.81	70.08	74.80	74.27	77.21	75.42	76.05	76.05	**78.46**	76.85
MulDA-mBART	**72.37**	**68.19**	**74.59**	**71.72**	**75.04**	**74.56**	**77.78**	**75.79**	**77.54**	**76.32**	78.21	**77.36**
En + Tgt-Tran	69.16	64.57	71.40	68.38	73.63	69.81	75.83	73.09	74.45	75.88	78.40	76.24
BiDA-LSTM	**72.51**	**68.77**	**72.65**	**71.31**	**74.97**	**73.69**	**77.51**	**75.39**	**76.59**	**76.47**	**78.97**	**77.34**

Table 4: Cross-lingual NER results of models trained on multilingual augmented data.

Method	**af**	**ar**	**bg**	**bn**	**de**	**el**	**en**	**es**	**et**	**eu**	**fa**	**fi**	**fr**	**he**	**hi**	**hu**	**id**	**it**	**ja**	**jv**
En	70.87	40.45	73.18	67.96	72.86	69.91	74.81	**67.47**	70.38	56.17	48.91	72.92	72.10	41.76	58.96	72.62	47.28	73.42	9.29	59.32
En + Multi-Tran	74.01	42.77	75.54	73.21	74.25	71.38	77.27	66.13	73.23	56.11	**51.28**	74.51	75.21	53.75	**67.52**	73.58	**54.23**	**76.73**	34.51	60.56
Weak Tagger	73.75	38.54	76.12	**74.52**	**75.22**	72.80	78.18	65.77	**73.81**	58.52	43.57	75.00	74.78	53.80	66.75	**75.09**	50.11	76.52	36.13	59.38
MulDA-LSTM	74.25	44.95	76.54	74.19	74.95	71.43	78.23	65.88	73.31	**61.94**	48.40	**75.56**	75.17	55.04	67.49	74.64	50.94	75.73	36.15	**62.03**
MulDA-mBART	**74.58**	**53.62**	**76.99**	74.29	73.80	**73.66**	**78.79**	66.88	72.63	55.66	48.05	74.66	**75.53**	**55.11**	67.46	74.57	53.44	76.37	**37.05**	60.80
Method	**ka**	**kk**	**ko**	**ml**	**mr**	**ms**	**my**	**nl**	**pt**	**ru**	**sw**	**ta**	**te**	**th**	**tl**	**tr**	**ur**	**vi**	**yo**	**zh**
En	53.10	42.70	46.49	55.63	54.66	56.73	44.91	77.04	72.68	55.62	64.59	48.37	43.81	2.56	67.26	73.07	51.08	65.07	44.62	13.46
En + Multi-Tran	64.27	45.10	50.86	60.51	59.84	67.48	50.71	78.17	74.42	60.81	67.81	56.79	48.90	3.67	72.87	73.51	**55.70**	68.54	49.75	39.40
Weak Tagger	64.98	**46.50**	50.13	58.42	59.37	67.79	**53.54**	**79.29**	73.87	63.18	**69.17**	57.07	**51.14**	4.37	73.11	**78.41**	50.34	**71.04**	52.28	38.57
MulDA-LSTM	67.27	46.10	**52.69**	**62.53**	**63.54**	**68.79**	52.62	78.22	74.56	64.28	68.77	58.98	50.88	5.13	**74.97**	76.05	52.37	69.22	48.09	**41.77**
MulDA-mBART	**67.68**	43.12	52.46	58.47	61.49	67.70	52.06	78.86	**76.15**	**65.00**	67.40	**59.30**	48.95	**5.31**	74.57	74.75	48.86	70.25	**52.97**	41.30

Table 5: Cross-lingual NER F1 for Wikiann when only 1k annotated English sentences are available. We assume MT models are only available for the languages highlighted with green background.

arately, and then use them to augment the data processed in §3.2. We concatenate English gold data and the filtered multilingual translated data to train/finetune the modified LMs, where LSTM-LM is trained from scratch and mBART is intialized with the mBART CC25 checkpoint[8] for finetuning. mBART CC25 is a model with 12 encoder and decoder layers trained on 25 languages. We follow the steps described in §2.4 to post-process the augmented data, and concatenate them with the corresponding English gold and translated multilingual data to train the NER models. The size of the augmented data used in each setting is the same as the size of the corresponding English gold data. MulDA-LSTM and MulDA-mBART are used to denote the methods that use LSTM-LM and mBART augmented data respectively. In addition, we also report a bilingual version of our method, denoted with BiDA-LSTM, which performs data augmentation on English and the translated target-language data only. We follow the same settings as above to evaluate cross-lingual performance of the NER models trained on different data.

Results Average results of 5 runs are reported in Table 4. Note that MulDA-LSTM and MulDA-mBART train a single model for all the target languages in each setting, while BiDA-LSTM trains one model for each target language in each setting. Therefore, we compare BiDA-LSTM with En + Tgt-Tran only. As we can see, the proposed multilingual data augmentation methods further improve cross-lingual NER performance consistently. For the 1k and 2k setting, MulDA-LSTM achieves comparable average performance as BiDA-LSTM.

3.4 Evaluation on More Distant Languages

We evaluate the proposed method on a wider range of target languages in this section.

Experimental settings The Wikiann NER data (Pan et al., 2017) processed by Hu et al. (2020) is used in these experiments. 1k English sentences ($\mathcal{D}^S_{1k}$) are sampled from the gold train data to simulate the low resource scenarios. We also assume MT models are not available for all of the target languages, so we only translate the sampled English sentences to 6 target languages: ar, fr, it, ja, tr and zh. $\mathcal{D}^T_{trans}$ is used to denote the translated target-language sentences by following steps described in §2.1. The low quality translated sentences are filtered out in the same way as §3.2. To evaluate our method in the semi-supervised setting, we also sample 5,000 sentences from the training data of the 6 target languages and then remove the NER tags to create unlabeled data $\mathcal{D}^T_{unlabeled}$. We follow the steps described in §2.3 to annotate $\mathcal{D}^T_{unlabeled}$ with one NER model trained on $\{\mathcal{D}^S_{1k}, \mathcal{D}^T_{trans}\}$, and then filter the pseudo labeled data with two other NER models trained on the same data but with different random seeds. We use $\mathcal{D}^T_{semi}$ to denote the data generated with this

[8] https://github.com/pytorch/fairseq/blob/master/examples/mbart/README.md

semi-supervised approach. Finally, we concatenate $\{\mathcal{D}^S_{1k}, \mathcal{D}^T_{trans}, \mathcal{D}^T_{semi}\}$ to generate augmented data $\mathcal{D}^T_{aug}$ following the steps in §2.2 and §2.4. With the augmented data above, we train NER models on the concatenated data of $\{\mathcal{D}^S_{1k}, \mathcal{D}^T_{trans}, \mathcal{D}^T_{aug}\}$ for cross-lingual NER evaluation. We also train an NER model on $\{\mathcal{D}^S_{1k}, \mathcal{D}^T_{trans}, \mathcal{D}^T_{semi}\}$ for comparison, denoted as **Weak Tagger**. The other settings are same as the above experiments.

Method	En	Tran-Train	Zero Shot	All
En	74.81	47.10	57.47	56.35
En + Multi-Tran	77.27	56.91	61.70	61.37
Weak Tagger	78.18	57.19	61.81	61.52
MulDA-LSTM	78.23	58.37	**62.58**	**62.34**
MulDA-mBART	**78.79**	**59.62**	62.24	62.26

Table 6: Summary of the cross-lingual NER performance on Wikiann.

Results We summarize the results in Table 6. Tran-Train is the average performance of the 6 languages that have corresponding training data translated from English. Zero Shot is the average performance of the other target languages. MulDA-LSTM demonstrates promising performance improvements on both the Tran-Train and Zero Shot languages. The performance of MulDA-mBART is slightly lower, one possible reason is the noise introduced by the sentences labeled at character level. We follow the gold data format to label translated zh and ja sequences at character level, which is inconsistent with how mBART is pretrained. Please refer to Table 5 for the detailed cross-lingual NER results of each language.

3.5 Case Study

3.5.1 Effectiveness in Label Projection

The label projection step of the previous methods needs to locate the entities and determine their boundaries, which is vulnerable to many problems, such as word order change, long entities, etc. Our method effectively avoids these problems with placeholders. In the two examples shown in Figure 6, Jain et al. (2019) either labeled only part of the whole entity or incorrectly split the entity into two, Li et al. (2020) incorrectly split the entities into two in both examples, while our method can correctly map the labels.

3.5.2 Multilingual Data Augmentation

We look into the data generated by our multilingual data augmentation method. During LM training, the NER tags can be viewed as a shared vocabulary between different languages. As a result, we find that some generated sentences contain tokens from multiple languages, which are useful to help improve cross-lingual transfer (Tan and Joty, 2021). Two examples are shown in Figure 7.

Example 1
Gold EN: . . . (ORG Association for Relations Across the Taiwan Straits) . . .
Jain et al. (2019): . . . (ORG Vereinigung für Beziehungen) über die Taiwanstraße . . .
Li et al. (2020): . . . (ORG Vereinigung für Beziehungen) über (ORG die Taiwanstraße) . . .
Ours: . . . (ORG Vereinigung für Beziehungen über die Taiwanstraße) . . .

Example 2
Gold EN: . . . (LOC U.S. Midwest) . . .
Jain et al. (2019): . . . (LOC Mittlerer Westen) der (LOC USA) . . .
Li et al. (2020): . . . Mittlerer (LOC Westen) der (LOC USA) . . .
Ours: . . . (LOC Mittlerer Westen der USA) . . .

Figure 6: Two examples that the previous methods fail to find the correct entity boundaries.

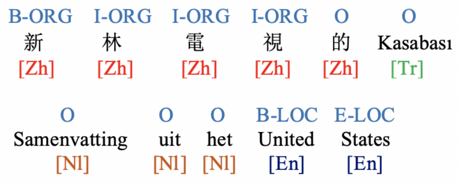
B-ORG I-ORG I-ORG I-ORG O O
新 林 電 視 的 Kasabası
[Zh] [Zh] [Zh] [Zh] [Zh] [Tr]
O O O B-LOC E-LOC
Samenvatting uit het United States
[Nl] [Nl] [Nl] [En] [En]

Figure 7: Examples of multilingual sentences.

4 Related Work

Cross-lingual NER There has been growing interest in cross-lingual NER. Prior approaches can be grouped into two main categories, instance-based transfer and model-based transfer. Instance-based transfer translates source-language training data to target language, and then apply label projection to annotate the translated data (Tiedemann et al., 2014; Jain et al., 2019). Instead of MT, some earlier approaches also use parallel corpora to construct pseudo training data in the target language (Yarowsky et al., 2001; Fu et al., 2014). To minimize resource requirement, Mayhew et al. (2017) and Xie et al. (2018) design frameworks that only rely on word-to-word/phrase-to-phrase translation with bilingual dictionaries. Besides, there are also many studies on improving label projection quality with additional feature or better mapping methods (Tsai et al., 2016; Li et al., 2020). Different from these methods, our labeled sentence translation approach leverages placeholders to determine the position of entities after translation, which effectively avoids many issues during label projection, such as word order change, entity span determination, noise-sensitive similarity metrics and so on.

Model-based transfer directly applies the model trained on the source language to the target-language test data (Täckström et al., 2012; Ni et al., 2017; Joty et al., 2017; Chaudhary et al., 2018), which heavily relies on the quality of cross-lingual representations. Recent methods have achieved significant performance improvement by fine-tuning large scale pretrained multilingual LMs (Devlin et al., 2019; Keung et al., 2019; Conneau et al., 2020). Besides, there are also some approaches that combine instance-based and model-based transfer (Xu et al., 2020; Wu et al., 2020). Compared with these methods, our approach leverages MT models and LMs to add more diversity to the training data, and prevents over-fitting on language-specific features by fine-tuning NER models on multilingual data.

Data augmentation Data augmentation (Simard et al., 1998) adds more diversity to training data to help improve model generalization, which has been widely used in many fields, such as computer vision (Zhang et al., 2018), speech (Cui et al., 2015; Park et al., 2019), NLP (Wang and Eisner, 2016; Sun et al., 2020) and so on. For NLP, back translation (Sennrich et al., 2016) is one of the most successful data augmentation approaches, which translates target-language monolingual data to the source language to generate more parallel data for MT model training. Other popular approaches include synonym replacement (Kobayashi, 2018), random deletion/swap/insertion (Sun et al., 2020; Kumar et al., 2020), generation (Ding et al., 2020), etc. Data augmentation has also been proven to be useful in the cross-lingual settings (Zhang et al., 2019; Singh et al., 2020; Riabi et al., 2020; Qin et al., 2020; Bari et al., 2021; Mohiuddin et al., 2021), but most of the exiting methods overlook the better utilization of multilingual training data when such resources are available.

5 Conclusions

We have proposed a multilingual data augmentation framework for low resource cross-lingual NER. Our labeled sequence translation method effectively avoids many label projection related problems by leveraging placeholders during MT. Our generation-based multilingual data augmentation method generates high quality synthetic training data to add more diversity. The proposed framework has demonstrated encouraging performance improvement in various low-resource settings and across a wide range of target languages.

Acknowledgements

This research is partly supported by the Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University. Linlin Liu would like to thank the support from Interdisciplinary Graduate School, Nanyang Technological University. We would like to thank the help from our Alibaba colleagues, Ruidan He and Qingyu Tan in this work as well.

References

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. *Wireless Personal Communications*, 108(3):1909–1931.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-Resource Cross-Lingual Named Entity Recognition. In *Proceedings of the 34th AAAI Conference on Artifical Intelligence*, AAAI '20, New York, USA. AAAI.

M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. 2021. UXLA: A Robust Unsupervised Data Augmentation Framework for Cross-Lingual NLP. In *Proceedings of The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, Online. Association for Computational Linguistics.

Antonia Baumann. 2019. Multilingual language models for named entity recognition in German and English. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 21–27, Varna, Bulgaria. INCOMA Ltd.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R Mortensen, and Jaime G Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. *arXiv preprint arXiv:1808.09500*.

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, Online. Association for Computational Linguistics.

Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury. 2015. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1469–1477.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online. Association for Computational Linguistics.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2020. Filter: An enhanced fusion method for cross-lingual language understanding.

Ruiji Fu, Bing Qin, and Ting Liu. 2014. Generating chinese named entity data from parallel corpora. *Frontiers of Computer Science*, 8(4):629–641.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

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, Hong Kong, China. Association for Computational Linguistics.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks: Application to community question answering. In *Proceedings of The SIGNLL Conference on Computational Natural Language Learning*, CoNLL'17, pages 226–237, Vancouver, Canada. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1355–1360, Hong Kong, China. Association for Computational Linguistics.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the Third Workshop on Neural Generation and Translation*, Hong Kong. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaoyu Kou, Yaming Yang, Yujing Wang, Ce Zhang, Yiren Chen, Yunhai Tong, Yan Zhang, and Jing Bai. 2020. Improving bert with self-supervised attention.

Canasai Kruengkrai. 2019. Better exploiting latent variables in text modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5527–5532, Florence, Italy. Association for Computational Linguistics.

Canasai Kruengkrai, Thien Hai Nguyen, Sharifah Mahani Aljunied, and Lidong Bing. 2020. Improving low-resource named entity recognition using joint sentence and token labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5898–5905.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020. Unsupervised cross-lingual adaptation for sequence tagging and beyond.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. 2021. Augvic: Exploiting bitext vicinity for low-resource nmt. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3853–3860. International Joint Conferences on Artificial Intelligence Organization. Main track.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2020. Synthetic data augmentation for zero-shot cross-lingual question answering.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Shijing Si, Rui Wang, Jedrek Wosik, Hao Zhang, David Dov, Guoyin Wang, and Lawrence Carin. 2020. Students need more attention: Bert-based attention model for small data with application to automatic patient message triage. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 436–456, Virtual. PMLR.

Patrice Y. Simard, Yann A. LeCun, John S. Denker, and Bernard Victorri. 1998. *Transformation Invariance in Pattern Recognition — Tangent Distance and Tangent Propagation*, pages 239–274. Springer Berlin Heidelberg, Berlin, Heidelberg.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2020. {XLDA}: Cross-lingual data augmentation for natural language inference and question answering.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for NLP tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2012)*.

Samson Tan and Shafiq Joty. 2021. Code-mixing on sesame street: Dawn of the adversarial polyglots. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'21, Mexico City, Mexico. ACL.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140, Ann Arbor, Michigan. Association for Computational Linguistics.

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 Seventh 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.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. 2020. Unitrans: Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2018. Conditional bert contextual augmentation.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2019. Cross-lingual dependency parsing using code-mixed TreeBank. 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 997–1006, Hong Kong, China. Association for Computational Linguistics.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. 2020. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*.

A Appendix

A.1 Translation with Placeholders

Figure 8 shows more examples of translating the sequence *"PER0 was born in LOC1."* to different languages. We can see that the placeholders are all well kept. Meanwhile, the translation quality is also good.

Source sentence:
en: PER0 was born in LOC1.

Translations:
de: PER0 wurde in LOC1 geboren.
es: PER0 nació en LOC1.
nl: PER0 is geboren in LOC1.
vi: PER0 được sinh ra ở LOC1.
fr: PER0 est né en LOC1.
zh: PER0出生于LOC1。

Figure 8: Translations of *"PER0 was born in LOC1."* to different languages with Google translation system.

A.2 Number of Entities in Translated Data

We count the total number of entities in gold EN data and the translated data. As shown in Table 7, the number of entities in our translated data is the most close to that of the gold EN data.

Method	**de**	**es**	**nl**
Jain et al. (2019)†	23068	23442	23275
Li et al. (2020)†	23844	23335	23930
ours	**23418**	**23473**	**23475**
Gold En		23499	

Table 7: Number of entities in translated data. The bold text denotes the numbers most to that of the gold EN data. † denotes the reproduced results.

A.3 Visualization of Entity Representations

We visualize the last layer transformer outputs of the finetuned NER model with t-SNE. We finetune two XLM-R initialized NER models on English and MulDA-LSTM respectively, and generate last layer representations with Chinese test data. Only the token representations corresponding to the B and I tags are saved. The two dimensional t-SNE visualizations are shown in Figures 9 and 10. As we can see, the representation clusters corresponding to different NER entities in Figure 10 (MulDA-LSTM) are further separated than that in Figure 9 (English).

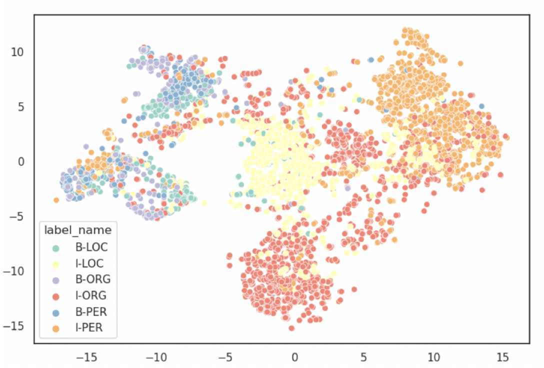

Figure 9: Entity representation distribution of the NER model trained on English.

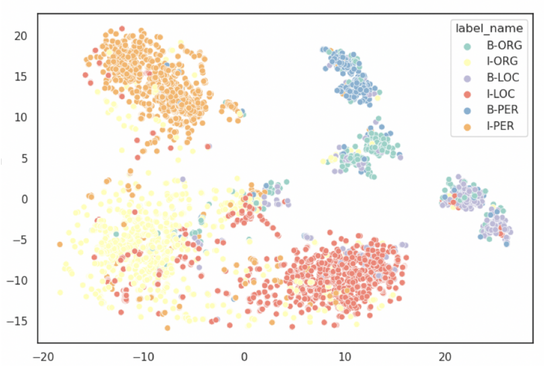

Figure 10: Entity representation distribution of the NER model trained on MulDA-LSTM augmented data.

A.4 Parameters

The parameters used for NER model fine-tuning are shown in Table 8.

Parameters	Values
Batch Size	16
Optimizer	AdamW
Learning Rate	2e-5
Betas	(0.9, 0.999)
Max Number of Epochs	10

Table 8: Parameters used for NER model fine-tuning.