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]

[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, Hefei, China
[4]Jilin Kexun Information Technology Co., Ltd.

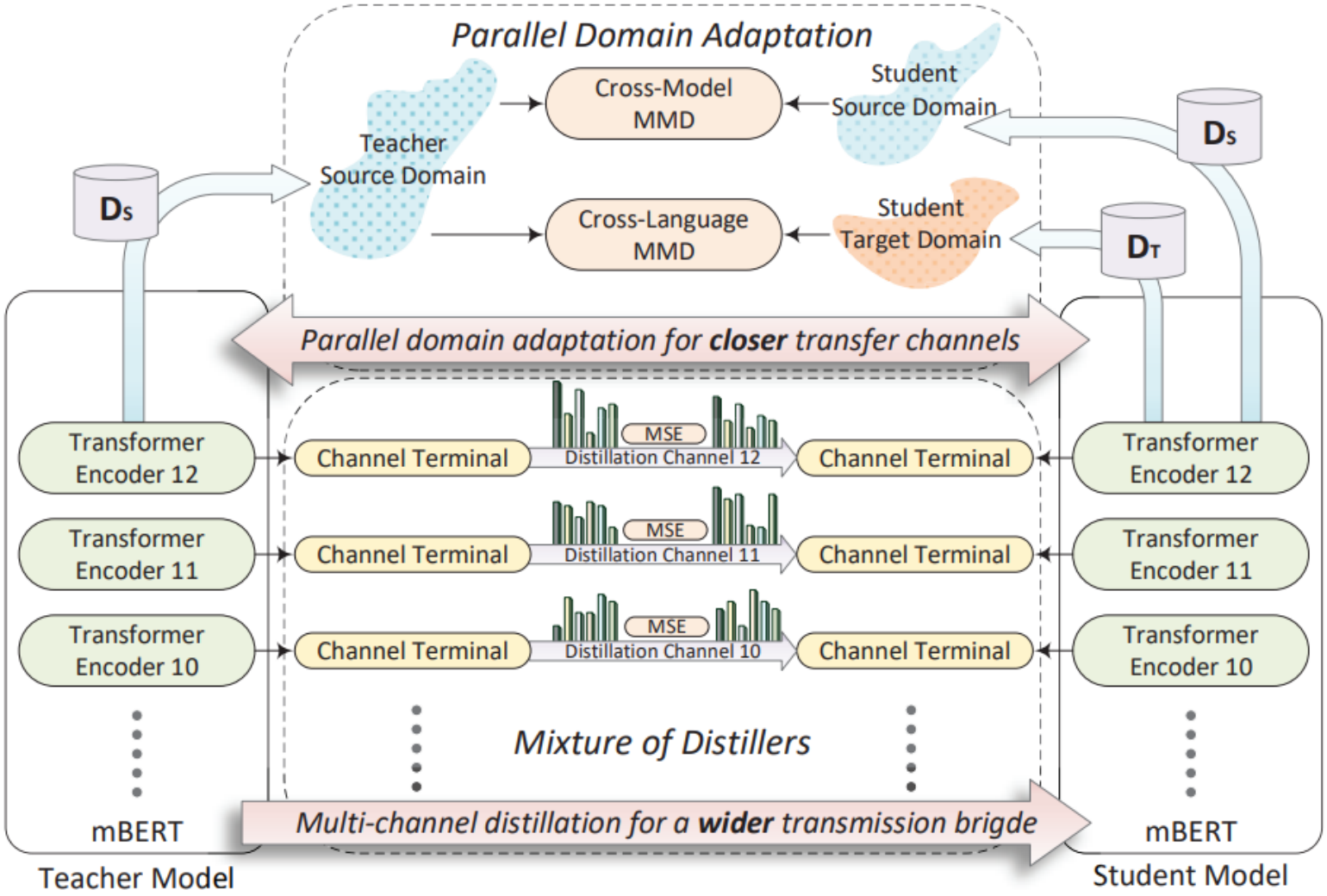
Outline

- **Introduction**
- Datasets & Basic System
- Mixture of Distillers
- Parallel Domain Adaptation
- Experiments

Introduction

- Task: zero-shot cross-lingual named entity recognition(NER)
- Challenge: few annotated data are available for some languages
- Method: the mixture of short-channel distillers (MSD).

The Architechture of the MSD



Outline

- Introduction
- **Datasets & Basic System**
- Mixture of Distillers
- Parallel Domain Adaptation
- Experiments

Datasets & Basic System

- Datasets
  - CoNLL-2002 : Spanish and Dutch;
  - CoNLL-2003 : English and German;
  - WikiAnn : English, Arabic, Hindi and Chinese;
  - mLOWNER : English,Korean, Farsi, and Turkish.
- Basic System: the mBERT is used as the pre-trained language model with a Softmax classifier predicts the tag of each token.

Outline

- Introduction
- Datasets & Basic System
- **Mixture of Distillers**
- Parallel Domain Adaptation
- Experiments

Mixture of Distillers

Each layer of the pre-trained mBERT is appended with a classifier. For the teacher, given a sentence x of length L with labels y from source language data D^S:

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

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

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

For the following knowledge distillation, a student model Θ_{stu} is distilled based on the unlabeled target language data D^T. Given a sentence x′ of length L from D^T train, these could be described as:

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

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

Outline

- Introduction
- Datasets & Basic System
- Mixture of Distillers
- **Parallel Domain Adaptation**
- Experiments

Parallel Domain Adaptation

Two MMD losses L^M_{MMD} and L^L_{MMD} are proposed to minimize the cross-model and cross-language discrepancies respectively.

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, $D^{T_{stu}}_{train}$ is obtained by applying the student model to the unlabeled target language data.

These losses could be formulated as:

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

Parallel Domain Adaptation

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

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

Parallel Domain Adaptation

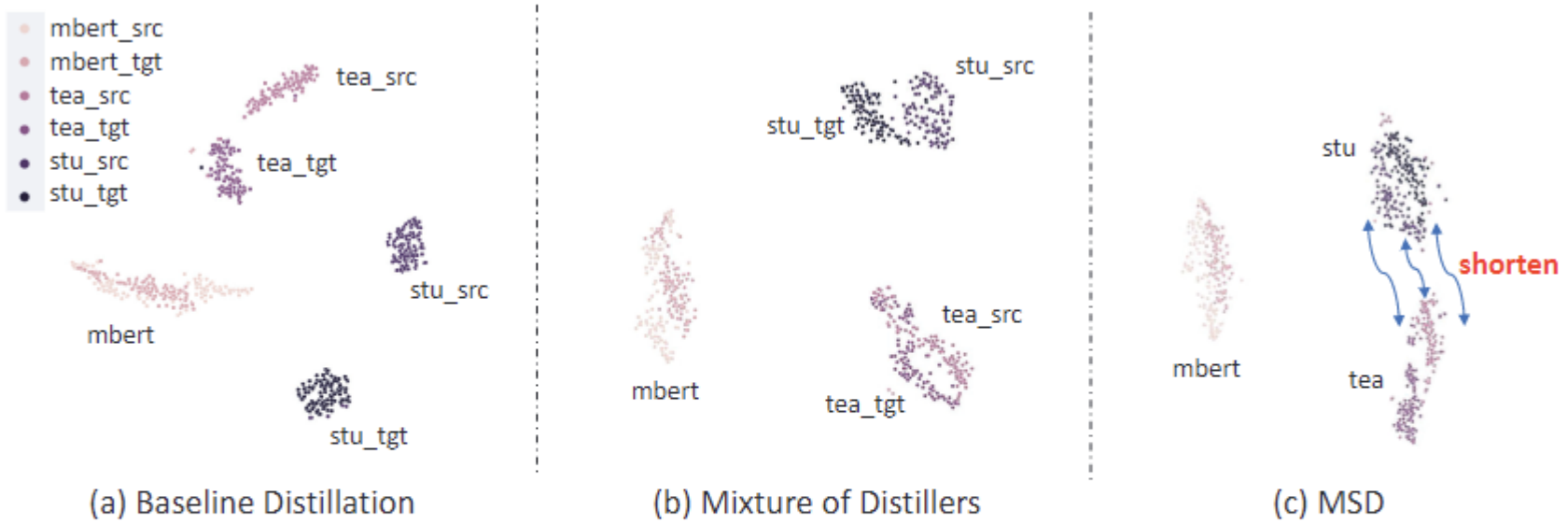


Figure 2. T-SNE visualization of semantic domains of different models by randomly sampling 100 unannotated English (source) and German (target) sentences from the training set of the CoNLL datasets.

Outline

- Introduction
- Datasets & Basic System
- Mixture of Distillers
- Parallel Domain Adaptation
- **Experiments**

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).

Thank You !!

Code at GitHub