A Robust Self-Learning Framework for Cross-Lingual Text Classification

Xin Dong
Rutgers University
New Brunswick, NJ, USA
xd48@rutgers.edu

Gerard de Melo
Rutgers University
New Brunswick, NJ, USA
gdm@demelo.org

Abstract

Based on massive amounts of data, recent pretrained contextual representation models have made significant strides in advancing a number of different English NLP tasks. However, for other languages, relevant training data may be lacking, while state-of-the-art deep learning methods are known to be data-hungry. In this paper, we present an elegantly simple robust self-learning framework to include unlabeled non-English samples in the fine-tuning process of pretrained multilingual representation models. We leverage a multilingual model's own predictions on unlabeled non-English data in order to obtain additional information that can be used during further fine-tuning. Compared with original multilingual models and other cross-lingual classification models, we observe significant gains in effectiveness on document and sentiment classification for a range of diverse languages.

1 Introduction

Owing to notable advances in deep learning and representation learning, important progress has been achieved on text classification, reading comprehension, and other NLP tasks. Recently, pretrained language representations with self-supervised objectives (Peters et al., 2018; Devlin et al., 2018; Radford et al., 2018) have further pushed forward the state-of-the-art on many English tasks. While these sorts of deep models can be trained on different languages, deep models typically require substantial amounts of labeled data for the specific domain of data.

Unfortunately, the cost of acquiring new custom-built resources for each combination of language and domain is very high, as it typically requires human annotation. Available resources for domain-specific tasks are often imbalanced between different languages. The scarcity of non-English annotated corpora may preclude our ability to train language-specific machine learning models. In contrast, English-language annotations are often readily available to train deep models. Although translation can be an option, human translation is very costly and for many language pairs, any available domain-specific parallel corpora are too small to train high-quality machine translation systems.

Cross-lingual systems rely on training data from one language to train a model that can be applied to other languages (de Melo and Siersdorfer, 2007), alleviating the training bottleneck issues for low-resource languages. This is facilitated by recent advances in learning joint multilingual representations (Lample and Conneau, 2019; Artetxe and Schwenk, 2018; Devlin et al., 2018).

In our work, we propose a self-learning framework to incorporate the predictions of the multilingual BERT model (Devlin et al., 2018) on non-English data into an English training procedure. The initial multilingual BERT model was simultaneously pretrained on 104 languages, and has shown to perform well for cross-lingual transfer of natural language tasks (Wu and Dredze, 2019). Our model begins by learning just from available English samples, but then makes predictions on unlabeled non-English samples and a part of those samples with high confidence prediction scores are repurposed to serve as labeled examples for a next iteration of fine-tuning until the model converges.

Based on this multilingual self-learning technique, we demonstrate the superiority of our framework on Multilingual Document Classification (MLDoc) (Schwenk and Li, 2018) in comparison with several strong baselines. Our study then proceeds to show that our method is better on Chinese sentiment classification than other cross-lingual methods that also consider unla-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6306–6310,
Hong Kong, China, November 3–7, 2019.

beled non-English data. This shows that our method is more effective at cross-lingual transfer for domain-specific tasks, using a mix of labeled and unlabeled data via a multilingual BERT sentence model.

2 Method

Our proposed framework consists of three parts as shown in Figure 1. The first part is the pretrained multilingual encoder, denoted as $f_n(\cdot;\theta_n)$. The encoder is assumed to have been pretrained across different languages with appropriate strategies, such as WordPiece, Byte Pair Encoding modeling, and Cross-lingual Word Alignment, to allow the model to share representations across languages to a certain degree. Hence, we obtain a universal sentence representation $\mathbf{h} \in \mathbb{R}^d$ from this encoder, where d is the dimensionality of the sentence representation. Subsequently, a task-specific classification module $f_{\text{cl}}(\cdot;\theta_{\text{cl}})$ is applied for fine-tuning on top of the pretrained model $f_n(\cdot;\theta_n)$. This module consists of a linear function mapping $\mathbf{h} \in \mathbb{R}^d$ into $\mathbb{R}^{|\mathcal{Y}|}$ and a softmax function, where $\mathcal{Y}$ is the set of target classes.

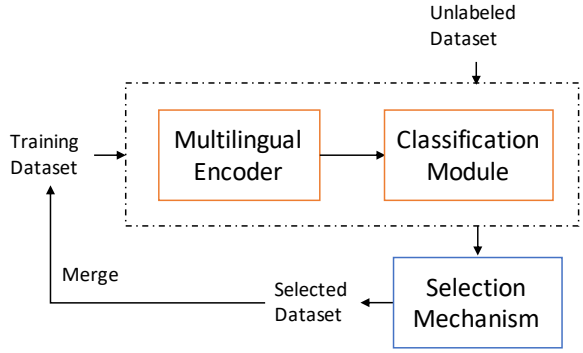

Figure 1: Illustration of self-learning process for cross-lingual classification.

For the overall process, we first train the whole network $f(\cdot;\theta)$ in K epochs using a set of the labeled data $L = \{(x_i, y_i) \mid i = 1, ..., n\}$, where n is the number of labeled instances, $x_i \in \mathcal{X}$ are instances, and $y_i \in \mathcal{Y}$ are the corresponding ground truth labels. The next step is to make predictions for the unlabeled instances in $U = \{x_u \mid u = 1, ..., m\}$. We assume that $f(\cdot;\theta)$ yields a class label as well as a confidence score.

To better take advantage of the pretrained multilingual model, a selection mechanism is invoked to repurpose unlabeled data with high confidence scores for incorporation into the training data. There are several variations of such selection mechanisms (Abney, 2007), and we rely on a *balancing* approach that considers the same number of instances for each class. Thus, we select a subset $\{x_s \mid s = 1, ..., K_t\}$ of the unlabeled data for each class, containing the top K_t highest confidence items based on the current trained model. The union set U_{s} of selected items is merged into the training set L and then we retrain the model and repeat this process iteratively. The detailed process is described in Algorithm 1, where for each class $y \in \mathcal{Y}$, $\mathcal{D}_y$ denotes the set of tuples pairing unlabeled data with corresponding confidence scores c.

Algorithm 1 Self-learning on cross-lingual tasks

repeat
  Fine-tune $f(\cdot;\theta)$ for K epochs using L
  for $y \in \mathcal{Y}$ **do**
    $\mathcal{D}_y \leftarrow \emptyset$
  for $x_{\text{u}} \in U$ **do**
    $(y, c) \leftarrow f(x_u;\theta)$
    $\mathcal{D}_y \leftarrow \mathcal{D}_y \cup \{(x_{\text{u}}, c)\}$
  for $y \in \mathcal{Y}$ **do**
    $S_y \leftarrow \underset{S \subseteq \mathcal{D}_y,\ |S| \leq K_t}{\text{argmax}} \sum_{(x_{\text{u}}, c) \in S} c$
    $U \leftarrow U \setminus S_y$
    $U_{\text{s}} \leftarrow U_{\text{s}} \cup S_y$
  $L \leftarrow L \cup U_{\text{s}}$
until stopping criterion is true

3 Experiments

We evaluate our self-learning framework on two cross-lingual document and sentiment classification tasks to show the effectiveness of self-learning for multilingual BERT-based cross-lingual transfer.

3.1 Experimental Setup

Datasets. For evaluation, we first rely on MLDoc (Schwenk and Li, 2018), a balanced subset of the Reuters corpus covering 8 languages for document classification, with 1,000 training and validation documents and 4,000 test documents for each language. The 4-way topic classification scheme consists of CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets) as targets. For cross-lingual classification, 1,000 target language training documents are used as unlabeled data for self-learning.

Table 1: Accuracy (in %) on MLDoc experiments. Bold denotes the best on cross-lingual transfer.

Approach	*en*	*de*	*zh*	*es*	*fr*	*it*	*ja*	*ru*
In-language supervised learning								
Schwenk and Li (2018)	92.2	93.7	87.3	94.5	92.1	85.6	85.4	85.7
BERT (Wu and Dredze, 2019)	94.2	93.3	89.3	95.7	93.4	88.0	88.4	87.5
Cross-lingual transfer								
Schwenk and Li (2018)	92.2	81.2	74.7	72.5	72.4	69.4	67.6	60.8
Artetxe and Schwenk (2018)	89.9	84.8	71.9	77.3	78.0	69.4	60.3	67.8
BERT	94.2	78.9	75.4	79.5	80.0	68.7	73.2	70.7
Our Approach	94.2	**90.0**	**87.0**	**85.3**	**88.4**	**75.2**	**76.8**	**79.3**

Table 2: Accuracy (in %) on Chinese sentiment classification without using labeled Chinese data. CLD and CLTC represent cross-lingual distillation and cross-lingual text classification.

Approach	ACC
Domain Adaptation	
mSDA (Chen et al., 2012)	31.44
Machine Translation	
DAN + MT (Chen et al., 2018)	39.66
CLD-based CLTC	
CLD-KCNN (Xu and Yang, 2017)	40.96
CLDFA-KCNN (Xu and Yang, 2017)	41.82
Cross-lingual Adversarial	
ADAN (Chen et al., 2018)	42.49
Cross-lingual transfer	
BERT	40.73
Our Approach	**43.88**

We further evaluate our method on cross-lingual sentiment classification from English to Chinese. For English, we use a balanced dataset of 700k Yelp reviews from Zhang et al. (2015) with their ratings as labels (scale 1–5) and adopting their training–validation split: 650k reviews for training and 50k as a validation set. For Chinese, we use the same dataset configuration as Chen et al. (2018), consisting of 150k unlabeled Chinese hotel reviews and 10k balanced Chinese hotel reviews as a validation set. The results are reported on a separate test set of another 10k hotel reviews. The data are also annotated with 5 labels (1–5). Both classification tasks are evaluated in terms of classification accuracy (ACC).

Model Details. We tune the hyper-parameters for our neural network architecture based on each non-English validation set. For the encoder, we invoke the multilingual BERT model (Devlin et al., 2018), which supports 104 languages[1]. It relies on a shared 110k WordPiece vocabulary across all languages and yields sentence representations in a common multilingual space. Most model hyper-parameters are the same as in pretraining, with the exception of the batch size, max. sequence length, and number of training epochs. The batch size, max. sequence length and number of training epochs used for the MLDoc task are 128, 32, and 4, respectively, while they are 128, 96, and 3 for sentiment classification. Another hyper-parameter involved in self-learning is K_t, which is 40 for MLDoc and 100 for sentiment classification. We rely on early stopping as a termination criterion.

[1]2018-11-23 version from https://github.com/google-research/bert/blob/master/multilingual.md

3.2 Results and Analysis

Cross-lingual Document Classification. Three recent strong baselines are included in our MLDoc experiments. Schwenk and Li (2018) use MultiCCA, multilingual word embeddings trained with a bilingual dictionary and convolution neural networks. Artetxe and Schwenk (2018) pretrain a multilingual sentence representation with a massively multilingual sequence-to-sequence NMT model, where the encoder is used for fine-tuning downstream tasks. We also considered multilingual BERT without self-learning as one of our baselines. As shown in Table 1, our framework significantly outperforms all baselines in 7 languages on cross-lingual document classification, including for phylogenetically unrelated languages. We also show the respective percentages of instances added into the training set that are correct for the MLDoc data based on our method in Table 3. The high percentages of correctly labeled incorporated instances in 7 languages further show the effectiveness of self-learning in our framework.

Table 3: Percentages of instances added into the training set that are correct for the MLDoc data using our method.

de	*zh*	*es*	*fr*	*it*	*ja*	*ru*
93%	92 %	89 %	92%	80%	84%	89%

Cross-lingual Sentiment Classification. To evaluate the robustness of our framework on cross-lingual sentiment classification, we consider several diverse baselines as listed in Table 2. mSDA (Chen et al., 2012) is a very effective method for cross-domain sentiment classification on Amazon reviews, which can also be used in cross-lingual tasks, but it has the worst performance. Deep Averaging Networks (DANs) by Iyyer et al. (2015) consider an arithmetic mean of word vectors as a sentence representation and pass it to a classification module, while Chen et al. (2018) translate the Chinese test text into English as a machine translation baseline. The third category of baselines includes Xu and Yang (2017), who propose a cross-lingual distillation (CLD) method that makes use of soft source predictions on a parallel corpus to train a target model (CLD-KCNN). They further propose an improved variant (CLDFA-KCNN) that utilizes adversarial training for domain adaptation within a single language. Adversarial DAN (ADAN) by Chen et al. (2018) is another state of the art baseline that improves cross-lingual generalization by means of adversarial training. We also run experiments on multilingual BERT and observe that it does not outperform CLD-based CLTC and ADAN, while our approach achieves the new state-of-the-art result, indicating that our self-learning method for cross-lingual transfer can be more effective than a diverse range of other approaches. In addition, we evaluate the proximity between incorrect prediction and the corresponding correct label in our sentiment task by means of mean squared error. The error of our method is 1.37, while for regular multilingual BERT it is 1.42, which also shows the superiority of our method.

Comparison of Selection Mechanisms. We ran experiments on two selection mechanisms. One is *balancing*, as described in Section 2. An alternative is *throttling* by selecting the top n unlabeled examples without considering specific classes (Abney, 2007). Our experiments on MLDoc show that the results suffer from a rapid decline during self-learning with *throttling*. This is because selecting from all samples leads to an imbalance between different classes and due to repeated error amplification this means that samples are increasingly likely to be assigned to the majority class in each self-learning iteration.

4 Related Work

Semi-supervised Learning. There is a long history of research on semi-supervised Learning to exploit unlabeled data. Self-learning (also known as self-training) was successfully applied to NLP tasks in early work such as on word sense disambiguation (Yarowsky, 1995) and parsing (McClosky et al., 2006). In recent work, Artetxe et al. (2018) show that self-learning can iteratively improve unsupervised cross-lingual word embeddings. Clark et al. (2018) presents Cross-View Training, a new self-training algorithm that works well for neural sequence modeling. Other semi-supervised methods, such as co-training (Blum and Mitchell, 1998) and tri-training (Zhou and Li, 2005), have as well been used for sentiment analysis. Ruder and Plank (2018) propose a novel multi-task tri-training method that reduces the time and space complexity of classic tri-training for sentiment analysis. For cross-lingual sentiment analysis, Wan (2009) uses machine translation to directly convert English training data to Chinese, which provides two views for co-training. Xu and Yang (2017) propose to use soft probabilistic predictions for the documents in a label-rich language as the (induced) supervisory labels in a parallel corpus of documents, while there is no need to use parallel corpora in our work. Chen et al. (2018) propose an Adversarial Deep Averaging Network to learn invariance across languages, which is another baseline considered in our experiments.

Cross-lingual Representation Learning. With models such as ELMo (Peters et al., 2018), GPT-2 (Radford et al., 2018), and BERT (Devlin et al., 2018), important progress has been made in learning improved sentence representations with context-specific encodings of words via a language modeling objective. The latter two approaches both rely on Transformer encoders, but BERT is trained using masked language modeling instead of right-to-left or left-to-right language modeling. Additionally, BERT also optimizes a next sentence classification objective. Recent work has also investigated cross-lingual extensions. Devlin et al. (2018) themselves published a multilingual version of BERT, following the same model architecture and training procedure, except that the union of 104 different language editions of Wikipedia serves as the training input. Lample and Conneau (2019) incorporate parallel text into BERT's architecture by

training on a new supervised learning objective. Artetxe and Schwenk (2018) also show that the encoder from a pretrained sequence-to-sequence model can be used to produce cross-lingual sentence embeddings. All these methods are compatible with our self-learning framework, since they provide a shared sentence meaning representation across languages as needed by our approach.

5 Conclusion

In this work, we propose a self-learning framework for cross-lingual text classification. Based on the cross-lingual prediction ability of pretrained multilingual model, this elegantly simple framework makes the most of unlabeled text to improve cross-lingual transfer for text classification. We achieve new state-of-the-art results on cross-lingual document and sentiment classification and demonstrate that self-learning is an effective method for improved classification accuracy without target language training data.

References

Steven Abney. 2007. *Semisupervised learning for computational linguistics*. Chapman and Hall/CRC.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*.

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.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. 2018. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of ACL-IJCNLP 2015*, volume 1.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of NAACL-HLT 2006*.

Gerard de Melo and Stefan Siersdorfer. 2007. Multilingual text classification using ontologies. In *Proceedings of ECIR 2007*. Springer.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. *arXiv preprint arXiv:1804.09530*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. *arXiv preprint arXiv:1805.09821*.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL-IJCNLP 2009*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. *arXiv preprint arXiv:1705.02073*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge & Data Engineering*, (11):1529–1541.