Can Monolingual Pretrained Models Help Cross-Lingual Classification?

Zewen Chi[†*], **Li Dong**[‡], **Furu Wei**[‡], **Xian-Ling Mao**[†], **Heyan Huang**[†]
[†]Beijing Institute of Technology
[‡]Microsoft Research
{czw,maoxl,hhy63}@bit.edu.cn
{lidong1,fuwei}@microsoft.com

Abstract

Multilingual pretrained language models (such as multilingual BERT) have achieved impressive results for cross-lingual transfer. However, due to the constant model capacity, multilingual pre-training usually lags behind the monolingual competitors. In this work, we present two approaches to improve zero-shot cross-lingual classification, by transferring the knowledge from monolingual pretrained models to multilingual ones. Experimental results on two cross-lingual classification benchmarks show that our methods outperform vanilla multilingual fine-tuning.

1 Introduction

Supervised text classification heavily relies on manually annotated training data, while the data are usually only available in rich-resource languages, such as English. It requires great effort to make the resources available in other languages. Various methods have been proposed to build cross-lingual classification models by exploiting machine translation systems (Xu and Yang, 2017; Chen et al., 2018; Conneau et al., 2018), and learning multilingual embeddings (Conneau et al., 2018; Yu et al., 2018; Artetxe and Schwenk, 2019; Eisenschlos et al., 2019).

Recently, multilingual pretrained language models have shown surprising cross-lingual effectiveness on a wide range of downstream tasks (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Chi et al., 2020a,b). Even without using any parallel corpora, the pretrained models can still perform zero-shot cross-lingual classification (Pires et al., 2019; Wu and Dredze, 2019; Keung et al., 2019). That is, these models can be fine-tuned in a source language, and then directly evaluated in other target languages. Despite the effectiveness of cross-lingual transfer, the multilingual pretrained language models have their own drawbacks. Due to the constant number of model parameters, the model capacity of the rich-resource languages decreases if we adds languages for pre-training. The curse of multilinguality results in that the multilingual models usually perform worse than their monolingual competitors on downstream tasks (Arivazhagan et al., 2019; Conneau et al., 2020). The observations motivate us to leverage monolingual pretrained models to improve multilingual models for cross-lingual classification.

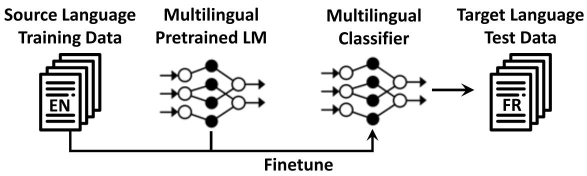
In this paper, we propose a multilingual fine-tuning method (MonoX) based on the teacher-student framework, where a multilingual student model learns end task skills from a monolingual teacher. Intuitively, monolingual pretrained models are used to provide supervision of downstream tasks, while multilingual models are employed for knowledge transfer across languages. We conduct experiments on two widely used cross-lingual classification datasets, where our methods outperform baseline models on zero-shot cross-lingual classification. Moreover, we show that the monolingual teacher model can help the student multilingual model for both the source language and target languages, even though the student model is only trained in the source language.

2 Background: Multilingual Fine-Tuning

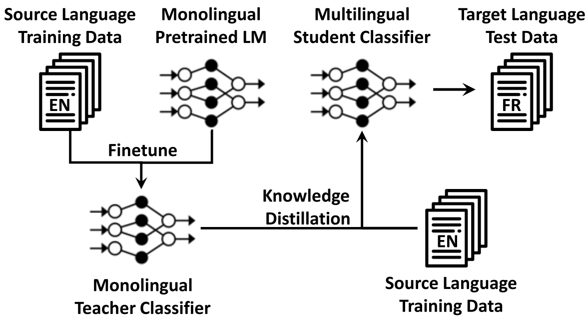
We use multilingual BERT (Devlin et al., 2019) for multilingual pretrained language models. The pretrained model uses the BERT-style Transformer (Vaswani et al., 2017) architecture, and follows the similar fine-tuning procedure as BERT for text classification, which is illustrated in Figure 1(a). To be specific, the first input token of the models is always a special classification token

[*]Contribution during internship at Microsoft Research.

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 12–17



(a) Multilingual LM finetuning for cross-lingual classification



(b) Finetuning procedure of MonoX-KD

Figure 1: Illustration of multilingual LM fine-tuning. (a) The original multilingual LM fine-tuning procedure for cross-lingual classification. (b) The fine-tuning procedure of our proposed MONOX via knowledge distillation (MONOX-KD). Notice that MONOX does not use any target language data during fine-tuning.

`[CLS]`. During fine-tuning, the final hidden state of the special token is used as the sentence representation. In order to output predictions, an additional softmax classifier is built on top of the sentence representation. Denoting $\mathcal{D}$ as the training data in the source language, the pretrained models are fine-tuned with standard cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}}(\theta; \mathcal{D}) = -\sum_{(x,y)\in\mathcal{D}} \log p(y|x;\theta)$$

where θ represents model parameters. Then the model is directly evaluated on other languages for cross-lingual classification.

3 Methods

As shown in Figure 1(b), we first fine-tune the monolingual pretrained model in the source language. Then we transfer task knowledge to the multilingual pretrained model by soft (Section 3.1) or hard (Section 3.2) labels. We describe two variants of our proposed method (MONOX) as follows.

3.1 Knowledge Distillation

In order to transfer task-specific knowledge from monolingual model to multilingual model, we propose to use knowledge distillation (Hinton et al., 2015) under our MONOX framework, where a student model s is trained with soft labels generated by a better-learned teacher model t. The loss function of the student model is:

$$\mathcal{L}_{\mathrm{KD}}(\theta_s; \mathcal{D}, \theta_t) =$$
$$-\sum_{(x,y)\in\mathcal{D}} \sum_{k=1}^{K} q(y=k|x;\theta_t) \log p(y=k|x;\theta_s)$$

where $p(\cdot)$ and $q(\cdot)$ represent the probability distribution over K categories, predicted by the student s and the teacher t, respectively. Notice that only the student model parameters θ_s are updated during knowledge distillation. As shown in Figure 1(b), we first use the fine-tuned monolingual pretrained model as a teacher, which is learned by minimizing $\mathcal{L}_{\mathrm{CE}}(\theta_t; \mathcal{D})$. Then we perform knowledge distillation for the student model with $\mathcal{L}_{\mathrm{KD}}(\theta_s; \mathcal{D}_C, \theta_t)$ as the loss function, where $\mathcal{D}_C$ is the concatenation of training dataset and the unlabeled dataset in the source language. We denote this implementation as MONOX-KD.

3.2 Pseudo-Label

In addition to knowledge distillation, we also consider implementing MONOX by training the student multilingual model with pseudo-label (Lee, 2013). Specifically, after fine-tuning the monolingual pretrained model on the training data as teacher, we apply the teacher model on the unlabeled data in the source language to generate pseudo labels. Next, we filter the pseudo labels by a prediction confidence threshold, and only keep the examples with higher confidence scores. Notice that the pseudo training data are assigned with hard labels. Finally, we concatenate the original training data and the pseudo data as the final training set for the student model. We denote this implementation as MONOX-PL.

4 Experiments

4.1 Experimental Setup

In the following experiments, we consider the zero-shot cross-lingual setting, where models are trained with English data and directly evaluated on all target languages.

Datasets We conduct experiments on two widely used datasets for cross-lingual evaluation: (1) Cross-Lingual Sentiment (CLS) dataset (Prettenhofer and Stein, 2010), containing Amazon reviews in three domains and four languages; (2) Cross-Lingual NLI (XNLI) dataset (Conneau et al., 2018), containing development and test sets in 15 languages and a training set in English for the natural language inference task.

Pretrained Language Models We use multilingual BERT$_{\text{BASE}}$[1] for cross-lingual transfer. For monolingual pretrained language model, the English-version RoBERTa$_{\text{LARGE}}$[2] is employed. All the pretrained models used in our experiments are cased models.

Baselines We compare our methods (MonoX-Kd, and MonoX-Pl) with the following models:

- mBert: directly fine-tuning the multilingual BERT$_{\text{BASE}}$ with English training data.
- mBert-St: fine-tuning the multilingual BERT$_{\text{BASE}}$ by self-training, i.e., alternately fine-tuning mBERT and updating the training data by labeling English unlabeled examples.

4.2 Configuration

For the CLS dataset, we randomly select 20% examples from training data as the development set and use the remaining examples as the training set. For XNLI, we randomly sample 20% examples from training data as the training set, and regard the other examples as the unlabeled set. We use the vocabularies provided by the pretrained models, which are extracted by Byte-Pair Encoding (Sennrich et al., 2016). The input sentences are truncated to 256 tokens. For both datasets, we use Adam optimizer with a learning rate of 5×10^{-6}, and a batch size of 8. We train models with epoch size of 200 and 2,500 steps for CLS and XNLI, respectively. For MonoX-Kd, the softmax temperature of knowledge distillation is set to 0.1. For MonoX-Pl, the confidence threshold is set to zero, which means all of the generated pseudo labels are used as training data.

4.3 Results and Discussion

Preliminary Experiments To see how much monolingual pretrained models is better than multilingual pretrained models, we finetune several different pretrained language models on the two datasets under the aforementioned configuration, and only evaluate them in English. As shown in Table 1, the gap between multilingual and monolingual pretrained models is large, even when using the same size of parameters. It is not hard to explain because mBert is trained in 104 languages, where different languages tend to confuse each other.

	Parameters	CLS	XNLI
Multilingual Pretrained Models			
mBert	110M	86.37	77.07
Monolingual Pretrained Models			
BERT$_{\text{BASE}}$	110M	90.10	80.46
RoBERTa$_{\text{BASE}}$	125M	93.82	85.09
RoBERTa$_{\text{LARGE}}$	355M	95.77	89.24

Table 1: Preliminary experiments results. Models are finetuned with English training data of CLS and XNLI under the configuration (see Section 3.2), and only evaluated in English. The results on CLS are averaged over three domains.

Sentiment Classification We evaluate our method on the zero-shot cross-lingual sentiment classification task. The goal of sentiment classification is to classify input sentences to positive or negative sentiments. In Table 2 we compare the results of our methods with baselines on CLS. It can be observed that our MonoX method outperforms baselines in all evaluated languages and domains, providing 4.91% improvement of averaged accuracy to the original multilingual BERT fine-tuning method. Notice that mBert-St is trained under the same condition with our method, i.e., using the same labeled and unlabeled data as ours. However, we only observe a slight improvement over mBert, which demonstrates that the performance improvement of MonoX mainly benefits from its end task knowledge transfer rather than the unlabeled data.

Natural Language Inference We also evaluate our method on the zero-shot cross-lingual NLI task, which is more challenging than sentiment classification. The goal of NLI is to identify the relationship of a pair of input sentences, including a premise and a hypothesis with an *entailment*, *contradiction*, or *neutral* relationship between them. As shown in Table 3, we present the evaluation results on XNLI. Unsurprisingly, both MonoX-Pl and MonoX-Kd perform better than baseline methods, showing that our method success-

[1] https://github.com/google-research/bert/blob/master/multilingual.md
[2] https://github.com/pytorch/fairseq/tree/master/examples/roberta

	en			de			fr			ja			
	Books	DVD	Music	Books	DVD	Music	Books	DVD	Music	Books	DVD	Music	avg
MBERT	87.75	86.60	84.75	79.55	75.90	77.05	81.45	80.35	80.35	75.15	76.90	75.90	80.14
MBERT-ST	88.20	85.50	88.00	79.65	76.70	80.00	84.85	83.25	80.55	74.60	75.80	76.90	81.17
MONOX-PL	**94.00**	**92.75**	91.80	83.20	79.25	82.95	**86.00**	84.95	**84.55**	78.85	**80.00**	79.35	84.80
MONOX-KD	93.90	91.40	**92.25**	**84.20**	**81.50**	**83.65**	85.40	**85.90**	83.95	**78.95**	79.15	**80.30**	**85.05**

Table 2: Evaluation results of zero-shot cross-lingual sentiment classification on the CLS dataset.

	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
MBERT	61.2	67.4	65.8	61.6	77.1	70.7	68.6	**53.4**	67.0	50.6	**44.6**	56.3	57.8	43.6	67.8	60.9
MBERT-ST	60.9	67.6	65.4	61.0	77.6	70.4	68.9	53.1	65.9	50.6	41.8	55.2	56.8	43.6	67.9	60.5
MONOX-PL	**63.5**	**70.1**	**69.8**	61.7	**80.9**	**74.1**	**72.1**	52.5	68.4	51.2	42.3	**57.9**	58.0	44.0	70.2	**62.5**
MONOX-KD	62.2	69.3	69.3	**62.1**	79.6	72.9	72.0	52.8	**68.6**	**52.3**	41.7	**57.9**	**58.5**	**45.9**	**70.8**	62.4

Table 3: Evaluation results of zero-shot cross-lingual NLI on the XNLI dataset. Note that 20% of the original training data are used as training set, and the other 80% are used as unlabeled set.

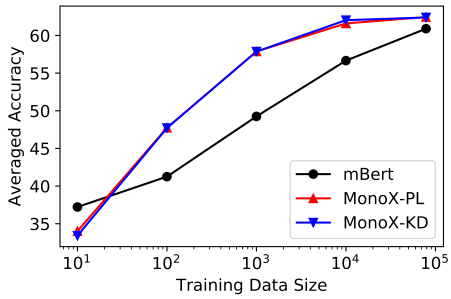


Figure 2: Averaged accuracy scores on zero-shot XNLI with different training data sizes. (20% and 80% of the training data are regraded training and unlabeled set.)

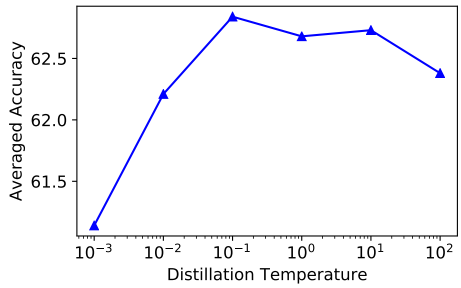


Figure 3: Averaged accuracy scores on the development set for zero-shot XNLI with different softmax temperatures of MONOX-KD.

fully helps the multilingual pretrained model gain end task knowledge from the monolingual pretrained model for cross-lingual classification. It is also worth mentioning that the performance of MBERT-ST is similar to MBERT. We believe the reason is that XNLI has more training data than CLS, which wakens the impact of self-training.

Effects of Training Data Size We conduct a study on how much multilingual pretrained model can learn from monolingual pretrained model for different training data size. We cut the training data to 10, 100, 1K, 10K and 78K (full training data in our setting) examples, and keep other hyper-parameters fixed. In Figure 2, we show the averaged accuracy scores for zero-shot XNLI with different training data sizes. We observe that MONOX outperforms MBERT on all data sizes except the 10-example setting. When the training data is relatively small ($\leq 10^4$), our method shows a great improvement.

Effects of Distillation Temperature Figure 3 presents XNLI averaged accuracy scores of MONOX-KD with different softmax temperatures in knowledge distillation. Even though the temperature varies from 10^{-3} to 10^2, all of the results are higher than baseline scores, which indicates MONOX-KD is nonsensitive to the temperature. When the temperature is set to 10^{-1}, we observe the best results on the development set. Therefore we set temperature as 0.1 in other experiments.

5 Conclusion

In this work, we investigated whether a monolingual pretrained model can help cross-lingual classification. Our results have shown that, with a RoBERTa model pretrained in English, we can boost the classification performance of a pretrained multilingual BERT in other languages. For future work, we will explore whether monolingual pretrained models can help other cross-lingual NLP tasks, such as natural language generation (Chi et al., 2020a).

Acknowledgements

Prof. Heyan Huang is the corresponding author. The work is supported by National Key R&D Plan (No. 2016QY03D0602), NSFC (No. U19B2020, 61772076, 61751201 and 61602197) and NSFB (No. Z181100008918002).

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, M. Gatu Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020a. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7570–7577. AAAI Press.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020b. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. *ArXiv, abs/2007.07834*.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. 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 5702–5707, Hong Kong, China. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

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.

Dong-Hyun Lee. 2013. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Vancouver, Canada. Association for Computational Linguistics.

Katherine Yu, Haoran Li, and Barlas Oguz. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 175–179.