DualNER: A Dual-Teaching framework for Zero-shot Cross-lingual Named Entity Recognition

Jiali Zeng[1]*, **Yufan Jiang**[1], **Yongjing Yin**[2], **Xu Wang**[1], **Binghuai Lin**[1], **Yunbo Cao**[1]
[1]Tencent Cloud Xiaowei, Beijing, China
[2]Zhejiang University, Westlake University, Zhejiang, China
{lemonzeng,garyyfjiang,frostwu,ethanlli}@tencent.com
yinyongjing@westlake.edu.cn

Abstract

We present DualNER, a simple and effective framework to make full use of both annotated source language corpus and unlabeled target language text for zero-shot cross-lingual named entity recognition (NER). In particular, we combine two complementary learning paradigms of NER, i.e., sequence labeling and span prediction, into a unified multi-task framework. After obtaining a sufficient NER model trained on the source data, we further train it on the target data in a *dual-teaching* manner, in which the pseudo-labels for one task are constructed from the prediction of the other task. Moreover, based on the span prediction, an entity-aware regularization is proposed to enhance the intrinsic cross-lingual alignment between the same entities in different languages. Experiments and analysis demonstrate the effectiveness of our DualNER. Code is available at https://github.com/lemon0830/dualNER.

1 Introduction

Aiming at classifying entities in un-structured text into pre-defined categories, named entity recognition (NER) is an indispensable component for various downstream neural language processing applications such as information retrieval (Banerjee et al., 2019) and question answering (Fabbri et al., 2020). Current supervised methods have achieved great success with sufficient manually labeled data, but the fact remains that most of the annotated data are constructed for high-resource languages like English and Chinese, posing a big challenge to low-resource scenarios (Mayhew et al., 2017; Bari et al., 2021).

To address this issue, zero-shot cross-lingual NER is proposed to transfer knowledge of NER from high-resource languages to low-resource languages. The knowledge can be acquired in either of the following two ways: 1) from aligned cross-lingual word representations or multilingual pre-trained encoder fine-tuned on high-resource languages (Conneau et al., 2020; Bari et al., 2021). 2) from translated target language data with label projection (Mayhew et al., 2017; Jain et al., 2019; Liu et al., 2021). These two kinds of methods can be unified into a knowledge distillation (KD) framework, to further improve the cross-lingual NER performance (Wu et al., 2020; Fu et al., 2022). Though widely used, the transfer process still suffers from poor translation quality, label projection error and over-fitting of large-scale multilingual language models.

In this paper, we present a simple and effective framework, named DualNER, alleviating the above problems from a different angle. We combine two popular complementary learning paradigms of NER, sequence labeling and span prediction, into a single framework. Specifically, we first train a teacher NER model by jointly exploiting sequence labeling and span prediction with the annotated source language corpus. Unlike the previous KD-based methods that produce pseudo labels for the corresponding paradigms, we propose a dual-teaching strategy to make the two paradigms complement each other. More concretely, the model prediction for sequence labeling is used to construct the pseudo-labels for span prediction and vice versa. Furthermore, we propose a multilingual entity-aware regularization forcing same entities in different languages to have similar representations. By doing this, the trained model is able to leverage the intrinsic cross-lingual alignment across different languages to enhance the cross-lingual transfer ability.

Experiments and analysis conducted on XTREME for 40 target languages well validate the effectiveness of DualNER [1].

[*] Corresponding author.

[1] We will release code upon acceptance

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1837–1843
December 7-11, 2022 ©2022 Association for Computational Linguistics

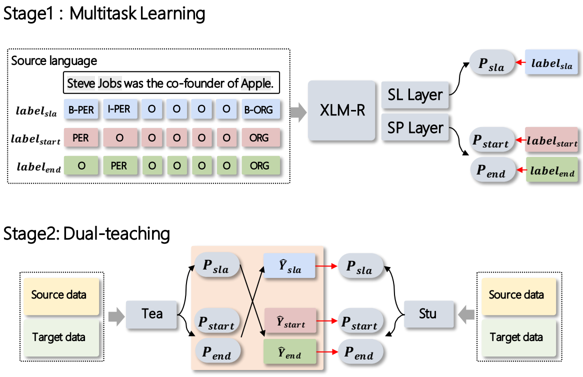

Figure 1: The overall framework of DualNER. SL denotes *Sequence Labeling* and SP denotes *Span Prediction*.

2 Framework

Recently, the dominate paradigm for NER shifts from *sequence labeling* (Ma and Hovy, 2016; Lample et al., 2016; Devlin et al., 2019; Xia et al., 2019; Luo et al., 2020; Lin et al., 2020) to *span-level prediction* (Jiang et al., 2020; Ouchi et al., 2020; Li et al., 2020; Xue et al., 2020; Fu et al., 2021). We combine these two formulations into a unified multitask framework for complementarity. As shown in Figure 1, our DualNER consists of three major modules: *Token Representation Layer*, *Sequence Labeling Layer*, and *Span Prediction Layer*.

2.1 Model

Given an example of training data (X, Y_{sla}), where X=$\{x_1, x_i, ..., x_n\}$ is the input sequence and Y_{sla}=$\{y_1, y_i, ..., y_n\}$ is the corresponding label (e.g., "B-ORG", "I-PER", "O") sequence, we can extract the start and end index sequence, Y_{start} and Y_{end}, as reference for span prediction, and convert the training instance to a quadruple $(X, Y_{sla}, Y_{start}, Y_{end})$.

Token Representation Layer. For the input sequence X, we use a multilingual pre-trained language models (PLM), e.g., XLM-R, to obtain the contextualized representations H=$\{h_1, ..., h_i, ..., h_n\}$.

Sequence Labeling Layer. Formally, we stack a softmax classifier layer on H, and the objective of sequence labeling is

$$\mathcal{J}_{sla} = -log(P_{sla}(Y_{sla}|H; \theta, \theta_{sla})), \quad (1)$$

where θ and θ_{sla} denote the parameters of PLM and the classifier respectively.

Span Prediction Layer. For the formulation of span prediction, we adopt two $(C+1)$-class classifiers, where C denotes the number of NER entities (e.g., LOC, PER, ORG, 3 entities in XTREME-40 dataset), and one is used to predict whether each token is the start of an entity, and the other is used to predict whether each token is the end. Formally, given the representations H and two label sequences Y_{start} and Y_{end} of length n, the losses for start and end index predictions are defined as:

$$\mathcal{J}_{start} = -log(P_{start}(Y_{start}|H; \theta, \theta_{start})) \quad (2)$$

$$\mathcal{J}_{end} = -log(P_{end}(Y_{end}|H; \theta, \theta_{end})). \quad (3)$$

2.2 Training

To achieve zero-shot cross-language NER, we adopt a two-stage training strategy.

Stage 1: Multitask Learning. At the first stage, we fine-tune a multilingual pre-trained model on the labeled source language data in a multi-task manner:

$$\mathcal{J}^{src} = \mathcal{J}^{src}_{sla} + \mathcal{J}^{src}_{start} + \mathcal{J}^{src}_{end}. \quad (4)$$

Stage 2: Dual-teaching. At the stage two, we focus on generating pseudo labels for both labeled and unlabeled data with the trained NER model θ^{tea}. In particular, the pseudo labels for the sequence labeling task are converted by the model prediction for the span prediction task, and vice versa. Specifically, based on the predictions P_{sla}, P_{start} and P_{end} of an input sequence X^{src} (or X^{trg}), we construct the pseudo labels for sequence labeling and span prediction as follows:

$$\hat{Y}_{sla} = \text{Sequential}(P_{start}, P_{end}) \quad (5)$$

$$\hat{Y}_{start}, \hat{Y}_{end} = \text{ExtractSpan}(P_{sla}), \quad (6)$$

where *Sequential* and *ExtractSpan* are the corresponding transformation between sequence labels and span labels.

As a result, X^{src} is paired with six label sequences $\{Y^{src}_{sla}, Y^{src}_{start}, Y^{src}_{end}, \hat{Y}^{src}_{sla}, \hat{Y}^{src}_{start}, \hat{Y}^{src}_{end}\}$, and X^{trg} is paired with three pseudo label sequences $\hat{Y}^{trg}_{sla}$, $\hat{Y}^{trg}_{start}$, and $\hat{Y}^{trg}_{end}\}$. Using the constructed data, we train a student model θ^{stu} initialized with θ^{tea} with the following objective:

$$\mathcal{J}^{src} = 0.5 * \mathcal{J}^{src}(X^{src}, Y^{src}_{sla}, Y^{src}_{start}, Y^{src}_{end}) \quad (7)$$
$$+ 0.5 * \mathcal{J}^{src}(X^{src}, \hat{Y}^{src}_{sla}, \hat{Y}^{src}_{start}, \hat{Y}^{src}_{end})$$
$$\mathcal{J}^{trg} = \mathcal{J}^{trg}(X^{trg}, \hat{Y}^{trg}_{sla}, \hat{Y}^{trg}_{start}, \hat{Y}^{trg}_{end}). \quad (8)$$

XLM-R$_{base}$

Method	en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hl	hu	id	it	ja	jv
CLA	81.7	75.3	49.4	78.2	70.2	74.4	74.7	69.5	71.5	58.4	50.6	75.6	75.3	52.2	68.1	77.0	47.7	77.2	21.5	54.8
SPAN	83.1	75.8	51.0	77.8	70.7	74.4	75.4	**80.6**	67.8	61.8	44.5	75.3	79.7	50.3	68.8	76.6	47.4	77.9	25.5	57.4
MLT	82.9	75.1	**79.3**	77.6	71.3	73.6	75.2	74.3	67.8	59.8	46.6	75.3	76.5	49.7	68.3	75.1	48.8	77.0	20.1	54.2
FILTER (Fang et al., 2021)	83.3	78.7	56.2	83.3	75.4	79.0	79.7	75.6	80.0	67.0	**70.3**	80.1	79.6	55.0	72.3	80.2	**52.7**	**81.6**	25.2	61.8
DualNER																				
+TRG$_{Trans}$	**84.1**	78.0	59.0	80.0	**75.5**	78.1	74.7	74.5	77.1	62.4	49.8	78.6	78.9	55.7	73.3	79.5	51.0	80.0	33.8	56.4
+TRG$_{Gold}$	83.6	**80.0**	62.4	**84.6**	73.9	**81.9**	**80.8**	79.1	**80.4**	**68.0**	66.4	**82.9**	**81.8**	**69.6**	**74.3**	**84.5**	52.5	80.3	**38.4**	**63.5**

Method	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh
CLA	66.1	42.3	49.3	59.3	61.3	67.2	55.2	79.8	77.0	64.3	67.9	54.1	34.0	0.04	69.7	76.7	56.0	68.5	37.5	26.3
SPAN	65.8	47.8	47.7	57.4	59.4	54.3	43.0	80.9	80.2	61.4	70.7	53.8	47.1	0.03	72.3	79.3	66.3	66.8	**55.8**	31.9
MLT	64.5	42.8	46.7	56.3	58.8	62.3	44.2	80.5	79.3	60.8	67.6	54.6	46.2	0.01	73.0	78.1	52.9	64.9	50.2	28.4
FILTER (Fang et al., 2021)	70.0	50.6	**63.8**	67.3	66.4	68.1	60.7	83.7	81.8	71.5	68.0	62.8	56.2	**1.5**	74.5	80.9	71.2	76.2	40.4	35.9
DualNER																				
+TRG$_{Trans}$	71.7	50.0	55.7	69.4	63.6	**70.8**	**67.7**	83.1	80.3	63.3	70.5	58.8	50.7	0.08	78.0	81.6	63.7	73.4	37.9	43.6
+TRG$_{Gold}$	**74.1**	**52.2**	62.7	**71.1**	**72.2**	70.4	66.2	**84.6**	**85.2**	**64.8**	**70.7**	**66.5**	**61.6**	0.1	**80.2**	**86.0**	**84.3**	**77.2**	35.6	**46.1**

InfoXLM$_{large}$

Method	en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hl	hu	id	it	ja	jv
CLA	84.5	81.4	64.2	83.2	78.3	81.2	81.5	77.6	79.0	67.0	68.9	80.9	80.1	58.0	74.1	82.0	61.0	82.4	28.0	61.2
SPAN	85.6	81.7	64.2	84.6	77.3	81.9	81.9	74.3	76.6	66.4	61.2	82.2	80.3	54.9	75.1	81.4	56.8	82.4	31.9	59.0
MLT	85.4	83.5	62.5	84.1	78.0	80.9	82.9	80.5	78.1	**68.3**	62.6	82.4	81.6	61.2	74.8	82.5	55.8	82.4	29.9	63.9
DualNER																				
+TRG$_{Trans}$	85.7	**83.2**	65.6	84.1	80.3	**82.8**	80.9	74.5	**82.6**	61.7	67.3	**82.7**	82.1	66.3	77.7	82.3	**68.1**	**83.9**	52.8	**68.3**
+TRG$_{Gold}$	**85.9**	82.5	**73.5**	**86.5**	**82.9**	81.8	**83.0**	**83.2**	79.4	67.8	**71.9**	82.3	**86.3**	**74.9**	**79.5**	**85.3**	55.1	83.7	**55.1**	68.2

Method	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh
CLA	71.7	59.9	60.6	65.7	68.5	71.1	51.8	85.4	83.3	72.0	71.6	61.5	57.1	0.01	75.0	83.8	78.2	76.6	36.4	35.0
SPAN	71.2	**60.2**	57.8	64.2	70.1	73.3	43.9	86.2	83.8	73.6	72.5	60.0	51.5	0.02	78.0	85.7	75.1	80.3	**56.2**	38.9
MLT	74.7	56.1	61.6	68.5	67.9	72.1	46.2	86.0	83.1	74.2	72.8	61.3	54.1	0.02	76.3	84.6	75.6	78.1	43.8	36.7
DualNER																				
+TRG$_{Trans}$	78.6	56.1	69.3	72.6	**71.5**	72.1	**68.6**	86.3	85.0	72.4	**73.3**	66.2	56.5	0.05	**80.6**	83.2	80.2	**83.1**	42.3	55.5
+TRG$_{Gold}$	**79.0**	51.5	**74.2**	**75.3**	68.6	**74.0**	66.0	**86.7**	**86.5**	**80.4**	71.4	**73.6**	**62.4**	**0.08**	76.9	**85.9**	**87.5**	82.3	42.9	**57.3**

Table 1: **Experimental results on the test sets of XTREME-40 NER**. We highlight better results between *CLA* and *SPAN* with gray and highlight best results among all methods with pink.

Furthermore, in order to strengthen the correlation of the same entities across languages, we present an entity-aware regularization term. We illustrate an example in Appendix A. More concretely, for the j-th entity, we extract the start token and the end token by applying *argmax* to the distributions P_{start} and P_{end}, and obtain its representation r_j by concatenating the representations of the two tokens. We use a mean square error (MSE) loss to pull the representations of the same entities across different languages together:

$$\mathcal{J}_{mse} = -\frac{1}{C}\sum_{c=1}^{C}\frac{1}{|R_c|}\sum_{(r_m,r_q)\in R_c, m\neq q}(r_m - r_q)^2, \tag{9}$$

where C is the number of NER entities and R_c is the representation set of the c-th entity in a mini-batch.

The overall training objective is defined as:

$$\mathcal{J} = \mathcal{J}^{src} + \mathcal{J}^{trg} + \alpha \cdot \mathcal{J}_{mse}, \tag{10}$$

where α is a hyper-parameter to balance the effect of MSE loss. During training, we update the teacher NER model θ^{tea} using the better student model θ^{stu} based on the validation performance. At inference time, we only use the prediction of *Span Prediction Layer*.

3 Experiments & Analysis

3.1 Setup

The proposed method is evaluated on the cross-lingual NER dataset from the XTREME-40 benchmark (Hu et al., 2020). Named entities in Wikipedia are annotated with LOC, PER, and ORG tags in BOI-2 format. We try two types of unlabeled target language data: **Natural Language Text**, the target language text in the training set of XTREME-40; and **Translation Text** (Fang et al., 2021). We take XLM-R-base (Conneau et al., 2020) and InfoXLM-large (Chi et al., 2021) as our backbones, and set α as 0.5. Detailed experimental setups are shown in Appendix B. We use entity-level F1-score of all language development sets to choose the best checkpoint, and report the F1-score on each test set of each language.

3.2 Main Result

We compare DualNER to the following baselines: 1) *FILTER* (Fang et al., 2021), which feeds paired language input into PLM and is trained with self-teaching; 2) *CLA*, which formulates NER as a sequence labeling problem; 3) *SPAN*, which formulates NER as a span prediction problem; and 4) *MLT*, the model trained after our Stage 1. Besides, we name DualNER trained on unlabeled target natural language text as *DualNER+TRG*$_{Gold}$, while denote DualNER trained on target language trans-

Model	F1
MLT	61.53±0.53
DualNER+TRG$_{Gold}$	**68.64**±0.06
w/o $\mathcal{J}_{mse}$	68.05±0.49
w/ selfKL	65.18±0.65
w/o TRG	62.17±0.52

Table 2: **Ablation Study**. We run 3 times with different random seeds and report mean and standard deviation on all the validation sets.

lation text as *DualNER+TRG$_{Trans}$*.

Table 1 reports the zero-shot cross-lingual NER results. The conclusions are as follows: 1) *CLA* and *SPAN* have no obvious advantages over each other. 2) *DualNER* significantly outperforms the baselines on almost all of the languages, demonstrating the effectiveness of our proposed method. 3) Directly combining *CLA* and *SPAN* into a multitask learning framework (i.e., *MLT*) fails to achieve consistent improvement. This observation shows that the gain of DualNER entirely comes from the proposed dual-teaching training strategy, rather than the usage of multitask learning. 4) As expected, using natural language text (i.e., *DualNER+TRG$_{Gold}$*) achieves better performance compared to translation text (i.e., *DualNER+TRG$_{Trans}$*), since translations possibly lose the idiomatic expressions of some entities.

3.3 Ablation Study

To analyze the impact of different components of DualNER, we investigate the following three variants: 1) *DualNER w/o $\mathcal{J}_{mse}$*, removing the entity-aware regularization; 2) *DualNER w/ selfKL*, where *Dual-teaching* is replaced by *Self-teaching* with KL loss at the Stage 2. 3) *DualNER w/o TRG*, where we only use the source language data in the Stage 2. We take XLM-R$_{base}$ as the backbone. The results are listed in Table 2. Compared with *DualNER w/ selfKL*, *DualNER* obtains a significant improvement of 3.46 points, validating our motivation in making use of complementarity of different task paradigms of NER. The degradation of *DualNER w/o $\mathcal{J}_{mse}$* and *DualNER w/o TRG* confirm the intrinsic cross-lingual alignment and the importance of task-related target language information.

3.4 Visualization

We choose English, Korean, and Arabic, which comes from different language families, and visualize the entity representations r in Eq. 9 with hypertools (Heusser et al., 2017). As shown in Figure 2, the representations of different entities in the same language are clearly distributed in different regions, while the representations of the same entity across different languages are concentrated.

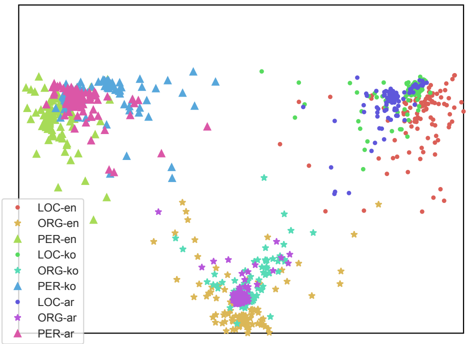

Figure 2: The visualization of the entity representations in different languages, where the triangle-shaped, circle-shaped, and pentagonal-shaped(blue) points denote location, organization, and person entities, respectively.

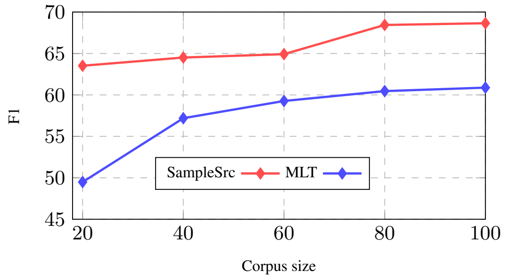

Figure 3: **Effect of Source Language Corpus Size.** We report F1-score on all the validation sets.

3.5 Effect of Source Language Corpus Size

In this experiment, we study the impact of annotated source language corpus size on DualNER by sampling different percentages of annotated source language corpus for the Stage 1. Meanwhile, we remove the labels of the remaining source data, and mix it with the unlabeled target language text for the Stage 2. Figure 3 shows the comparison between *DualNER* and *MLT*. Surprisingly, DualNER trained with only 20% of annotated source data achieves better performance than MLT trained using complete data, demonstrating the data-efficiency of our proposed method.

4 Conclusion

In this paper, we propose a simple and effective dual-teaching framework, coined DualNER, for zero-shot cross-lingual named entity recognition. In particular, DualNER makes full use of the exchangeability of the labels in span prediction and sequence labeling, and generates abundant pseudo data for available labeled and unlabeled data. Experiments and analysis validate the effectiveness of our DualNER.

5 Limitations

The performance of DualNER relies on the capability of cross-lingual transfer of multilingual pretrained models. In practice, for an adequate quality of the pseudo-labels generated in the stage 2, it is necessary to ensure that the NER model has acquired certain ability to conduct cross-lingual transfer in the stage 1.

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. *Wirel. Pers. Commun.*, 108(3):1909–1931.

M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. 2021. UXLA: A robust unsupervised data augmentation framework for zero-resource cross-lingual NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1978–1992, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

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.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. 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. 2021. FILTER: an enhanced fusion method for cross-lingual language understanding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12776–12784. AAAI Press.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195, Online. Association for Computational Linguistics.

Yingwen Fu, Nankai Lin, Ziyu Yang, and Shengyi Jiang. 2022. A dual-contrastive framework for low-resource cross-lingual named entity recognition. *CoRR*, abs/2204.00796.

Andrew C. Heusser, Kirsten Ziman, Lucy L. W. Owen, and Jeremy R. Manning. 2017. Hypertools: a python toolbox for gaining geometric insights into high-dimensional data. *J. Mach. Learn. Res.*, 18:152:1–152:6.

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 generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 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.

Zhengbao Jiang, Wei Xu, Jun Araki, and Graham Neubig. 2020. Generalizing natural language analysis through span-relation representations. In *Proceedings of the 58th Annual Meeting of the Association* *for Computational Linguistics*, pages 2120–2133, Online. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020. Triggerner: Learning with entity triggers as explanations for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8503–8511. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Ying Luo, Fengshun Xiao, and Hai Zhao. 2020. Hierarchical contextualized representation for named entity recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8441–8448. AAAI Press.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

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.

Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui. 2020. Instance-based learning of span representations: A case study through named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6452–6459, Online. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jianguang Lou. 2020. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3926–3932. ijcai.org.

Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip S. Yu. 2019. Multi-grained named entity recognition. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1430–1440. Association for Computational Linguistics.

Mengge Xue, Bowen Yu, Zhenyu Zhang, Tingwen Liu, Yue Zhang, and Bin Wang. 2020. Coarse-to-fine pre-training for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6345–6354. Association for Computational Linguistics.

A Example of Entity-aware Regularization

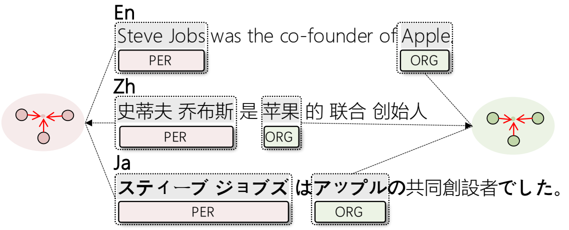
En
Steve Jobs was the co-founder of Apple.
PER
ORG
Zh
史蒂夫 乔布斯 是 苹果 的 联合 创始人
PER
ORG
Ja
スティーブ ジョブズ はアップルの共同創設者でした。
PER
ORG

Figure 4: Entity-aware Regularization.

Figure 4 illustrates an example of entity-aware regularization.

B Settings for Different Pretrained Models

In this paper, we fine-tuned different pretrained models including XLM-R-base and InfoXLM-large. We evaluate the model each 250 steps. The batch size, training epoch, warmup steps and learning rate in two-stage training are list in Table 3.

Model	**Batch Size**	**Epoch**	**Warmup**	**lr**
		Stage 1		
XLM-R$_{base}$	128	8	300	2e-5
InfoXLM$_{large}$	128	8	300	2e-5
		Stage 2		
XLM-R$_{base}$	500	8	300	2e-5
InfoXLM$_{large}$	128	8	300	2e-5

Table 3: Hyper-parameters settings for different pre-trained models.