Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training

Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, Kai-Wei Chang
University of California, Los Angeles
{khhuang, wasiahmad, violetpeng, kwchang}@cs.ucla.edu

Abstract

Pre-trained multilingual language encoders, such as multilingual BERT and XLM-R, show great potential for zero-shot cross-lingual transfer. However, these multilingual encoders do not precisely align words and phrases across languages. Especially, learning alignments in the multilingual embedding space usually requires sentence-level or word-level parallel corpora, which are expensive to be obtained for low-resource languages. An alternative is to make the multilingual encoders more robust; when fine-tuning the encoder using downstream task, we train the encoder to tolerate noise in the contextual embedding spaces such that even if the representations of different languages are not aligned well, the model can still achieve good performance on zero-shot cross-lingual transfer. In this work, we propose a learning strategy for training robust models by drawing connections between adversarial examples and the failure cases of zero-shot cross-lingual transfer. We adopt two widely used robust training methods, adversarial training and randomized smoothing, to train the desired robust model. The experimental results demonstrate that robust training improves zero-shot cross-lingual transfer on text classification tasks. The improvement is more significant in the *generalized* cross-lingual transfer setting, where the pair of input sentences belong to two different languages.

1 Introduction

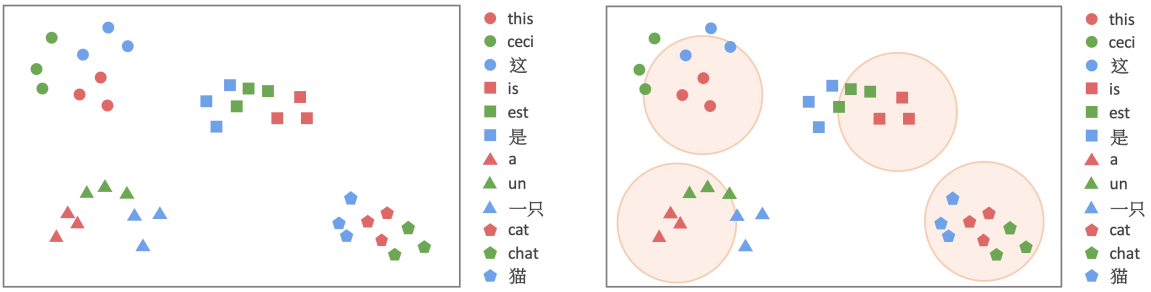
Zero-shot cross-lingual transfer learning aims to learn models with data available in one or more source languages and use them in other target languages for which there is no data (zero-resource) available. The zero-shot cross-lingual transfer has a great practical value for low-resource languages since it reduces the requirement of labeled data to learn models for downstream tasks, e.g., text classification (Conneau et al., 2018; Yang et al., 2019) and question answering (Lewis et al., 2020).

Recently, pre-trained multilingual language encoders, such as multilingual BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a), demonstrate promising performance on zero-shot cross-lingual transfer learning for a wide range of downstream tasks (Hu et al., 2020; Liang et al., 2020). These language encoders learn a shared multilingual contextual embedding space; they are able to represent word pairs in parallel sentences with similar contextual representations. However, the multilingual encoders fail to capture this similarity when the source and target languages are less similar at levels of morphology, syntax, and semantics (Ahmad et al., 2019a,b).

Prior studies (Cao et al., 2020; Pan et al., 2021; Dou and Neubig, 2021) have shown that aligning the representations of different languages in the multilingual embedding space plays an important role for zero-shot cross-lingual transfer learning. As illustrated in Figure 1a, words with similar meanings (e.g. *this*, *ceci*, and 这) have similar representations in the contextual multilingual embedding space, even though these words are in different languages. This alignment helps models transfer the learned knowledge from source languages to target languages. Therefore, several works focus on improving the quality of alignments in the multilingual embedding space (Cao et al., 2020; Chi et al., 2020; Pan et al., 2021; Dou and Neubig, 2021). Nevertheless, learning such alignments usually requires sentence-level or word-level parallel corpora, which are expensive to be obtained for low-resource languages. In addition, because the meanings of words in different languages are usually not exactly matched, learn a perfect alignment could be impossible.

In this work, we start from another point of view to improve zero-shot cross-lingual transfer performance. We aim to make the multilingual encoders robust such that they can tolerate a certain amount of noise in the input embeddings. More specifi-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1684–1697
November 7–11, 2021.



(a) Contextual representations of different words.

(b) Robust regions try to cover neighbor embeddings.

Figure 1: An illustration of different words in the multilingual contextual embedding space. (a) Words with similar meanings in different languages have similar representations but they are not exactly aligned. (b) We aim to learn a robust classifier whose robust regions (orange circles) that cover as many neighbor words as possible.

cally, as shown in Figure 1b, we target to construct *robust regions* (orange circles) for embeddings in the multilingual embedding space. During training, the robust model is expected to output similar predictions for embeddings in the same robust region. Therefore, as long as similar words in different languages fall into the same robust region, even if they are not perfectly aligned, the model can still have similar predictions for them.

To learn the robust model, we first draw connections between adversarial examples (Li et al., 2020; Garg and Ramakrishnan, 2020; Jin et al., 2020) and the failure cases of zero-shot cross-lingual transfer, and then study two widely used robust training methods to learn the robust model: (1) adversarial training (Goodfellow et al., 2015; Madry et al., 2018) and (2) randomized smoothing (Cohen et al., 2019; Ye et al., 2020). Both of them can make the model robust against perturbations in the input embeddings by modifying the training objective when fine-tuning model for the downstream task. For randomized smoothing, we also adopt the data augmentation approach (Ye et al., 2020) to learn the robust model.

We perform experiments on two cross-lingual text classification tasks, paraphrase identification and natural language inference[1]. The experimental results demonstrate that robust training indeed improves the performance of zero-shot cross-lingual transfer on the classification benchmarks: PAWS-X (Yang et al., 2019) and XNLI (Conneau et al., 2018). On average the cross-lingual transfer performance improves by 2.1 and 1.6 points on PAWS-X and XNLI, respectively. In addition, we show that robust training remarkably improves *generalized* cross-lingual transfer (Lewis et al., 2020). In this setting, the pair of input sentences in the text classification tasks belong to two different languages, e.g., paraphrase prediction for a pair of sentences in English and Korean.

2 Related Work

Zero-shot cross-lingual transfer learning. In recent years, several pre-trained multilingual language models are proposed for zero-shot cross-lingual transfer, including multilingual BERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), and XLM-R (Conneau et al., 2020a; Goyal et al., 2021). Many studies put attentions on the rationales that make zero-shot cross-lingual transfer work (K et al., 2020; Lauscher et al., 2020; Conneau et al., 2020b; Artetxe et al., 2020; Dufter and Schütze, 2020). Various tasks and datasests are presented to facilitate zero-shot cross-lingual transfer learning (Conneau et al., 2018; Yang et al., 2019; Clark et al., 2020; Artetxe et al., 2020; Lewis et al., 2020). XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) further provide benchmarks for zero-shot cross-lingual transfer learning.

Embedding space alignments. Learning to align embedding spaces have always been an important research topic to improve multilinguality. Early works focus on word embedding spaces (Mikolov et al., 2013; Smith et al., 2017; Artetxe et al., 2017). Recently, many approaches are proposed to align contextual word embedding spaces, such as learning rotation projections (Schuster et al., 2019; Aldarmaki and Diab, 2019; Conneau et al., 2020b) and fine-tuning pre-trained multilingual language models (Chi et al., 2020; Feng et al., 2020; Cao et al., 2020; Qin et al., 2020; Liu et al., 2020; Dou and Neubig, 2021; Wei et al., 2021).

[1]Our code is available at https://github.com/uclanlp/Robust-XLT

However, most of them require additional supervision signals, such as parallel sentence pairs (Chi et al., 2020; Feng et al., 2020; Wei et al., 2021), bilingual dictionary (Cao et al., 2020; Qin et al., 2020; Liu et al., 2020), or both (Pan et al., 2021). These additional supervised corpora are usually expensive for low-resource languages.

Embedding misalignment handling. Instead of directly aligning the representations, there is a line of research making the model be aware of the embedding misalignment issues by considering additional syntactic features, such as part-of-speech (Kozhevnikov and Titov, 2013) and dependency parse trees (Ahmad et al., 2019b; Subburathinam et al., 2019; Zhang et al., 2019; Liu et al., 2019; Ahmad et al., 2021a,b), and other syntactic features (Meng et al., 2019). However, those syntactic features require large human efforts to obtained.

Robust training. Recently, adversarial attacks are presented to check the robustness of NLP models, such as character manipulation (Ebrahimi et al., 2018; Gil et al., 2019), word replacements (Alzantot et al., 2018; Li et al., 2020; Garg and Ramakrishnan, 2020; Jin et al., 2020), and syntactic rearrangements (Iyyer et al., 2018). To against those attacks, various robust training methods are proposed. For example, Alzantot et al. (2018) trains a robust model by data augmentation with generated adversarial examples. Other works (Ebrahimi et al., 2018; Dong et al., 2021; Zhou et al., 2021) consider adversarial training, which includes the adversarial accuracy to the training objective. A few studies propose transformations on inputs before feeding them to models (Edizel et al., 2019; Jones et al., 2020). Randomized smoothing (Cohen et al., 2019; Ye et al., 2020) is presented to make models robust against noise in input representations. Another line of research aims at providing theoretical guarantee of robustness, including interval bound propagation methods (Jia et al., 2019; Huang et al., 2019) and verification methods (Shi et al., 2020). Most of those robust training methods focus on defending adversarial attacks, while we propose to apply robust training methods to improve the zero-shot cross-lingual transfer performance.

3 Zero-Shot Cross-Lingual Transfer with Robust Training

In this work, we focus on zero-shot cross-lingual transfer for text classification tasks. Our goal is to learn a classifier f from a set of training examples in source languages $X_{src} = \{(x_i, y_i)\}_{i=1}^{N}$. At test time, we directly use the classifier f to conduct inference on a set of test examples in target languages $X_{tgt} = \{x_i\}_{i=1}^{M}$. We expect that the classifier f can transfer the learned knowledge from the source languages to the target languages.

3.1 Connection with Adversarial Examples

The aligned representations of different languages have been shown as a crucial factor (Cao et al., 2020; Chi et al., 2020; Pan et al., 2021) for multilingual embeddings to be effective for zero-shot cross-lingual transfer. For example, assuming the source language and the target language are English and French, respectively, and considering a pair of parallel sentences "*this is a cat*" (in English) and "*Ceci est un chat*" (in French), we can get the contextual representations of the source English sentence $\mathbf{E}_{src} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4)$ and the target French sentence $\mathbf{E}_{tgt} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4)$. Let $\boldsymbol{\delta}$ denote the difference between the source and the target contextual representations as follows.[2]

$$\begin{aligned}\boldsymbol{\delta} &= \mathbf{E}_{src} - \mathbf{E}_{tgt} \\ &= (\mathbf{v}_1 - \mathbf{u}_1, \mathbf{v}_2 - \mathbf{u}_2, \mathbf{v}_3 - \mathbf{u}_3, \mathbf{v}_4 - \mathbf{u}_4) \\ &= (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3, \boldsymbol{\delta}_4).\end{aligned}$$

Since words with similar meanings have similar representations, the norm of their differences $\|\boldsymbol{\delta}_i\|$ is supposed to be small. Therefore, if $f(\mathbf{E}_{src}) = c$, we have a high probability for $f(\mathbf{E}_{tgt}) = c$ as well, which means that the classifier f is able to transfer the learned knowledge from the source language to the target language. If unfortunately, the transfer fails, we have

$$\begin{aligned} &f(\mathbf{E}_{tgt}) = f(\mathbf{E}_{src} + \boldsymbol{\delta}) \neq f(\mathbf{E}_{src}), \\ &\text{where } \|\boldsymbol{\delta}_i\| \text{ is small.} \end{aligned} \tag{1}$$

We observe that Eq. (1) is very similar to the definition of *adversarial examples* (Alzantot et al., 2018; Li et al., 2020; Garg and Ramakrishnan, 2020; Jin et al., 2020). The goal of adversarial examples is to find a small perturbation $\boldsymbol{\Delta}$ for an instance $\mathbf{x}$ such that a classifier h changes the prediction on $\mathbf{x}$, as illustrated by the following equation.

$$\begin{aligned} &h(\tilde{\mathbf{x}}) = h(\mathbf{x} + \boldsymbol{\Delta}) \neq h(\mathbf{x}), \\ &\text{where } \|\boldsymbol{\Delta}\| \text{ is small.} \end{aligned} \tag{2}$$

[2]For the ease of describing our idea, we assume the word orders in different languages are the same. Later in experiments, we relax this condition and present a preliminary study on the influence of word orders in Section 4.4.

For the case that cross-lingual transfer fails, the difference between the source and target representations $\boldsymbol{\delta}$ behaves like an adversarial perturbation. This inspires us to consider robust training methods, which are designed for defending adversarial examples, to improve the zero-shot cross-lingual transfer performance. More specifically, our goal is to train a robust classifier that can tolerate small perturbations on input embeddings. As shown in Figure 1b, we aim to train a robust classifier f that has robust regions (orange circles) such that the robust classifier f outputs similar values for input embeddings are in the same robust region.

We study two widely used robust training methods in literature: (1) adversarial training and (2) randomized smoothing, as they have been successfully used for defending adversarial attacks (Ebrahimi et al., 2018; Jia et al., 2019; Huang et al., 2019; Cohen et al., 2019).

3.2 Adversarial Training

The main idea of adversarial training is considering the most effective adversarial perturbation in each optimization iteration. More precisely, in normal training, we learn a classifier f by solving the following optimization problem

$$\min_{f} \sum_{(x,y)\in X_{src}} \mathcal{L}(f(\text{Enc}(x)), y),$$

where $\text{Enc}(\cdot)$ is the multilingual encoder and $\mathcal{L}$ is the cross-entropy loss. When considering adversarial training, we solve the following min-max optimization problem instead

$$\min_{f} \sum_{(x,y)\in X_{src}} \max_{\|\boldsymbol{\delta}_i\|\leq\varepsilon} \mathcal{L}(f(\text{Enc}(x)+\boldsymbol{\delta}), y),$$

where ε is a hyper-parameter to control the size of robust regions which are described by several norm balls $\|\boldsymbol{\delta}_i\|$. The inner maximization finds the most effective perturbation to change the prediction, while the outer minimization tries to ensure the correct prediction against the perturbation. With this min-max optimization, the classifier f is aware of perturbations within the robust regions $\|\boldsymbol{\delta}_i\|$ and becomes more robust.

3.3 Randomized Smoothing

Unlike adversarial training, which always considers the most effective perturbation, randomized smoothing focuses on the expectation case and aims to guarantee the local smoothness of the classifier at the same time. Following previous work (Cohen et al., 2019; Ye et al., 2020), we let f be the classifier learned by solving the normal optimization problem and learn a smoothed classifier g such that

$$g(\text{Enc}(x)) = \arg\max_{c\in\mathcal{Y}} \mathbb{P}_{\boldsymbol{\delta}}(f(\text{Enc}(x)+\boldsymbol{\delta}) = c),$$

where $\mathbb{P}_{\boldsymbol{\delta}}$ is a prior distribution of the perturbation $\boldsymbol{\delta}$ and $\mathcal{Y}$ is the label space. In other words, we want that $g(\text{Enc}(x))$ has a similar output value (label predictions) to $f(\text{Enc}(x))$. The random perturbation $\boldsymbol{\delta}$ is introduced to ensure the local smoothness of g. That is, $g(\text{Enc}(x)+\boldsymbol{\delta})$, the output for the perturbed input, is similar to the output value of $g(\text{Enc}(x))$. Compared to the original classifier f, the smoothed classifier g is more robust against local perturbations.

We consider two different ways to learn the smoothed classifier g: (1) random perturbation and (2) data augmentation.

Random perturbation (RP). Specifically, we focus on the following objective

$$\min_{g} \sum_{(x,y)\in X_{src}} \mathbb{P}_{\boldsymbol{\delta}}(\mathcal{L}(g(\text{Enc}(x)+\boldsymbol{\delta}), y)).$$

In each optimization step, we randomly sample a perturbation $\boldsymbol{\delta}$ from $\mathbb{P}_{\boldsymbol{\delta}}$ and add it to $\text{Enc}(x)$. Then, we use the perturbed representation as the input to calculate the loss and update the classifier g.

Data augmentation (DA). Another common way to approximate the smoothed classifier g is data augmentation (Ye et al., 2020). Instead of randomly sampling the perturbation $\boldsymbol{\delta}$, we consider a predefined synonym set (Alzantot et al., 2018). For every example $x = (w_1, w_2, ..., w_n)$ in X_{src}, we generate m augmented examples by replacing each word w_i in x with one of its synonym words (including w_i itself). We allow multiple replacements in one example. Then, we use the augmented data to train a smoothed classifier g.

It is worth noting that the predefined synonym set is required for *only* source languages. Unlike previous work (Qin et al., 2020; Liu et al., 2020), which uses bilingual dictionary of *both* source languages and target languages, the proposed method does not need any additional annotations of target languages.

Model	en	de	es	fr	ja	ko	zh	avg.
mBERT*	94.0	85.7	87.4	87.0	73.0	69.6	77.0	82.0
mBERT (reproduce)	93.7	85.4	88.2	87.8	75.3	74.2	79.1	83.4
mBERT-ADV	93.7	86.5	88.5	87.8	76.1	75.3	80.4	84.0
mBERT-RS-RP	**94.5**	87.4	**90.0**	**89.5**	77.9	77.5	**82.0**	**85.5**
mBERT-RS-DA	93.5	**87.8**	88.8	88.8	**79.3**	**78.3**	81.5	85.4

Table 1: Averaged results of zero-shot cross-lingual transfer on PAWS-X with 10 different random seeds. Highest scores are in bold. Underlines denote that the improvement is significant with $p \leq 0.05$ for the bootstrapped paired t-test. *We report the numbers in the previous paper (Hu et al., 2020).

Model	en	ar	bg	de	el	es	fr	hi
mBERT*	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9
mBERT (reproduce)	82.3	64.8	68.2	70.8	66.4	74.3	73.7	59.7
mBERT-ADV	81.9	64.9	68.3	71.7	66.5	74.4	74.5	59.6
mBERT-RS-RP	**82.6**	65.4	68.7	70.5	67.2	**75.0**	74.1	59.8
mBERT-RS-DA	81.0	**66.4**	**69.9**	**71.8**	**68.0**	74.7	**74.2**	**62.7**
Model	ru	sw	th	tr	ur	vi	zh	avg.
mBERT*	67.8	49.7	54.1	60.9	57.2	69.3	67.8	65.4
mBERT (reproduce)	68.7	50.0	53.0	60.9	57.7	70.3	69.2	66.0
mBERT-ADV	68.8	48.8	50.6	61.7	59.2	70.0	69.4	66.0
mBERT-RS-RP	69.5	48.4	50.5	59.7	57.9	70.5	69.7	66.0
mBERT-RS-DA	**70.6**	**51.1**	**55.7**	**62.9**	**60.9**	**71.8**	**71.4**	**67.6**

Table 2: Averaged results of zero-shot cross-lingual transfer on XNLI with 10 different random seeds. Highest scores are in bold. Underlines denote that the improvement is significant with $p \leq 0.05$ for the bootstrapped paired t-test. *We report the numbers in the previous paper (Hu et al., 2020).

4 Experiments

We conduct experiments to verify that robust training indeed improves the performance of zero-shot cross-lingual transfer.

4.1 Setup

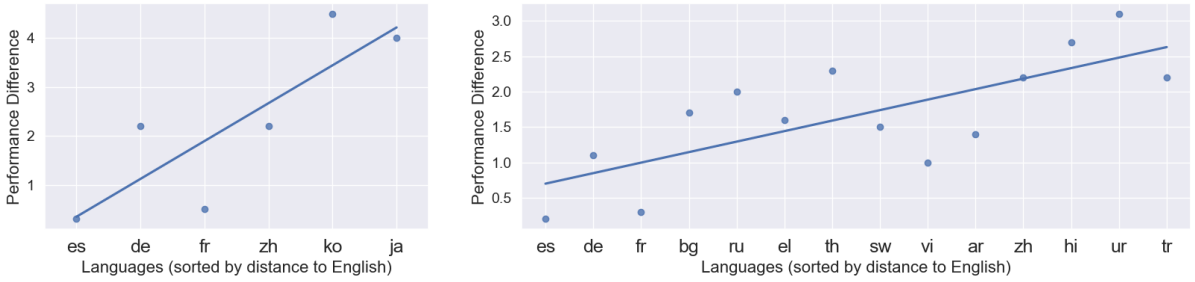
We consider two cross-lingual text classification datasets: Cross-lingual Paraphrase Adversaries from Word Scrambling (PAWS-X) (Yang et al., 2019) and Cross-lingual Natural Language Inference (XNLI) (Conneau et al., 2018). The goal of PAWS-X is to determine whether two sentences are paraphrases to each other or not. XNLI is designed for natural language inference; given a premise and a hypothesis, the classifier predicts the relation of the two sentences from {*entailment*, *neutral*, *contradiction*}.

For both datasets, we consider English as the source language and treat other languages as the target languages. We use the train, validation, and test splits provided by XTREME framework (Hu et al., 2020). Specifically, we conduct 10 runs of experiments with 10 different random seeds. In each run, we train the classifier on the English training set, use the English validation set to search the best parameters, and record the results of the test sets. Finally, the averaged results of 10-run experiments are reported.

Compared models. We consider the following four different models:

- **mBERT**: the standard multilingual BERT (Devlin et al., 2019).
- **mBERT-ADV**: multilingual BERT with adversarial training.
- **mBERT-RS-RP**: multilingual BERT with randomized smoothing via random perturbation.
- **mBERT-RS-DA**: multilingual BERT with randomized smoothing via data augmentation.

Implementation details. For adversarial training, we consider L_∞-norm as the norm of perturbation $\|\boldsymbol{\delta}_i\|$. The size of robust regions is searched



(a) PAWS-X

(b) XNLI

Figure 2: Performance difference between mBERT-RS-DA and mBERT over different languages. We sort the languages according to their distances to English from left (small) to right (large). Performance on languages with larger distances to English is improved more with the robust training.

from $\{0.001, 0.01, 0.1, 1.0\}$. For the randomized smoothing via random perturbation, we consider uniform distribution over a L_∞-norm ball. The size of ball is searched from $\{0.001, 0.01, 0.1, 1.0\}$. For the randomized smoothing via data augmentation, we consider the synonym set provide by previous work (Alzantot et al., 2018), which is constructed by searching nearest neighbors of words in the GloVe embedding space (Pennington et al., 2014) post-processed by the counter-fitting method (Mrksic et al., 2016). The number of augmented examples m is set to 10 and 3 for PAWS-X and XNLI, respectively, while more discussion on m is shown in Section 4.2. For other parameters, such as the learning rate and the batch size, we follow the training scripts provided by XTREME framework (Hu et al., 2020).

4.2 Zero-Shot Cross-Lingual Transfer

Table 1 shows the averaged results of PAWS-X with 10 different random seeds. We first notice that all mBERT-ADV, mBERT-RS-RP, and mBERT-RS-DA perform better than the standard mBERT on average. Especially, robust training leads to up to 4.0% improvement on Japanese, up to 4.1% improvement on Korean, and up to 2.9% improvement on Chinese. The results suggest that robust training helps in improving the performance of zero-shot cross-lingual transfer learning.

We observe that randomized smoothing is usually better than adversarial training. The reason is that adversarial training always considers the most effective adversarial perturbation during the optimization process. Adversarial perturbations are suitable for defending adversarial examples as they are specifically designed for attacking the classifier. However, in the zero-shot cross-lingual transfer case, the perturbations are not explicitly designed but reflect the natural difference between languages. Therefore, randomized smoothing, which considers the average case, becomes the better choice.

We have a similar conclusion for the XNLI dataset. As shown in Table 2, robust training indeed leads to improvements on zero-shot cross-lingual transfer. Again, randomized smoothing performs better than the adversarial training approach.

Finally, we compare the two different ways (random perturbation and data augmentation) to learn the smoothed classifier. They have competitive performance on PAWS-X; however, data augmentation performs better than random perturbation on XNLI. We hypothesize that the ideal robust regions in practice may not be perfect norm balls. In fact, they are more like convex hulls composed by the neighbor words (Dong et al., 2021). By considering a predefined synonym set, mBERT-RS-DA can better capture the shapes of robust regions, leading to a more stable performance.

What languages are benefited most from robust training? We notice that cross-lingual transfer to some languages is significantly improved by robust training, especially those languages that are quite different from the source language (English). To verify this conjecture, we consider lang2vec (Littell et al., 2017), a tool that extracts features of different languages by querying the URIEL typological database[3], to calculate the distance between English and other languages. Then, we show the performance gaps between mBERT-RS-DA and mBERT over all languages as well as the least square regression line in Figure 2. Note that the

[3] http://www.cs.cmu.edu/~dmortens/projects/7_project

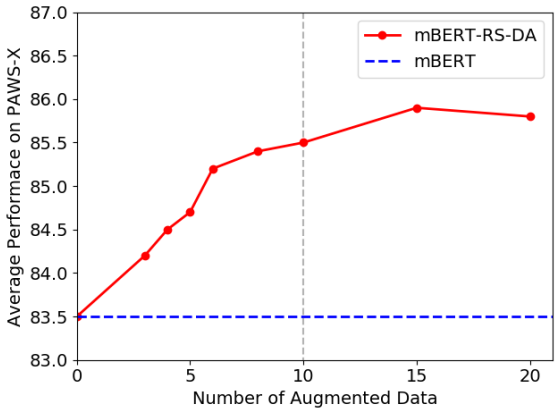


Figure 3: Performance of mBERT-RS-DA on PAWS-X over different m (the number of augmented instances generated by synonym replacements).

languages are sorted according to their distances to English from left to right.

From Figure 2a, we observe an obvious trend for PAWS-X that languages with larger distances to English have more performance gain with robust training. We posit that it is because languages with larger distances have more different representations from English in the multilingual embedding space. The norm of the perturbation $\boldsymbol{\delta}$ defined in Section 3 will be larger and thus the failure cases occur more often. By performing robust training, we reduce failure cases that lead to a larger improvement. Similar trend can be observed for XNLI (Figure 2b). Performance on languages with larger distances to English is improved more with the robust training.

How many augmented data needed for randomized smoothing? Since mBERT-RS-DA seems to be the most effective model for both PAWS-X and XNLI, we do further ablation on the number of augmented data for each example m. Figure 3 shows the average performance of mBERT-RS-DA on PAWS-X over different choices of m. We can observe that larger m leads to better performance in general because more augmented examples help the model better approximate the local smoothness, resulting in more accurate robust regions. Interestingly, when $m \leq 10$, increasing m can significantly improve the performance. When $m > 10$, increasing m only slightly improves the performance. This result suggests that setting m to 10 for PAWS-X. Interestingly, we observe that setting m to 3 is good enough for XNLI. This ablation study indicates that randomized smoothing with data augmentation can use just a few augmented instances per example to learn good robust regions.

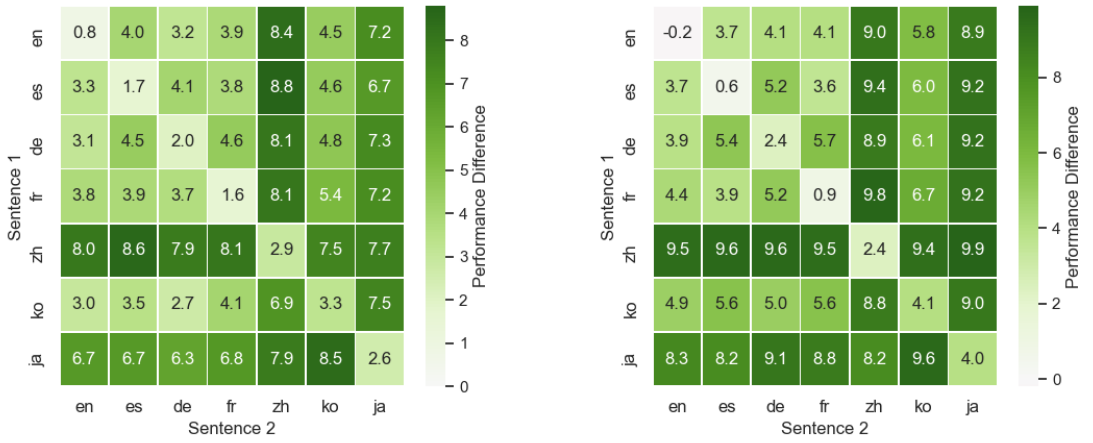
4.3 Zero-Shot Generalized Cross-Lingual Transfer Results

Next, we study the zero-shot cross-lingual transfer in a *generalized* setting. Lewis et al. (2020) proposed the generalized setting for the question answering task where the question and the context may belong to two different languages.[4] We consider the generalized setting for cross-lingual text classification since the input of PAWS-X and XNLI tasks are pairs of sentences. For example, consider XNLI on English-Arabic sentence pairs; the premises are in English, and the hypotheses are in Arabic. Note that due to the parallel nature of PAWS-X and XNLI dataset[5], we can pair up sentences from two different languages. Notice that we directly use the trained models in Section 4.2 to conduct inference in the generalized setting. In other words, all the classifiers are trained on English-English sentence pairs, without the consideration of target languages.

The results of mBERT-RS-RP and mBERT-RS-DA on PAWS-X and XNLI over all combinations of languages are shown in Figure 4 and Figure 5, respectively. While the diagonal numbers indicate the transfer results in the cross-lingual transfer settings, the non-diagonal entries present the generalized transfer performances. Note that we report the performance difference between the compared model and mBERT (exact numbers can be found in Appendix A) and the languages are sorted according to their distances to English. We observe that the non-diagonal numbers are much larger than the diagonal numbers, which suggests that robust training results in larger performance improvements in the generalized cross-lingual transfer setting. Given that the input sentences in training examples are in the same language (English), during inference, mBERT makes more mistakes in the classification tasks as the contextual representations for the input sentences may not be aligned accurately. However, mBERT-RS-RP and mBERT-RS-DA can tolerate a certain amount of noise in input embeddings. Therefore, they are more stable when the input sentences come from different languages, leading to a significant improvement.

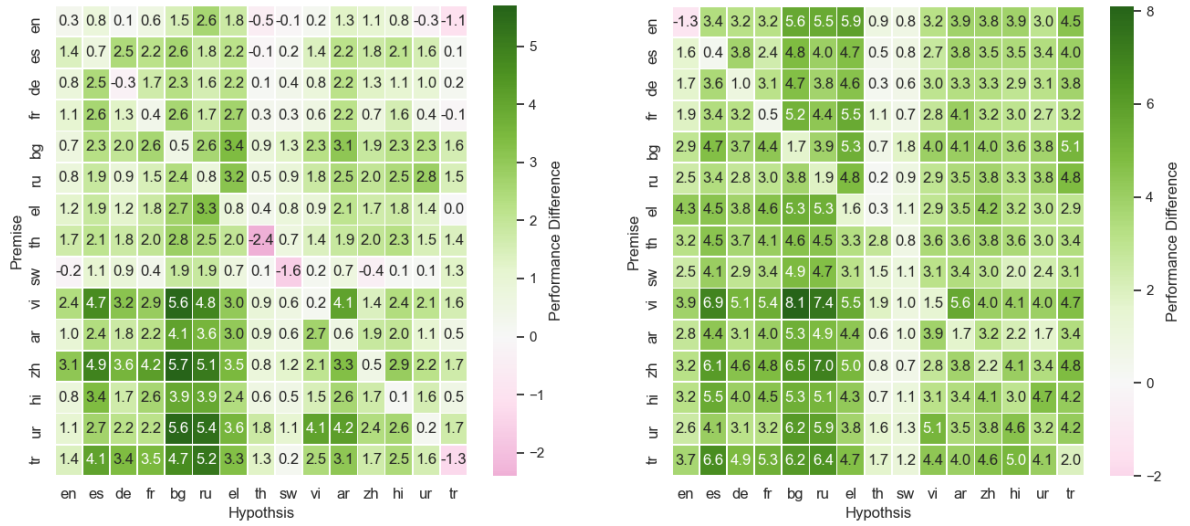
[4] QA systems should be able to answer questions written in French by reading an English context.

[5] PAWS-X and XNLI datasets consist of 7-way and 15-way parallel sentence pairs.



(a) mBERT-RS-RP (b) mBERT-RS-DA

Figure 4: Results for generalized zero-shot cross-lingual transfer on PAWS-X. We report the performance difference between the compared model and mBERT over different combinations of languages.



(a) mBERT-RS-RP (b) mBERT-RS-DA

Figure 5: Results for generalized zero-shot cross-lingual transfer on XNLI. We report the performance difference between the compared model and mBERT over different combinations of languages.

4.4 Study on Syntactic Perturbations

As mentioned in Section 3, our primary focus is on the perturbations in the multilingual embedding space and does not consider the influence of language syntax in cross-lingual transfer. Different languages have linguistic differences, such as word order. Differences in word order across languages affect the contextual embedding space that impacts cross-lingual transfer (Ahmad et al., 2019b). Therefore, we conduct a preliminary experiment to study the influence of syntax in robust training.

mBERT-RS-DA uses a predefined synonym set to generate perturbed examples for data augmentation. Following a similar strategy, we construct *syntactically perturbed examples* for data augmentation. More specifically, for every example $x = (w_1, w_2, ..., w_n)$ in X_{src}, we generate m syntactically perturbed examples by randomly swapping adjacent words with a probability $p = 0.1$. This random swapping may result in some examples with different word orders, which simulates the syntactic perturbations. Then, we use those syntactically perturbed examples to train the smoothed classifier g, called mBERT-RS-syntax.

Table 3 presents the preliminary results. The average performance of mBERT-RS-syntax is similar to the performance of standard mBERT. Interest-

Model	en	de	es	fr	ja	ko	zh	avg.
mBERT*	**94.0**	85.7	87.4	87.0	73.0	69.6	77.0	82.0
mBERT (reproduce)	93.7	85.4	88.2	87.8	75.3	74.2	79.1	83.4
mBERT-RS-DA	93.5	**87.8**	**88.8**	88.8	**79.3**	**78.3**	81.5	85.4
mBERT-RS-syntax	93.0	85.5	87.7	88.0	76.5	76.7	80.7	83.5

Table 3: Results of syntactic perturbations on PAWS-X. Highest scores are in bold. Underlines denote that the improvement is significant with $p \leq 0.05$ for the bootstrapped paired t-test. *We report the numbers in the previous paper (Hu et al. (2020)).

ingly, the zero-shot cross-lingual transfer performance drops when the target languages are more similar to the source language English (German, Spanish, and French), while the transfer performance increases when the target languages are more different from English (Japanese, Korean, and Chinese). This preliminary result suggests that it is possible to improve the zero-shot cross-lingual transfer by considering syntactic perturbations. One potential extension is adopting paraphrase generation models (Iyyer et al., 2018; Huang and Chang, 2021) to construct more sophisticated syntactic perturbations and we leave this direction for future work.

5 Conclusion

In this work, we propose a robust model by drawing connections between adversarial examples and the failure cases of zero-shot cross-lingual transfer. We adopt two robust training methods, adversarial training and randomized smoothing, to train the desired robust model. The experimental results demonstrate that robust training improves zero-shot cross-lingual transfer on text classification tasks. In addition, the improvement is more significant in the generalized cross-lingual transfer setting.

Acknowledgments

We thank anonymous reviewers for their helpful feedback. We thank UCLA-NLP group for the valuable discussions and comments. This work is supported in part by a Google Research Scholar Award, an Amazon Research Award, and the Intelligence Advanced Research Projects Activity (IARPA) via Contract No. 2019-19051600007.

References

Wasi Uddin Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021a. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021b. GATE: graph attention transformer encoder for cross-lingual relation and event extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019a. Cross-lingual dependency parsing with unlabeled auxiliary languages. In *The 2019 SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard H. Hovy, Kai-Wei Chang, and Nanyun Peng. 2019b. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Hanan Aldarmaki and Mona T. Diab. 2019. Context-aware cross-lingual mapping. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *8th International Conference on Learning Representations (ICLR)*.

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

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguistics*, 8:454–470.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS)*.

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 (EMNLP)*.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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 (NAACL-HLT)*.

Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. 2021. Towards robustness against natural language word substitutions. In *9th International Conference on Learning Representations (ICLR)*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Philipp Dufter and Hinrich Schütze. 2020. Identifying necessary elements for bert's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Bora Edizel, Aleksandra Piktus, Piotr Bojanowski, Rui Ferreira, Edouard Grave, and Fabrizio Silvestri. 2019. Misspelling oblivious word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yotam Gil, Yoav Chai, Or Gorodissky, and Jonathan Berant. 2019. White-to-black: Efficient distillation of black-box adversarial attacks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, (ICLR)*.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. 2019. Achieving verified robustness to symbol substitutions via interval bound propagation. 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)*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. 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)*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations (ICLR)*.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavas. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, (EMNLP)*.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori S. Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019. Neural cross-lingual event detection with minimal parallel resources. 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*.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations(ICLR)*.

Tao Meng, Nanyun Peng, and Kai-Wei Chang. 2019. Target language-aware constrained inference for cross-lingual dependency parsing. 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)*.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. Counter-fitting word vectors to linguistic constraints. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Mo Yu, and Saloni Potdar. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. 2020. Robustness verification for transformers. In *8th International Conference on Learning Representations (ICLR)*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations (ICLR)*.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare R. Voss. 2019. Cross-lingual structure transfer for relation and event extraction. 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)*.

Xiangpeng Wei, Yue Hu, Rongxiang Weng, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *9th International Conference on Learning Representations (ICLR)*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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)*.

Mao Ye, Chengyue Gong, and Qiang Liu. 2020. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yue Zhang, Rui Wang, and Luo Si. 2019. Syntax-enhanced self-attention-based semantic role labeling. 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)*.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021. Defense against synonym substitution-based adversarial attacks via dirichlet neighborhood ensemble. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

A Detailed Results of Zero-Shot Generalized Cross-Lingual Transfer

Table 4, 5, and 6 show the result for mBERT, mBERT-RS-RP, mBERT-RS-DA on PAWS-X, respectively, while Table 7, 8, and 9 list the result for mBERT, mBERT-RS-RP, mBERT-RS-DA on XNLI, respectively. From those tables, we can observe that mBERT-RS-RP and mBERT-RS-DA lead to remarkable improvements.

	en	es	de	fr	zh	ko	ja	avg.
en	93.7	85.4	85.0	85.0	66.5	66.4	63.4	77.9
es	86.1	88.2	80.5	83.9	63.7	64.0	60.9	75.3
de	85.6	79.7	85.4	79.8	63.9	64.7	61.5	74.4
fr	84.8	83.0	80.3	87.8	63.7	63.9	61.1	74.9
zh	66.7	63.7	63.9	64.3	79.1	62.4	64.3	66.3
ko	67.1	64.9	65.3	65.0	62.7	74.2	65.1	66.3
ja	63.0	61.1	61.1	61.1	64.6	63.6	75.3	64.3
avg.	78.1	75.1	74.5	75.3	66.3	65.6	64.5	71.3

Table 4: Results for mBERT on PAWS-X.

	en	es	de	fr	zh	ko	ja	avg.
en	94.5	89.3	88.2	88.9	74.9	70.8	70.6	82.5
es	89.3	90.0	84.6	87.7	72.5	68.6	67.5	80.0
de	88.7	84.2	87.4	84.4	72.0	69.5	68.8	79.3
fr	88.6	86.9	83.9	89.5	71.8	69.3	68.3	79.7
zh	74.6	72.3	71.9	72.4	82.0	69.9	72.0	73.6
ko	70.1	68.4	68.0	69.0	69.6	77.5	72.5	70.7
ja	69.7	67.7	67.5	67.9	72.5	72.1	77.9	70.7
avg.	82.2	79.8	78.8	80.0	73.6	71.1	71.1	76.7

Table 5: Results for mBERT-RS-RP on PAWS-X.

	en	es	de	fr	zh	ko	ja	avg.
en	93.5	89.1	89.1	89.1	75.6	72.2	72.3	83.0
es	89.7	88.8	85.7	87.5	73.0	70.0	70.1	80.7
de	89.5	85.1	87.9	85.5	72.8	70.7	70.7	80.3
fr	89.2	86.9	85.5	88.8	73.4	70.5	70.3	80.7
zh	76.1	73.3	73.6	73.9	81.5	71.8	74.2	74.9
ko	72.0	70.4	70.3	70.5	71.5	78.3	74.1	72.4
ja	71.3	69.2	70.2	69.8	72.8	73.2	79.3	72.3
avg.	83.0	80.4	80.3	80.7	74.4	72.4	73.0	77.7

Table 6: Results for mBERT-RS-DA on PAWS-X.

	en	es	de	fr	bg	ru	el	th	sw	vi	ar	zh	hi	ur	tr	avg.
en	82.3	70.3	65.8	69.7	60.5	63.1	55.3	44.6	41.1	63.9	57.7	64.6	52.0	49.5	52.3	59.5
es	73.5	74.3	62.9	69.0	60.5	63.7	57.3	44.6	40.6	61.4	57.9	60.8	50.4	47.1	51.6	58.4
de	71.8	65.5	70.8	65.6	59.5	63.3	55.8	44.3	41.0	60.2	56.5	60.1	52.5	49.4	52.0	57.9
fr	73.6	69.0	64.0	73.8	59.5	63.1	55.7	44.1	40.5	62.2	57.3	61.6	51.1	48.5	51.8	58.4
bg	67.8	63.7	60.8	62.5	68.2	64.2	56.0	44.2	39.9	57.4	56.3	57.8	51.2	47.2	50.3	56.5
ru	69.1	65.2	62.6	64.4	62.7	68.7	55.0	44.2	39.9	59.0	56.7	58.6	50.6	46.8	50.0	56.9
el	62.7	61.4	58.0	60.2	57.1	57.7	66.4	44.4	40.5	56.4	55.6	54.0	49.6	46.8	50.7	54.8
th	54.8	52.0	49.9	51.3	49.1	50.4	49.0	53.0	39.4	51.1	49.9	49.3	45.9	44.8	45.4	49.0
sw	54.2	51.2	48.7	50.5	47.2	47.9	47.9	41.8	50.0	48.5	49.1	48.5	45.4	44.4	45.8	48.1
vi	67.4	60.3	57.4	61.2	52.9	57.1	52.9	44.2	39.8	70.3	53.3	62.0	49.2	45.9	47.5	54.8
ar	63.9	60.4	57.0	59.5	54.5	57.1	53.3	43.9	40.4	55.4	64.8	55.2	50.3	48.4	49.9	54.3
zh	67.9	59.9	57.2	59.9	53.4	56.5	50.4	42.7	39.6	60.8	53.5	69.2	48.0	45.7	48.0	54.2
hi	61.4	55.5	55.0	55.3	52.6	54.4	51.9	43.8	40.3	53.8	53.1	53.7	59.7	52.7	49.9	52.9
ur	60.1	54.0	53.9	55.1	48.8	51.5	49.6	41.9	39.7	50.0	52.1	52.3	54.4	57.7	48.2	51.3
tr	61.0	55.1	53.6	55.1	52.0	52.6	50.9	42.4	40.7	52.3	52.0	53.2	49.7	47.3	60.9	51.9
avg.	66.1	61.2	58.5	60.9	55.9	58.1	53.8	44.3	40.9	57.5	55.1	57.4	50.7	48.1	50.3	54.6

Table 7: Results for mBERT on XNLI.

	en	es	de	fr	bg	ru	el	th	sw	vi	ar	zh	hi	ur	tr	avg.
en	82.6	71.2	65.9	70.3	62.0	65.7	57.0	44.1	40.9	64.1	58.9	65.7	52.8	49.2	51.2	60.1
es	74.9	75.0	65.4	71.2	63.0	65.6	59.5	44.5	40.8	62.9	60.1	62.6	52.5	48.7	51.7	59.9
de	72.6	68.0	70.5	67.4	61.7	64.9	58.0	44.4	41.4	61.0	58.8	61.4	53.6	50.4	52.2	59.1
fr	74.7	71.6	65.2	74.1	62.1	64.8	58.4	44.4	40.8	62.7	59.5	62.4	52.7	48.9	51.7	59.6
bg	68.5	66.0	62.9	65.1	68.7	66.8	59.4	45.1	41.1	59.7	59.4	59.7	53.5	49.6	51.8	58.5
ru	69.9	67.1	63.5	65.9	65.0	69.5	58.2	44.7	40.9	60.8	59.2	60.6	53.1	49.5	51.5	58.6
el	63.9	63.3	59.3	62.0	59.8	61.0	67.2	44.7	41.3	57.3	57.7	55.7	51.4	48.2	50.7	56.2
th	56.4	54.1	51.7	53.3	51.9	52.9	51.0	50.5	40.1	52.5	51.8	51.3	48.2	46.3	46.8	50.6
sw	54.1	52.3	49.6	50.9	49.1	49.7	48.6	41.8	48.4	48.7	49.8	48.1	45.4	44.5	47.1	48.6
vi	69.9	65.0	60.5	64.1	58.6	61.8	56.0	45.1	40.4	70.5	57.4	63.4	51.5	48.0	49.1	57.4
ar	64.9	62.8	58.7	61.7	58.7	60.7	56.3	44.7	41.1	58.1	65.4	57.1	52.2	49.6	50.3	56.1
zh	71.1	64.8	60.8	64.1	59.0	61.6	53.9	43.5	40.8	63.0	56.8	69.7	50.9	47.8	49.7	57.2
hi	62.2	58.9	56.7	57.9	56.5	58.3	54.3	44.5	40.8	55.3	55.8	55.3	59.8	54.2	50.4	54.7
ur	61.2	56.7	56.1	57.3	54.4	57.0	53.2	43.7	40.8	54.1	56.3	54.6	56.9	57.9	49.9	54.0
tr	62.4	59.2	57.0	58.6	56.7	57.9	54.2	43.7	40.9	54.8	55.1	54.9	52.2	48.8	59.7	54.4
avg.	67.3	63.7	60.3	62.9	59.1	61.2	56.4	44.6	41.4	59.0	57.5	58.8	52.4	49.4	50.9	56.3

Table 8: Results for mBERT-RS-RP on XNLI.

	en	es	de	fr	bg	ru	el	th	sw	vi	ar	zh	hi	ur	tr	avg.
en	81.0	73.7	69.0	72.8	66.2	68.6	61.2	45.5	41.9	67.1	61.6	68.4	55.9	52.4	56.8	62.8
es	75.2	74.7	66.7	71.4	65.3	67.7	62.0	45.1	41.4	64.1	61.7	64.3	53.9	50.5	55.5	61.3
de	73.4	69.2	71.9	68.7	64.1	67.1	60.4	44.6	41.6	63.2	59.9	63.4	55.4	52.4	55.8	60.7
fr	75.4	72.4	67.2	74.2	64.7	67.5	61.2	45.2	41.2	65.0	61.4	64.8	54.2	51.2	55.0	61.4
bg	70.7	68.4	64.6	66.9	69.9	68.0	61.3	44.9	41.6	61.4	60.4	61.7	54.8	51.0	55.4	60.1
ru	71.6	68.6	65.4	67.4	66.4	70.6	59.7	44.4	40.9	61.9	60.2	62.4	53.9	50.5	54.8	59.9
el	67.0	65.9	61.9	64.8	62.3	63.0	68.0	44.6	41.6	59.3	59.1	58.2	52.7	49.7	53.7	58.1
th	58.0	56.4	53.6	55.4	53.7	54.9	52.4	55.8	40.3	54.7	53.5	53.0	49.5	47.8	48.8	52.5
sw	56.7	55.3	51.7	53.9	52.2	52.6	51.0	43.2	51.1	51.6	52.5	51.5	47.4	46.8	48.9	51.1
vi	71.3	67.2	62.4	66.5	61.0	64.5	58.5	46.1	40.8	71.8	58.9	66.0	53.3	49.9	52.2	59.4
ar	66.7	64.9	60.0	63.5	59.8	61.9	57.7	44.4	41.5	59.3	66.4	58.4	52.5	50.1	53.3	57.4
zh	71.2	66.0	61.9	64.7	59.9	63.5	55.4	43.5	40.3	63.6	57.3	71.5	52.1	49.1	52.8	58.2
hi	64.6	60.9	59.0	59.9	57.9	59.5	56.2	44.5	41.4	56.9	56.6	57.8	62.8	57.4	54.1	56.6
ur	62.8	58.1	57.0	58.3	55.0	57.5	53.5	43.6	41.0	55.1	55.6	56.1	58.9	60.9	52.4	55.0
tr	64.7	61.7	58.5	60.4	58.2	59.0	55.7	44.1	41.9	56.6	56.0	57.7	54.7	51.4	62.9	56.2
avg.	68.7	65.6	62.0	64.6	61.1	63.1	58.3	45.3	41.9	60.8	58.7	61.0	54.1	51.4	54.2	58.0

Table 9: Results for mBERT-RS-DA on XNLI.