Adversarial Learning with Contextual Embeddings for Zero-resource Cross-lingual Classification and NER

Phillip Keung, Yichao Lu, Vikas Bhardwaj
Amazon Inc.
{keung, yichaolu, vikab}@amazon.com

Abstract

Contextual word embeddings (e.g. GPT, BERT, ELMo, etc.) have demonstrated state-of-the-art performance on various NLP tasks. Recent work with the multilingual version of BERT has shown that the model performs very well in cross-lingual settings, even when only labeled English data is used to finetune the model. We improve upon multilingual BERT's zero-resource cross-lingual performance via adversarial learning. We report the magnitude of the improvement on the multilingual MLDoc text classification and CoNLL 2002/2003 named entity recognition tasks. Furthermore, we show that language-adversarial training encourages BERT to align the embeddings of English documents and their translations, which may be the cause of the observed performance gains.

1 Introduction

Contextual word embeddings (Devlin et al., 2019; Peters et al., 2018; Radford et al., 2019) have been successfully applied to various NLP tasks, including named entity recognition, document classification, and textual entailment. The multilingual version of BERT (which is trained on Wikipedia articles from 100 languages and equipped with a 110,000 shared wordpiece vocabulary) has also demonstrated the ability to perform 'zero-resource' cross-lingual classification on the XNLI dataset (Conneau et al., 2018). Specifically, when multilingual BERT is finetuned for XNLI with English data alone, the model also gains the ability to handle the same task in other languages. We believe that this zero-resource transfer learning can be extended to other multilingual datasets.

In this work, we explore BERT's[1] zero-resource performance on the multilingual MLDoc classification and CoNLL 2002/2003 NER tasks. We

find that the baseline zero-resource performance of BERT exceeds the results reported in other work, even though cross-lingual resources (e.g. parallel text, dictionaries, etc.) are not used during BERT pretraining or finetuning. We apply adversarial learning to further improve upon this baseline, achieving state-of-the-art zero-resource results.

There are many recent approaches to zero-resource cross-lingual classification and NER, including adversarial learning (Chen et al., 2019; Kim et al., 2017; Xie et al., 2018; Joty et al., 2017), using a model pretrained on parallel text (Artetxe and Schwenk, 2018; Lu et al., 2018; Lample and Conneau, 2019) and self-training (Hajmohammadi et al., 2015). Due to the newness of the subject matter, the definition of 'zero-resource' varies somewhat from author to author. For the experiments that follow, 'zero-resource' means that, during model training, we do not use labels from non-English data, nor do we use human or machine-generated parallel text. Only labeled English text and unlabeled non-English text are used during training, and hyperparameters are selected using English evaluation sets.

Our contributions are the following:

- We demonstrate that the addition of a language-adversarial task during finetuning for multilingual BERT can significantly improve the zero-resource cross-lingual transfer performance.

- For both MLDoc classification and CoNLL NER, we find that, even without adversarial training, the baseline multilingual BERT performance can exceed previously published results on zero-resource performance.

- We show that adversarial techniques encourage BERT to align the representations of En-

---

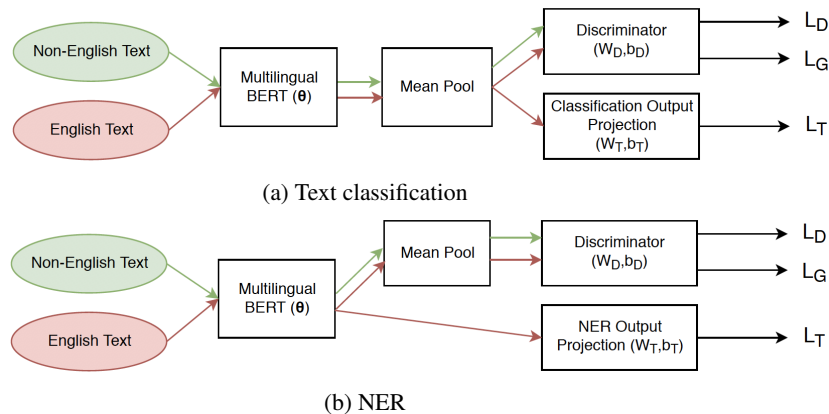[1] 'BERT' hereafter refers to multilingual BERT

Figure 1: Overview of the adversarial training process for classification and NER. All input text is in the form of a sequence of word pieces. L_D, L_G, L_T refer to the discriminator, generator and task-specific losses. Parameters of each component is in round brackets.

glish documents and their translations. We speculate that this alignment causes the observed improvement in zero-resource performance.

2 Related Work

2.1 Adversarial Learning

Language-adversarial training (Zhang et al., 2017) was proposed for generating bilingual dictionaries without parallel data. This idea was extended to zero-resource cross-lingual tasks in NER (Kim et al., 2017; Xie et al., 2018) and text classification (Chen et al., 2019), where we would expect that language-adversarial techniques induce features that are language-independent.

2.2 Self-training Techniques

Self-training, where an initial model is used to generate labels on an unlabeled corpus for the purpose of domain or cross-lingual adaptation, was studied in the context of text classification (Hajmohammadi et al., 2015) and parsing (McClosky et al., 2006; Zeman and Resnik, 2008). A similar idea based on expectation-maximization, where the unobserved label is treated as a latent variable, has also been applied to cross-lingual text classification in Rigutini et al. (2005).

2.3 Translation as Pretraining

Artetxe and Schwenk (2018) and Lu et al. (2018) use the encoders from machine translation models as a starting point for task-specific finetuning, which permits various degrees of multilingual transfer. Lample and Conneau (2019) add an additional masked translation task to the BERT pre-

training process, and the authors observed an improvement in the cross-lingual setting over using the monolingual masked text task alone.

3 Experiments

3.1 Model Training

We present an overview of the adversarial training process in Figure 1. We used the pretrained cased multilingual BERT model[2] as the initialization for all of our experiments. Note that the BERT model has 768 units.

We always use the labeled English data of each corpus. We use the non-English text portion (without the labels) for the adversarial training.

We formulate the adversarial task as a binary classification problem (i.e. English versus non-English.) We add a *language discriminator* module which uses the BERT embeddings to classify whether the input sentence was written in English or the non-English language. We also add a *generator* loss which encourages BERT to produce embeddings that are difficult for the discriminator to classify correctly. In this way, the BERT model learns to generate embeddings that do not contain language-specific information.

The pseudocode for our procedure can be found in Algorithm 1. In the description that follows, we use a batch size of 1 for clarity.

For language-adversarial training for the classification task, we have 3 loss functions: the task-specific loss L_T, the generator loss L_G, and the

---

[2]https://github.com/google-research/bert/blob/master/multilingual.md

discriminator loss L_D:

$$L_T(y^T; x) = \sum_{i=1}^{K} -y_i^T \log p(Y = i|x)$$

$$p(Y|x) = \text{Softmax}(W_T \bar{h}_\theta(x) + b_T)$$

$$L_G(y^A; x) = -(1 - y^A)\log p(E = 1|x)$$
$$- y^A \log p(E = 0|x)$$

$$L_D(y^A; x) = -(1 - y^A)\log p(E = 0|x)$$
$$- y^A \log p(E = 1|x)$$

$$p(E = 1|x) = \text{Sigmoid}(w_D \cdot \bar{h}_\theta(x) + b_D)$$

where K is the number of classes for the task, $p(Y|x)$ (dim: $K \times 1$) is the task-specific prediction, $p(E = 1|x)$ (dim: scalar) is the probability that the input is in English, $\bar{h}_\theta(x)$ (dim: 768×1) is the mean-pooled BERT output embedding for the input word-pieces x, θ is the BERT parameters, W_T, b_T, w_D, b_D (dim: $K \times 768$, $K \times 1$, 768×1, scalar) are the output projections for the task-specific loss and discriminator respectively, y^T (dim: $K \times 1$) is the one-hot vector representation for the task label and y^A (dim: scalar) is the binary label for the adversarial task (i.e. 1 or 0 for English or non-English).

In the case of NER, the task-specific loss has an additional summation over the length of the sequence:

$$L_T(y^T; x) = \sum_{i=1}^{K} \sum_{t=1}^{L} -y_{it}^T \log p(Y_t = i|x)$$

$$p(Y_t|x) = \text{Softmax}(W_T h_\theta(x)_t + b_T)$$

where $p(Y_t|x)$ (dim: $K \times 1$) is the prediction for the tth word, L is the number of words in the sentence, y^T (dim: $K \times L$) is the matrix of one-hot entity labels, and $h_\theta(x)_t$ (dim: 768×1) refers to the BERT embedding of the tth word.

The generator and discriminator losses remain the same for NER, and we continue to use the mean-pooled BERT embedding during adversarial training.

We then take the gradients with respect to the 3 losses and the relevant parameter subsets. The parameter subsets are $\theta_D = \{w_D, b_D\}$, $\theta_T = \{\theta, W_T, b_T\}$, and $\theta_G = \{\theta\}$. We apply the gradient updates sequentially at a 1:1:1 ratio.

During BERT finetuning, the learning rates for the task loss, generator loss and discriminator loss were kept constant; we do not apply a learning rate decay.

All hyperparameters were tuned on the English dev sets only, and we use the Adam optimizer in all experiments. We report results based on the average of 4 training runs.

3.2 MLDoc Classification Results
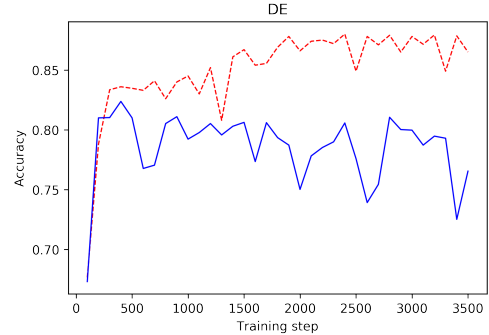
(a) German test accuracy vs steps taken
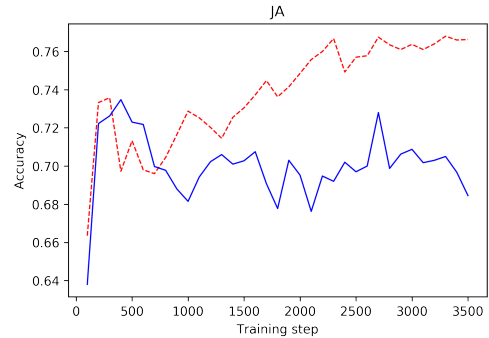
(b) Japanese test accuracy vs steps taken

Figure 2: German and Japanese MLDoc test accuracy versus the number of training steps, with and without adversarial training. The solid line shows the performance of the non-adversarial BERT baseline. The dashed line shows the performance with adversarial training.

We finetuned BERT on the English portion of the MLDoc corpus (Schwenk and Li, 2018). The MLDoc task is a 4-class classification problem, where the data is a class-balanced subset of the Reuters News RCV1 and RCV2 datasets. We used the `english.train.1000` dataset for the classification loss, which contains 1000 labeled documents. For language-adversarial training, we used the text portion of `german.train.10000`, `french.train.10000`, etc. without the labels.

1357

Algorithm 1 Pseudocode for adversarial training on the multilingual text classification task. The batch size is set at 1 for clarity. The parameter subsets are $\theta_D = \{w_D, b_D\}$, $\theta_T = \{\theta, W_T, b_T\}$, and $\theta_G = \{\theta\}$.

Input: pre-trained BERT model h_θ, data iterators for English and the non-English language L, learning rates η_D, η_G, η_T for each loss function, initializations for discriminator output projection w_D, b_D, task-specific output projection W_T, b_T, and BERT parameters θ

1: **while** not converged **do**
2:      $x_{En}, y^T \leftarrow \text{DataIterator}(En)$ ▷ get English text and task-specific labels
3:      $\bar{h}_{En} \leftarrow \text{MeanPool}(h_\theta(x_{En}))$
4:      $p^T \leftarrow \text{Softmax}(W_T \bar{h}_{En} + b_T)$ ▷ compute the prediction for the task
5:      $L_T \leftarrow -y^T \cdot \log p^T$ ▷ compute task-specific loss
6:      $\theta, W_T, b_T \mathrel{+}= -\eta_T \nabla_{\theta_T} L_T$ ▷ update model based on task-specific loss
7:      $x_L, x_{En} \leftarrow \text{DataIterator}(L), \text{DataIterator}(En)$ ▷ get non-English and English text
8:      $\bar{h}_L, \bar{h}_{En} \leftarrow \text{MeanPool}(h_\theta(x_L)), \text{MeanPool}(h_\theta(x_{En}))$
9:      $p_L^D \leftarrow \text{Sigmoid}(w_D \cdot \bar{h}_L + b_D)$ ▷ discriminator prediction on non-English text
10:      $p_{En}^D \leftarrow \text{Sigmoid}(w_D \cdot \bar{h}_{En} + b_D)$ ▷ discriminator prediction on English text
11:      $L_D \leftarrow -\log p_{En}^D - \log(1 - p_L^D)$ ▷ compute discriminator loss
12:      $w_D, b_D \mathrel{+}= -\eta_D \nabla_{\theta_D} L_D$ ▷ update model based on discriminator loss
13:      $x_L, x_{En} \leftarrow \text{DataIterator}(L), \text{DataIterator}(En)$
14:      $\bar{h}_L, \bar{h}_{En} \leftarrow \text{MeanPool}(h_\theta(x_L)), \text{MeanPool}(h_\theta(x_{En}))$
15:      $p_L^D \leftarrow \text{Sigmoid}(w_D \cdot \bar{h}_L + b_D)$
16:      $p_{En}^D \leftarrow \text{Sigmoid}(w_D \cdot \bar{h}_{En} + b_D)$
17:      $L_G \leftarrow -\log(1 - p_{En}^D) - \log p_L^D$ ▷ compute generator loss
18:      $\theta \mathrel{+}= -\eta_G \nabla_{\theta_G} L_G$ ▷ update model based on generator loss

	En	De	Es	Fr	It	Ja	Ru	Zh
Schwenk and Li (2018)	92.2	81.2	72.5	72.4	69.4	67.6	60.8	74.7
Artetxe and Schwenk (2018)	89.9	84.8	77.3	77.9	69.4	60.3	67.8	71.9
BERT En-labels	94.2	79.8	72.1	73.5	63.7	72.8	73.7	76.0
BERT En-labels + Adv.	-	**88.1**	**80.8**	**85.7**	**72.3**	**76.8**	**77.4**	**84.7**

Table 1: Classification accuracy on the MLDoc test sets. We present results for BERT finetuned on labeled English data and BERT finetuned on labeled English data with language-adversarial training. Our results are averaged across 4 training runs, and hyperparameters are tuned on English dev data.

We used a learning rate of 2×10^{-6} for the task loss, 2×10^{-8} for the generator loss and 5×10^{-5} for the discriminator loss.

In Table 1, we report the classification accuracy for all of the languages in MLDoc. Generally, adversarial training improves the accuracy across all languages, and the improvement is sometimes dramatic versus the BERT non-adversarial baseline.

In Figure 2, we plot the zero-resource German and Japanese test set accuracy as a function of the number of steps taken, with and without adversarial training. The plot shows that the variation in the test accuracy is reduced with adversarial training, which suggests that the cross-lingual performance is more consistent when adversarial training is applied. (We note that the batch size and learning rates are the same for all the languages in MLDoc, so the variation seen in Figure 2 are not affected by those factors.)

3.3 CoNLL NER Results

We finetuned BERT on the English portion of the CoNLL 2002/2003 NER corpus (Sang and De Meulder, 2003). We followed the text preprocessing in Devlin et al. (2019).

We used a learning rate of 6×10^{-6} for the task loss, 6×10^{-8} for the generator loss and 5×10^{-4} for the discriminator loss.

In Table 2, we report the F1 scores for all of the CoNLL NER languages. When combined with adversarial learning, the BERT cross-lingual F1 scores increased for German over the

	En	De	Es	Nl
Devlin et al. (2019)	92.4	-	-	-
Mayhew et al. (2017)	-	57.5	66.0	64.5
Ni et al. (2017)	-	58.5	65.1	65.4
Chen et al. (2019)	-	56.0	73.5	72.4
Xie et al. (2018)	-	57.8	72.4	71.3
BERT En-labels	91.1	68.6	**75.0**	77.5
BERT En-labels + Adv.	-	**71.9**	74.3	**77.6**

Table 2: F1 scores on the CoNLL 2002/2003 NER test sets. We present results for BERT finetuned on labeled English data and BERT finetuned on labeled English data with language-adversarial training. Our results are averaged across 4 training runs, and hyperparameters are tuned on English dev data.

non-adversarial baseline, and the scores remained largely the same for Spanish and Dutch. Regardless, the BERT zero-resource performance far exceeds the results published in previous work.

Mayhew et al. (2017) and Ni et al. (2017) do use some cross-lingual resources (like bilingual dictionaries) in their experiments, but it appears that BERT with multilingual pretraining performs better, even though it does not have access to cross-lingual information.

3.4 Alignment of Embeddings for Parallel Documents

Source	Target	Without Adv.	With Adv.
	De	0.74	**0.94**
	Es	0.72	**0.94**
	Fr	0.73	**0.94**
En	It	0.73	**0.92**
	Ja	0.65	**0.84**
	Ru	0.72	**0.89**
	Zh	0.69	**0.91**

Table 3: Median cosine similarity between the mean-pooled BERT embeddings of MLDoc English documents and their translations, with and without language-adversarial training. The median cosine similarity increased with adversarial training for every language pair, which suggests that the adversarial loss encourages BERT to learn language-independent representations.

If language-adversarial training encourages language-independent features, then the English documents and their translations should be close in the embedding space. To examine this hypothesis, we take the English documents from the MLDoc training corpus and translate them into German,

Spanish, French, etc. using Amazon Translate.

We construct the embeddings for each document using BERT models finetuned on MLDoc. We mean-pool each document embedding to create a single vector per document. We then calculate the cosine similarity between the embeddings for the English document and its translation. In Table 3, we observe that the median cosine similarity increases dramatically with adversarial training, which suggests that the embeddings became more language-independent.

4 Discussion

For many of the languages examined, we were able to improve on BERT's zero-resource cross-lingual performance on the MLDoc classification and CoNLL NER tasks. Language-adversarial training was generally effective, though the size of the effect appears to depend on the task. We observed that adversarial training moves the embeddings of English text and their non-English translations closer together, which may explain why it improves cross-lingual performance.

Future directions include adding the language-adversarial task during BERT pre-training on the multilingual Wikipedia corpus, which may further improve zero-resource performance, and finding better stopping criteria for zero-resource cross-lingual tasks besides using the English dev set.

References

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

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Mohammad Sadegh Hajmohammadi, Roliana Ibrahim, Ali Selamat, and Hamido Fujita. 2015. Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. *Information sciences*, 317:67–77.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks: Application to community question answering. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Conference on Machine Translation (WMT)*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.

Leonardo Rigutini, Marco Maggini, and Bing Liu. 2005. An em based training algorithm for cross-language text categorization. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 529–535. IEEE Computer Society.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.