Wider & Closer: Mixture of Short-channel Distillers for Zero-shot Cross-lingual Named Entity Recognition

Jun-Yu Ma[1*] Beiduo Chen[1*] Jia-Chen Gu[1] Zhen-Hua Ling[1] Wu Guo[1†] Quan Liu[2,3] Zhigang Chen[4] Cong Liu[1,3]

[1]National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China
[2]State Key Laboratory of Cognitive Intelligence [3]iFLYTEK Research [4]Jilin Kexun Information Technology Co., Ltd.

Introduction

- **Task** zero-shot cross-lingual named entity recognition(NER).
- **Challenge** few annotated data are available for some languages.
- **Solution** the mixture of short-channel distillers (MSD). This method first attempts to interact the rich hierarchical information in the teacher model and to transfer knowledge to the student model. Then it adopts parallel domain adaptation to shorten the channels between the teacher and student models to preserve domain-invariant features.
- **Results** experiments on four datasets across nine languages demonstrate that MSD achieves new *state-of-the-art* performance and shows great generalization and compatibility across languages and fields.

Mixture of Distillers

To establish multiple information transmission channels, each layer of the pre-trained mBERT is appended with a classifier. Given a sentence $\boldsymbol{x}$ of length L with labels $\boldsymbol{y}$ from source language data $\mathcal{D}^{S}_{\text{train}}$, these could be described as:

$$\boldsymbol{H}^m = \boldsymbol{f_\theta}^m(\boldsymbol{x}), \tag{1}$$

$$\boldsymbol{p}^m\left(x_i;\Theta\right) = \text{softmax}\left(\boldsymbol{W}^m \cdot \boldsymbol{h}_i^m + \boldsymbol{b}^m\right), \tag{2}$$

where $\boldsymbol{H^m}$ is the sentence representation from the m-th layer of mBERT and $\boldsymbol{p}^m\left(x_i;\Theta\right)$ is the probability distribution generated from the corresponding channel terminal.

For the teacher, the language model along with several channel terminals are jointly trained on the labeled source language data. Noting that the embedding layer and the bottom three layers of mBERT in the teacher and student models are frozen. So it is optimized as:

$$\mathcal{L}_{\text{main}} = \frac{1}{L}\sum_{i=1}^{L}\mathcal{L}_{\text{CE}}\ \left(\boldsymbol{p}^{12}\left(x_i;\Theta\right), y_i\right), \tag{3}$$

$$\mathcal{L}_{\text{aux}} = \frac{1}{L}\sum_{i=1}^{L}\sum_{m=4}^{11}\lambda_m\mathcal{L}_{\text{CE}}\ \left(\boldsymbol{p}^{m}\left(x_i;\Theta\right), y_i\right), \tag{4}$$

$$\mathcal{L}_{\text{tea}} = \mathcal{L}_{\text{main}} + \alpha\mathcal{L}_{\text{aux}}, \tag{5}$$

where $\lambda_m \in \mathbb{R}$ is a trainable parameter representing the contribution degree of the m-th layer and $\mathcal{L}_{\text{CE}}$ is cross entropy loss. α regulates the contribution of the auxiliary layers.

Next, for the following knowledge distillation, a student model Θ_{stu} is distilled based on the unlabeled target language data $\mathcal{D}^{T}_{\text{train}}$. Given a sentence $\boldsymbol{x'}$ of length L from $\mathcal{D}^{T}_{\text{train}}$, the distillation loss of the student model is as follows:

$$\mathcal{L}^{KD}_{\text{m}} = \frac{1}{L}\sum_{i=1}^{L}\text{MSE}\ \left(\boldsymbol{p}^{m}\left(x'_i;\Theta_{tea}\right), \boldsymbol{p}^{m}\left(x'_i;\Theta_{stu}\right)\right), \tag{6}$$

$$\mathcal{L}_{\text{stu}} = \mathcal{L}^{KD}_{\text{main}} + \beta\mathcal{L}^{KD}_{\text{aux}} = \mathcal{L}^{KD}_{12} + \sum_{m=4}^{11}\lambda'_m\mathcal{L}^{KD}_{m}, \tag{7}$$

where λ'_m and β have the same effect on the student as λ_m and α do on the teacher. MSE is the mean squared error loss.

Parallel Domain Adaptation

The parallel domain adaptation method based on MMD is proposed to preserve domain information between the teacher and student models at sentence-level during distillation.

Cross-model and cross-language MMD losses are proposed to minimize the cross-model and cross-language discrepancies respectively, which are denoted as $\mathcal{L}^{M}_{MMD}$ and $\mathcal{L}^{L}_{MMD}$. During distillation, the soft labels $D^{S_{tea}}_{train}$ and $D^{S_{stu}}_{train}$ are obtained by applying the teacher and student models to the source language data respectively. Meantime, the soft labels $D^{T_{stu}}_{train}$ is obtained by applying the student model to the unlabeled target language data. The $\mathcal{L}^{M}_{MMD}$ and $\mathcal{L}^{L}_{MMD}$ could be formulated as:

$$\mathcal{L}^{M}_{\text{MMD}}(D^{S_{tea}}_{train}, D^{S_{stu}}_{train}) = \text{MMD}^2(\boldsymbol{H}^{\boldsymbol{S_{tea}}}_{\boldsymbol{cls}}, \boldsymbol{H}^{\boldsymbol{S_{stu}}}_{\boldsymbol{cls}}), \tag{8}$$

$$\mathcal{L}^{L}_{\text{MMD}}(D^{S_{tea}}_{train}, D^{T_{stu}}_{train}) = \text{MMD}^2(\boldsymbol{H}^{\boldsymbol{S_{tea}}}_{\boldsymbol{cls}}, \boldsymbol{H}^{\boldsymbol{T_{stu}}}_{\boldsymbol{cls}}), \tag{9}$$

where $\boldsymbol{H_{cls}}$ denotes a set of `[CLS]` token embeddings $\boldsymbol{h_{cls}}$.

The training for the final student model contains two parts: the mixture of distillers and the parallel domain adaptation. The final loss is denoted as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{stu}} + \alpha'\mathcal{L}^{M}_{\text{MMD}} + \beta'\mathcal{L}^{L}_{\text{MMD}}, \tag{10}$$

where α' and β' are the weights to balance the contributions of the parallel adaptation methods.

Datasets & Basic System

- **Datasets** four Datasets: (1) CoNLL-2002 includes Spanish and Dutch; (2) CoNLL-2003 includes English and German; (3) WikiAnn includes English and three non-western languages (Arabic,Hindi, and Chinese); (4) mLOWNER includes four languages (English,Korean, Farsi, and Turkish). Different datasets have different entity types: CoNLL-2002 and CoNLL-2003 are annotated with 4 entity types: LOC, MISC, ORG, and PER. WikiAnn is annotated with 3 entity types: LOC, ORG, and PER. mLOWNER is annotated with 6 entity types: LOC, ORG, PER , CW, GRP, PROD. All datasets were annotated with the BIO entity labelling scheme.
  Following previous work, English is employed as the source language in all experiments, and the other languages are employed as target languages.
- **Basic System** the mBERT is used as the pre-trained language model with a Softmax classifier predicts the tag of each token. Then a source-language model is used as teacher to train a student model on unlabeled data in the target language for cross-lingual NER.

Parallel Domain Adaptation

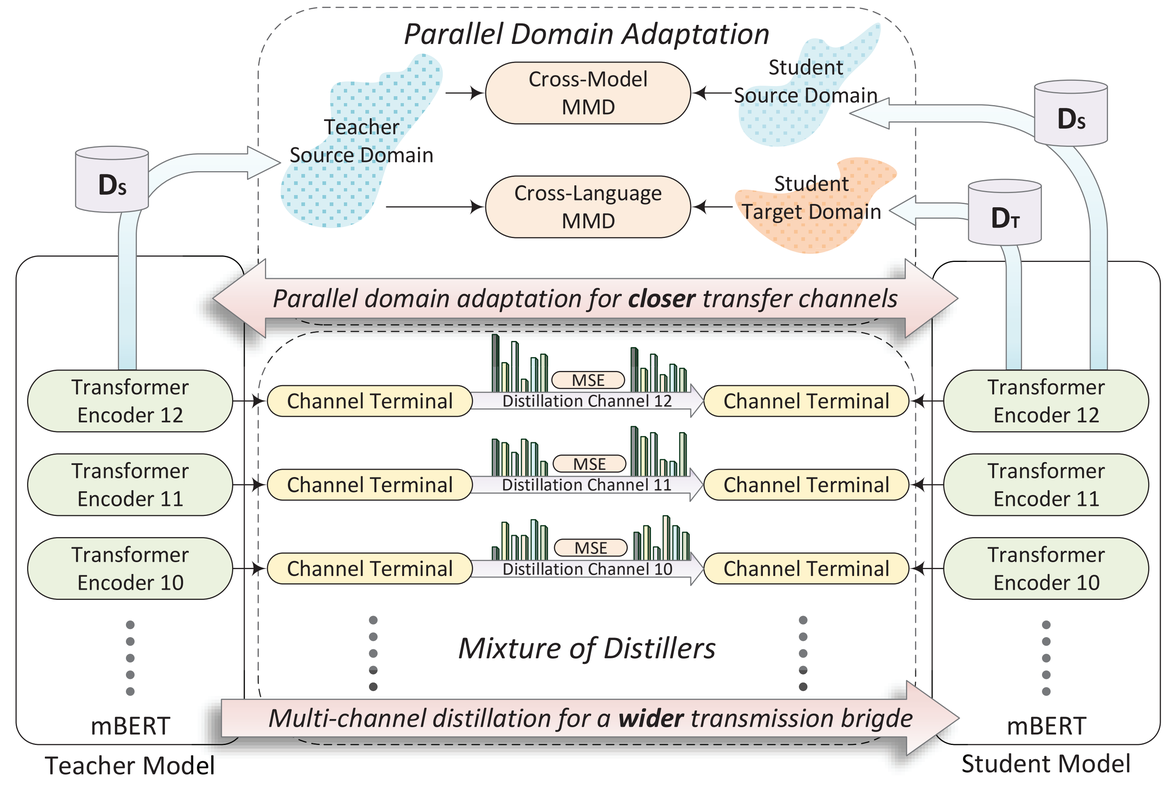

Figure 1. The overall structure of the proposed MSD.

Parallel Domain Adaptation

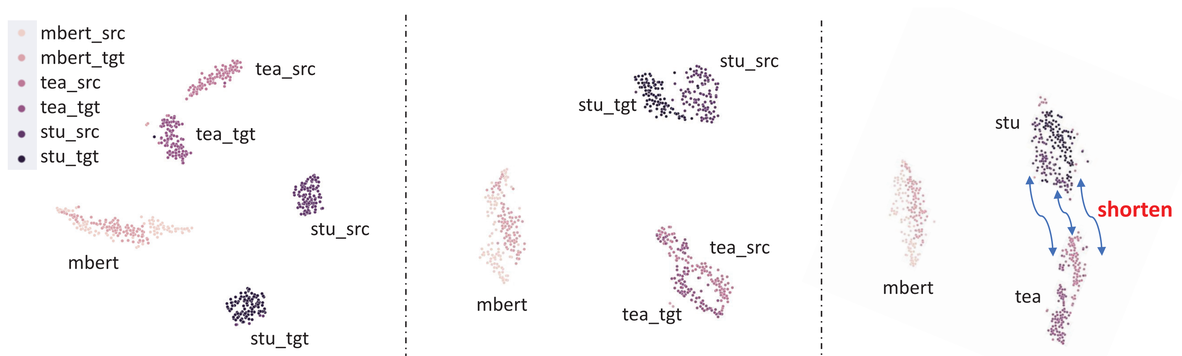

Figure 2. T-SNE visualization of semantic domains of different models by randomly sampling 100 unannotated English (source) and German (target) sentences from the training set of the CoNLL datasets.

Experiments

Method	de	es	nl	Avg
Wiki	48.12	60.55	61.56	56.74
WS	58.50	65.10	65.40	63.00
BWET	57.76	72.37	71.25	67.13
ADV	71.90	74.30	77.60	74.60
BS	69.59	74.96	77.57	73.57
TSL	73.16	76.75	80.44	76.78
Unitrans	74.82	79.31	82.90	79.01
AdvPicker	75.01	79.00	82.90	78.97
RIKD	75.48	77.84	82.46	78.59
TOF	76.57	80.35	82.79	79.90
MTMT	76.80	81.82	83.41	80.67
MSD	**77.56**	**81.92**	**85.11**	**81.53**
MSD w/o. distillers	75.31	79.34	83.16	79.27
MSD w/o. $\mathcal{L}^{L}_{\text{MMD}}$	76.68	80.27	84.07	80.34
MSD w/o. $\mathcal{L}^{M}_{\text{MMD}}$	77.12	79.81	84.36	80.43
MSD w/o. all	74.17	77.82	81.31	77.76

Table 1. Evaluation results (%) of entity-level F1-score on the test set of the CoNLL datasets. Results except ours were cited from the published literature.

Method	ar	hi	zh	Avg
BS	42.30	67.60	52.90	54.27
TSL	43.12	69.54	48.12	53.59
RIKD	45.96	70.28	50.40	55.55
MTMT	52.77	70.76	52.26	58.60
MSD	**62.88**	**73.43**	**57.06**	**64.46**
MSD w/o. distillers	54.52	70.22	52.46	59.06
MSD w/o. $\mathcal{L}^{L}_{\text{MMD}}$	56.93	71.50	56.68	61.70
MSD w/o. $\mathcal{L}^{M}_{\text{MMD}}$	58.65	72.11	56.53	62.43
MSD w/o. all	43.17	68.07	49.25	53.49

Table 2. Evaluation results (%) of entity-level F1-score on the test set of the WikiAnn dataset. Results except ours were cited from the published literature.

Method	ko	ru	tr	Avg
BS	51.78	52.33	58.85	54.32
TSL	53.91	54.26	61.15	56.44
AdvPicker	56.22	55.65	63.17	58.34
MSD	**61.67**	**58.06**	**67.80**	**62.51**
MSD w/o. distillers	57.23	56.81	65.14	59.72
MSD w/o. $\mathcal{L}^{L}_{\text{MMD}}$	57.88	57.24	67.83	60.98
MSD w/o. $\mathcal{L}^{M}_{\text{MMD}}$	59.12	58.08	67.41	61.53
MSD w/o. all	54.37	54.03	61.55	56.65

Table 3. Evaluation results (%) of entity-level F1-score on the test set of the mLOWNER dataset. Results except ours were obtained by re-implementing these baseline models with the source code provided by the original authors. 5 experiments under the same configuration were conducted for all the methods and the average results were taken as the final numbers. Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value <0.05).

The proposed MSD method significantly outperforms the baseline method TSL and achieves new state-of-the-art performance on all target languages.

The multi-channel distillation framework fully leverages the complementary and hierarchical information in the teacher model, and the unsupervised parallel domain adaptation method effectively pulls the domains between teacher and student models closer.

Mail to: {mjy1999,beiduo}@mail.ustc.edu.cn

The 2022 Conference on Empirical Methods in Natural Language Processing

Code at: https://github.com/Mckysse/MSD