Data Augmentation with Adversarial Training for Cross-Lingual NLI

Xin Dong[1], Yaxin Zhu[1], Zuohui Fu[1], Dongkuan Xu[2], Gerard de Melo[3]
[1] Rutgers University
[2] The Pennsylvania State University
[3] Hasso Plattner Institute / University of Potsdam
{xd48, yz956, zuohui.fu}@rutgers.edu
dux19@psu.edu, gdm@demelo.org

Abstract

Due to recent pretrained multilingual representation models, it has become feasible to exploit labeled data from one language to train a cross-lingual model that can then be applied to multiple new languages. In practice, however, we still face the problem of scarce labeled data, leading to subpar results. In this paper, we propose a novel data augmentation strategy for better cross-lingual natural language inference by enriching the data to reflect more diversity in a semantically faithful way. To this end, we propose two methods of training a generative model to induce synthesized examples, and then leverage the resulting data using an adversarial training regimen for more robustness. In a series of detailed experiments, we show that this fruitful combination leads to substantial gains in cross-lingual inference.

1 Introduction

There is a growing need for NLP systems that support low-resource languages, for which task-specific training data may be lacking, while domain-specific parallel corpora may be too scarce to train a reliable machine translation engine. To overcome this, zero-shot cross-lingual systems can be trained on a source language L_{S} and subsequently also be applied to other languages L_{T} despite a complete lack of labelled training data for those target languages. In the past, such systems typically drew on translation dictionaries, lexical knowledge graphs, or parallel corpora, to build a cross-lingual model that exploits simple connections between words and phrases across different languages (de Melo and Siersdorfer, 2007; Fu et al., 2020). Recently, pretrained language model architectures such as BERT (Devlin et al., 2019) have been shown capable of learning joint multilingual representations with self-supervised objectives under a shared vocabulary, simply by combining the input from multiple languages (Devlin et al., 2019; Artetxe and Schwenk, 2019; Conneau and Lample, 2019; Conneau et al., 2019). Such representations greatly facilitate cross-lingual applications. Still, the success of such cross-lingual transfer hinges on how close the involved languages are, with substantial drops observed for some more distant language pairs (Lauscher et al., 2020).

For our study, we focus on natural language inference (NLI), i.e., classifying whether a premise sentence entails, contradicts, or is neutral with regard to a hypothesis sentence (Williams et al., 2017). This is a useful building block for applications involving semantic understanding (Zhu et al., 2018; Reimers and Gurevych, 2019). However, the task is also very challenging, as it not only requires accounting for very subtle differences in meaning but also inferring presuppositions and implications that are not explicitly stated. Due to these intricate subtleties, zero-shot cross-lingual models are often fairly brittle, while obtaining in-language training data is fairly costly.

Data Augmentation. To boost the performance of cross-lingual models, an intuitive thought is to draw on unlabeled data from the target language so as to enable the model to better account for the specifics of that language, rather than just being fine-tuned on the source language. A natural way of exploiting unlabeled data is to consider standard semi-supervised learning methods that leverage a model's own predictions on unlabeled target language inputs (Dong and de Melo, 2019). However, this strategy fails when the predictions are too noisy to serve as reliable training signals. In this paper, we hence explore *data augmentation* to circumvent this problem. The idea, widespread in computer vision and speech recognition, is to generate new training data from existing labeled data. For images, a common approach is to apply

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing*, pages 5158–5167
August 1–6, 2021. ©2021 Association for Computational Linguistics

transformations such as rotation and flipping, as these typically preserve the original label assigned to an image (Krizhevsky et al., 2012). For text, in contrast, data augmentation is more challenging, and straightforward techniques include simple operations on words within the original training sequences, such as synonym replacement, random insertion, random swapping, or random deletion (Wei and Zou, 2019). In practice, however, there are two notable problems. One is that the synthesized data from data augmentation techniques may as well be noisy and unreliable. Second, new examples may diverge from the distribution of the original data.

On NLI, these problems are particularly pronounced, as the very nature of this task is to account for subtle differences between sentences. Modified versions of the original sentences may no longer have the same meaning and entailments. Hence, existing data augmentation techniques often fail to boost the result quality.

Overview and Contributions. In this paper, we propose a novel data augmentation scheme to synthesize controllable and much less noisy data for cross-lingual NLI. This augmentation consists of two parts. One serves to encourage language adaptation by means of reordering source language words based on word alignments to better cope with typological divergency between languages, denoted as Reorder Augmentation (RA). Another seeks to enrich the set of semantic relationships between a *premise* and pertinent *hypotheses*, denoted as Semantic Augmentation (SA). Both are achieved by learning corresponding sequence-to-sequence (Seq2Seq) models.

The resulting samples along with their new labels serve as an enriched training set for the final cross-lingual training. During this phase, we invoke a special adversarial training regimen that enables the model to better learn from such automatically induced training samples and transfer more information to the target languages while better bridging the gap between typologically distinct languages. Our empirical study demonstrates the necessity of incorporating adversarial training into training with synthetic samples and the superiority of our new augmentation method on cross-lingual Natural Language Inference (Conneau et al., 2018). Remarkably, our cross-lingual approach even outperforms in-language supervised learning.

2 Method

Our proposed method consists of two steps. The first involves inducing training examples with two data augmentation models. Next, a task-specific classifier is trained on both the original and the newly generated training instances, with adversarial perturbation for improved robustness and generalization.

2.1 Data Augmentation Model

2.1.1 Reorder Augmentation

Reorder augmentation is based on the intuition of making a model more robust with respect to differences in word order typology. If our training examples consist entirely of instances from a language L_{S} with a fairly strict subject–verb–object (SVO) word order such as English, the model will be less well equipped to pay attention to subtle semantic differences between sentences from a target language L_{T} obeying subject–object–verb (SOV) order. To alleviate this problem, we can rely on auxiliary data to diversify the training data. For this, we obtain word alignments for unannotated bilingual parallel sentence pairs covering L_{S} and an auxiliary language L_{A} that need not be the same as L_{T}. We then reorder all source sentences to match the word order of L_{A} based on the alignments, and train a model to apply such reordering on the NLI training instances.

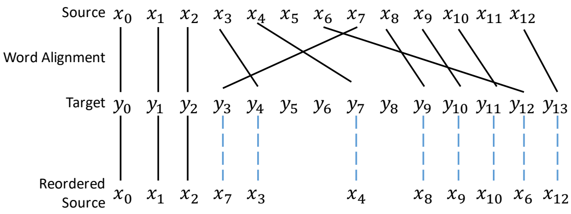
Source x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} x_{11} x_{12}
Word Alignment
Target y_0 y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8 y_9 y_{10} y_{11} y_{12} y_{13}
Reordered Source x_0 x_1 x_2 x_7 x_3 x_4 x_8 x_9 x_{10} x_6 x_{12}

Figure 1: Illustration of using a word-aligned parallel corpus for reordering a source language text.

Formally, suppose we have obtained l unlabelled parallel sentences in the source language L_{S} and in the auxiliary language L_{A}, $\mathcal{C} = \{(\langle s_i, a_i \rangle \mid i = 1, ..., l\}$, where $\langle s, a \rangle$ is a source–auxiliary language sentence pair. Based on a word alignment model, in our case *FastAlign* (Dyer et al., 2013), which uses Expectation Maximization to compute the lexical translation probabilities, we obtain a word pair table for each sentence pair $\langle s, t \rangle$, denoted as $A(s, a) = \{(i_1, j_1), ..., (i_m, j_m)\}$.

Following the word order of L_{A}, we then reorder the source sequence s by consulting the table

$A(s,t)$, yielding the new sentence pair $\langle s, \bar{s}\rangle$. Next, we consider a pretrained Seq2Seq model, denoted as $r(\cdot;\theta)$. The model is assumed to have been pretrained with an encoder and a decoder in the source language, and we fine-tune this generative model by training on the new parallel corpus $\bar{\mathcal{C}} = \{(\langle s_i, \bar{s}_i\rangle \mid i = 1, ..., l\}$. This generative Seq2Seq model can then reorder the sequences in the labeled training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, ..., n\}$, where n is the number of labeled instances, each $\mathbf{x}_i$ consists of a sequence pair $\langle s_1, s_2\rangle$, and each $y_i \in \mathcal{Y}$ is the corresponding ground truth label describing their relationship.

2.1.2 Semantic Augmentation

Our second augmentation strategy involves training a controllable model that, given a sentence and a label describing the desired relationship, seeks to emit a second sentence that stands in said relationship to the input sentence. Thus, given an existing training sentence pair, we can consider different variations of one sentence in the pair and invoke the model to generate a suitable second sentence. However, such automatically induced samples from SA are inordinately noisy, precluding their immediate use as training data, so we exploit a large pretrained *Teacher model* trained on available source language samples to rectify the labels of these synthetic samples with appropriate strategies.

Generation. As we wish to be able to control the label of a generated example, the requested label is prepended to the input as a (textual) prefix before it is fed into a Seq2Seq model. We adopt the ground-truth label of each example as the respective prefix, resulting in a new input sequence $(y_i : s_1)$ coupled with s_2 as the desired output forming a training pair for the generation model.

Given the resulting labeled training dataset $\mathcal{D}_{\mathrm{SA}}$, we can fine-tune a pretrained Seq2Seq model, denoted as $g(\cdot;\theta)$. This generative Seq2Seq model can then be invoked for semantic data augmentation to generate new training instances. For each $(\bar{y} : s_1)$ as a labeled input sequence, where $\bar{y} \in \mathcal{Y} \setminus \{y_i\}$, we generate an $\tilde{s}_2$ via the fine-tuned Seq2Seq model, yielding a new training instance $(\langle s_1, \tilde{s}_2\rangle, \bar{y})$.

Label Rectification. The semantic augmentation induces $\tilde{s}_2$ automatically based on s_1 and the requested label $\bar{y}$. However, the obtained $\tilde{s}_2$ may not always genuinely have the desired relationship $\bar{y}$ to s_1. Thus, we treat this data as inherently noisy and propose a rectifying scheme based on a *Teacher* model. We wish for this *Teacher* to be as accurate as possible, so we start off with a large pretrained language model specifically for the source language L_{S}, which we assume obtains a better performance on L_{S} than a pretrained multilingual model. We train the Teacher network $h(\cdot;\theta)$ in K epochs using the set of original labeled data $\mathcal{D}$. This teacher model is then invoked to verify and potentially rectify labels from the automatically induced augmentation data $\mathcal{D}_{\tilde{\mathrm{a}}} = \{(\tilde{\mathbf{x}}_{\mathbf{i}}, y_i) \mid i = 1, ..., m\}$ obtained in the previous step (where m is the number of instances). We assume $(\tilde{y_i}, c) = h(\tilde{\mathbf{x}}_{\mathbf{i}};\theta)$ denotes the predicted label along with the confidence score $c \in [0,1]$ emitted by the classifier, and assume a confidence threshold T has been predetermined. There are several strategies to determine the final labels.

- **Teacher Strategy:** We adopt $\mathcal{D}_{\mathrm{r}} = \{(\tilde{\mathbf{x}}_{\mathbf{i}}, \tilde{y_i}) \mid (\tilde{\mathbf{x}}_{\mathbf{i}}, y_i) \in \mathcal{D}_{\tilde{\mathrm{a}}}, (\tilde{y_i}, c) = h(\tilde{\mathbf{x}}_{\mathbf{i}}), c > T\}$, i.e., when the confidence score is above T, we believe the Teacher model is sufficiently confident to ensure a reliable label, while other instances are discarded.

- **TR Strategy:** An alternative scheme is to instead adopt $\mathcal{D}_{\mathrm{r}} = \{(\tilde{\mathbf{x}}_{\mathbf{i}}, \Phi(y_i, \tilde{y_i}, c)) \mid (\tilde{\mathbf{x}}_{\mathbf{i}}, y_i) \in \mathcal{D}_{\tilde{\mathrm{a}}}, (\tilde{y_i}, c) = h(\tilde{\mathbf{x}}_{\mathbf{i}})\}$, where
$$\Phi(y_i, \tilde{y_i}, c) = \begin{cases} \tilde{y_i} & c > T \\ y_i & \text{otherwise} \end{cases}$$
Here, labels remain unchanged when Teacher predictions match the originally requested labels. In case of an inconsistency, we adopt the Teacher model's label if it is sufficiently confident, and otherwise retain the requested label.

2.2 Adversarial Training

Upon completing the two kinds of data augmentation, we possess synthesized data that is substantially less noisy, denoted as $\mathcal{D}_{\mathrm{r}}$, which can be incorporated into the original training data $\mathcal{D}$ to yield the final augmented training set $\mathcal{D}_{\mathrm{a}} = \mathcal{D} \cup \mathcal{D}_{\mathrm{r}}$. With this, we proceed to train a new model $f(\cdot;\theta)$ for the final cross-lingual sentence pair classification.

As a special training regimen, we adopt adversarial training, which seeks to minimize the maximal loss incurred by label-preserving adversarial perturbations (Szegedy et al., 2014; Goodfellow et al., 2015), thereby promising to make the model more robust. Nonetheless, the gains observed from it in practice have been somewhat limited in both monolingual and cross-lingual settings. We conjecture that this is because it has previously merely been invoked as an additional form of monolingual regularization (Miyato et al., 2017).

In contrast, we hypothesize that adversarial training is particularly productive in a cross-lingual framework when used to exploit augmented data, as it encourages the model to be more robust towards the divergence among similar words and word orders in different languages and to better adapt to the new modestly noisy data. This hypothesis is later confirmed in our experimental results.

Adversarial training is based on the notion of finding optimal parameters θ to make the model robust against any perturbation $\mathbf{r}$ within a norm ball on a continuous multilingual (sub-)word embedding space. Hence, the loss function becomes:

$$\mathcal{L}_{\text{adv}}(\mathbf{x}_i, y_i) = \mathcal{L}(f(\mathbf{x}_i + \mathbf{r}_{\text{adv}}(\mathbf{x}_i, y_i); \theta), y_i) \tag{1}$$

$$\text{where } \mathbf{r}_{\text{adv}}(\mathbf{x}_i, y_i) = \underset{\mathbf{r}, ||\mathbf{r}|| \leq \epsilon}{\operatorname{argmax}} \mathcal{L}(f(\mathbf{x}_i + \mathbf{r}; \tilde{\theta}), y_i)$$

Generally, a closed form for the optimal perturbation $\mathbf{r}_{\text{adv}}(\mathbf{x}_i, y_i)$ cannot be obtained for deep neural networks. Goodfellow et al. (2015) proposed approximating this worst case perturbation by linearizing $f(\mathbf{x}_i; \tilde{\theta})$ around $\mathbf{x}_i$. With a linear approximation and an L_2 norm constraint in Equation 2, the adversarial perturbation is

$$\mathbf{r}_{\text{adv}}(\mathbf{x}_i, y_i) \approx \epsilon \frac{\mathbf{g}(\mathbf{x}_i, y_i)}{||\mathbf{g}(\mathbf{x}_i, y_i)||_2} \tag{2}$$

$$\text{where } \mathbf{g}(\mathbf{x}_i, y_i) = \nabla_{\mathbf{x_i}} \mathcal{L}(f(\mathbf{x}_i; \tilde{\theta}), y_i).$$

However, neural networks are typically not linear even over a relatively small region, so this approximation cannot guarantee to achieve the best optimal point within the bound. Madry et al. (2017) demonstrated that projected gradient descent (PGD) allows us to find a better perturbation $\mathbf{r}_{\text{adv}}(\mathbf{x}_i, y_i)$. In particular, for the norm ball constraint $||\mathbf{r}|| \leq \epsilon$, given a point r_0, $\Pi_{||\mathbf{r}|| \leq \epsilon}$ aims to find a perturbation $\mathbf{r}$ that is closest to r_0 as follows:

$$\Pi_{||\mathbf{r}|| \leq \epsilon}(r_0) = \underset{||\mathbf{r}|| \leq \epsilon}{\operatorname{argmin}} ||\mathbf{r} - r_0|| \tag{3}$$

To find more optimal points, K-step PGD is needed during training, which requires K forward–backward passes through the network. With a linear approximation and an L_2 norm constraint, PGD takes the following step in each iteration:

$$\mathbf{r}_{t+1} = \Pi_{||\mathbf{r}|| \leq \epsilon} \left(\mathbf{r}_t + \alpha \frac{\mathbf{g}(\mathbf{x}_i, y_i, \mathbf{r}_t)}{||\mathbf{g}(\mathbf{x}_i, y_i, \mathbf{r}_t)||_2} \right) \tag{4}$$

$$\text{where } \mathbf{g}(\mathbf{x}_i, y_i, \mathbf{r}_t) = \nabla_{\mathbf{r}_t} \mathcal{L}(f(\mathbf{x}_i + \mathbf{r}_t; \tilde{\theta}), y_i)$$

Here, α is the step size and t is the step index.

3 Experiments and Analysis

3.1 Experimental Setup

Tasks and Datasets. For evaluation, we used XNLI (Conneau et al., 2018), the most prominent cross-lingual Natural Language Inference corpus, which extends the MultiNLI dataset (Williams et al., 2017) to 15 languages. In our experiments, we considered 20k training data, i.e., ~5% of the original training size to study lower-resource settings requiring augmentation. Following previous work, we consider English as the source language in our experiments.

Model Details. To show that our reorder augmentation strategy does not require auxiliary data from a low-resource target language, we only give it access to parallel data for another closely related high-resource language. Specifically, we use the English–German bilingual parallel corpus from JW300 (Agić and Vulić, 2019). Like English, German commonly adopts an SVO word order, but in some instances also mandates SOV and is generally less rigid than English. This allows us to demonstrate the utility of reorder augmentation even in the absence of data from a language similar to the target language. We relied on *FastAlign*[1] to induce 200k training pairs for Seq2Seq fine-tuning on reordering.

As the pre-trained Seq2Seq model, we used Google's T5-base (Raffel et al., 2020), a unified text-to-text Transformer, to generate new training examples. During generation, we set the beam size as 1 and use sampling instead of greedy decoding. For the Teacher model in semantic augmentation, we relied on RoBERTa-Large (Liu et al., 2019), a robustly optimized BERT model, to fine-tune NLI on English. As the multilingual model, we employ XLM-RoBERTa-base (XLM-R) (Conneau et al., 2019), trained on over 100 different languages. For PGD, the step size α, norm constraint size ϵ, and number of steps K are 1.0, 3.0, 3, respectively. All hyperparameter tuning is conducted based on the

[1] https://github.com/clab/fast_align

Table 1: Hyper-parameters for pretrained models.

Parameter	RoBERTa	T5	XLM-R
max. sequence length	128	150	128
training batch size	16	8	32
learning rate	1e-5	3e-4	1e-5
max. grad. norm	-	1.0	-

accuracy on the English validation set. The Teacher strategy for XNLI then is used for the rectification of semantically augmented texts, as inference requires particularly clean data. The threshold T for this is 0.8. An overview of the basic network parameter values is given in Table 1. We rely on early stopping as a termination criterion. For all NLI classification results, we randomly repeat each experiment 5 times and report the averaged accuracy.

3.2 Main Results

Cross-lingual Inference Classification. Table 2 compares our approach against several strong baselines on XNLI. The first part considers in-language supervised learning, where we relied on genuine training data from the target language rather than a cross-lingual setting. These results are merely provided for comparison. The second part considers zero-shot cross-lingual transfer, i.e., the setting we are targeting in this paper: We first used English training data to train the XLM-R model and then applied it to non-English languages without any training data in the target language. We also trained the model with PGD adversarial training to assess how well PGD works without any data augmentation. Next, we evaluate XLM-R when trained on original and augmented examples from several augmentation methods, with and without adversarial training, respectively. The first of these is Easy Augmentation (EA) by Wei and Zou (2019), a state-of-the-art method for data augmentation in NLP. It mixes 4 strategies, namely synonym replacement, random insertion, random swapping, and random deletion, applying each of these to 20% of words in a sentence. Additionally, we consider our proposed RA and SA strategies, as well as combinations of EA or RA with SA.

Compared with vanilla XLM-R without adversarial training, XLM-R with PGD works better across a range of non-English languages, which shows the effectiveness of adversarial training for more robustness in cross-lingual settings. We observe that XLM-R, when trained with EA or RA, outperform the setting without augmentation for English and some non-English languages, though it does not achieve sufficiently stronger results in terms of the average accuracy across different languages. This suggests that XLM-R struggles to benefit from the augmented instances from RA for better generalizability. In contrast, when trained with SA, XLM-R performs better than without SA examples for most languages, confirming that our semantic augmentation is beneficial. Remarkably, XLM-R with SA examples even succeeds at outperforming in-language training with an average absolute improvement of about 1.1% in accuracy, suggesting that cross-lingual models trained with automatically generated English examples can be more informative with regard to inference than target language examples.[2] Next, we also observe that the accuracy of XLM-R with additional examples from EA, RA, SA is boosted with PGD. This suggests that adversarial training is particularly useful to boost generalizability and robustness when operating on artificial augmented examples.

Beyond this, our full zero-shot approach further outperforms all baselines across 14 languages, including in-language training. This demonstrates the value of improving generalizability and robustness by adding diverse forms of augmentation in an adversarial training framework that can cope with noisy examples.

3.3 Ablation Studies and Analysis

Comparisons on Different Rectifying Strategies. One key part of our method is the label rectification mechanism. We compare different rectification strategies in Table 3. The results show that the Teacher and TR methods introduced in Section 2.1.2 yield fairly similar results. This confirms the robustness of our approach with regard to the choice of strategy. The same also holds for an additional option, **Agreement**, which retains only those examples on which the prediction from the Teacher agrees with the originally requested label. Finally, for comparison, we evaluated yet another strategy, **Requested**, which always adopts the originally requested labels as chosen for generation. We find that this strategy introduces overly many unreliable labels, so the model is unable to work well. This confirms that rectifying labels with a Teacher model is a crucial ingredient.

Comparisons on Adversarial Perturbations. For assessing the value of PGD for adversarial per-

[2] Note that the in-language training data in XNLI was created using machine translation.

Table 2: Accuracy (in %) on XNLI with augmented examples used for cross-lingual transfer. The number of augmented examples from EA, RA and SA are 80k, 20k, 80k. EA (Wei and Zou, 2019) is Easy Data Augmentation. The best cross-lingual transfer results under XLM-R are given in boldface.

Approach	*en*	*de*	*es*	*zh*	*fr*	*ru*	*ar*	*sw*	*ur*	*bg*	*el*	*th*	*tr*	*vi*	*hi*	**avg**
In-Language Supervised Learning (Translate–Train)																
RoBERTa	88.2															
mBERT	73.3	65.2	69.0	66.5	66.5	64.8	61.7	57.7	56.3	65.8	63.4	49.3	61.5	66.9	59.3	63.1
XLM-R	77.7	70.6	73.0	68.1	72.8	70.6	67.4	61.8	60.5	73.2	71.0	68.9	69.3	70.2	64.9	69.3
Zero-Shot Cross-Lingual Transfer																
XLM-R	77.7	71.7	72.6	69.5	72.7	70.2	67.7	60.7	61.0	72.0	70.2	67.4	69.0	71..0	64.9	69.1
+PGD	78.9	71.8	74.5	70.2	73.5	71.1	67.3	60.7	62.0	72.9	71.3	68.7	69.2	71.3	64.9	69.9
+EA(80k)	77.8	70.3	73.1	69.2	72.9	70.3	67.5	61.6	63.5	72.1	70.1	68.1	68.7	69.5	65.1	69.3
+RA(20k)	78.4	71.0	73.1	67.3	73.0	70.2	67.1	61.5	61.1	71.9	70.3	65.5	67.5	69.5	64.7	68.8
+SA(80k)	79.5	72.0	74.4	69.6	74.1	71.9	67.5	63.6	62.7	73.6	71.9	69.0	69.2	71.0	66.1	70.4
+EA+PGD	77.9	71.9	74.4	71.1	73.5	71.5	68.8	63.3	64.4	74.1	68.3	69.5	68.9	70.4	66.9	70.3
+RA+PGD	78.9	72.5	74.7	71.1	74.5	72.0	68.6	63.1	63.6	73.3	72	69.0	69.9	71.7	65.9	70.7
+SA+PGD	80.4	73.4	75.7	71.8	74.0	73.1	69.3	64.5	63.7	74.5	73.2	70.3	70.2	72.3	66.9	71.5
+EA+SA+PGD	80.0	74.0	76.1	73.0	75.5	73.9	**70.2**	63.7	65.5	75.4	73.3	70.5	71.4	72.9	68.0	72.2
+RA+SA+PGD	**80.8**	**74.5**	**77.3**	**73.6**	**75.8**	**74.9**	70.0	**64.8**	**65.7**	**76.3**	**74.9**	**71.6**	**71.4**	**74.5**	**68.5**	**73.0**

Table 3: Accuracy (in %) on XLNI with different rectifying strategies, training on XLM-R with SA and PGD. T is the threshold. p denotes the percentage of initial augmented examples retained for training.

Approach	p	*en*	*de*	*es*	*zh*	*fr*	*ru*	*ar*	*sw*	*ur*	*bg*	*el*	*th*	*tr*	*vi*	*hi*	avg
Teacher ($T = 0$)	100%	79.7	72.8	75.6	71.7	73.9	73.0	69.3	64.5	63.8	74.0	72.6	69.8	70.0	71.8	66.5	**71.3**
Teacher ($T = 0.8$)	94%	80.4	73.4	75.7	71.8	74.0	73.1	69.3	64.5	63.7	74.5	73.2	70.3	70.2	72.3	66.9	**71.6**
TR ($T = 0.8$)	100%	79.1	72.9	75.3	71.4	74.1	73.1	68.8	64.1	63.6	73.9	73.1	70.4	70.4	72.0	66.6	**71.3**
Agreement	66%	78.7	71.3	74.5	70.8	72.7	71.7	68.7	63.8	62.6	73.0	72.0	69.7	69.4	71.1	65.9	**70.4**
Requested	100%	75.4	67.5	70.1	69.0	68.0	69.2	65.7	61.1	61.6	70.5	68.3	65.9	68.3	70.6	64.1	**67.7**

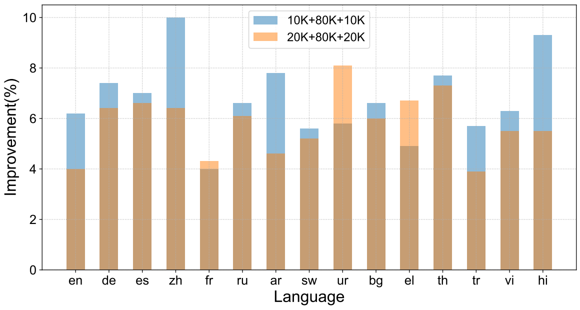

Figure 2: Relative improvements of XLM-R with augmentation and PGD over XLM-R. Blue refers to the improvement on 10k original instances plus 80k SA and 10k RA, while orange refers to the improvement on 20k original instances plus 80k SA and 20k RA, and brown designates the overlap between blue and orange.

turbation, Table 4 compares PGD with the standard Fast Gradient Method (FGM) for adversarial perturbation (Goodfellow et al., 2015) as introduced in Section 2.2. We ran experiments on XNLI with 10k and 20k training data, each augmented with 80k induced semantic examples. We observe that FGM obtains a lower average accuracy than PGD with the same amount of training data, confirming the superiority of PGD in providing better adversarial perturbations than FGM to improve both generalization and robustness.

Effectiveness on Different Training Sizes. Data augmentation is an important approach to deal with scarce labels. The results in Table 4 further show that when fine-tuning T5 using 10k XNLI training instances with 80k semantic and 10k reorder augmented examples, we obtain substantially better results than when using 20k training instances without augmentation. We can also observe the improvement of XLM-R with RA, SA, and adversarial training over vanilla XLM-R on each language as plotted in Figure 2. The relative gains with 10k training data are larger than with 20k training data across a range of languages, which shows that our method is consistently most beneficial when training data is scarce.

Influence of Amount of Augmentation. To assess the role of the amount of data augmentation, we conducted experiments on XNLI with 20k training examples, and evaluated the effect of adding either 20k or 80k augmented examples from EA, RA, SA. The results are given in Table 5. When trained without PGD, one can often benefit from using up to 80k augmented examples. Due to the inherent reordering differences between English and German, there are limits regarding the amount of such data one ought to incorporate. We find that 20k instances from RA can suffice. We observe

Table 4: Accuracy (in %) on XNLI experiments with different amounts of training and augmentation data, and different adversarial training methods.

Approach	*en*	*de*	*es*	*zh*	*fr*	*ru*	*ar*	*sw*	*ur*	*bg*	*el*	*th*	*tr*	*vi*	*hi*	**avg**
XLM-R (10k)	74.5	68.0	70.3	65.5	70.8	68.0	64.2	61.1	60.2	69.9	68.9	65.0	66.9	68.4	61.5	**66.9**
+SA +FGM	77.5	70.9	73.6	68.3	73.1	70.7	67.3	62.2	62.2	72.8	70.5	68.4	67.3	70.2	64.9	**69.3**
+SA +PGD	78.2	71.4	73.7	70.8	73.1	71.2	68.1	62.3	63.2	73.6	71.8	68.9	69.2	71.1	65.6	**70.1**
+RA +SA +PGD	79.1	73.0	75.2	72.3	73.6	72.5	69.2	64.5	63.7	74.5	72.3	70.0	70.7	72.7	67.2	**71.4**
Improvement(%)	6.2	7.4	7.0	10.0	4.0	6.6	7.8	5.6	5.8	6.6	4.9	7.7	5.7	6.3	9.3	**6.7**
XLM-R (20k)	77.7	70.0	72.5	69.2	72.7	70.6	66.9	61.6	60.8	72.0	70.2	66.7	68.7	70.6	64.9	**69.0**
+SA +FGM	79.3	72.4	74.7	70.6	73.7	71.8	67.6	63.5	63.0	72.9	71.9	68.3	69.3	71.6	66.6	**70.5**
+SA +PGD	80.4	73.4	75.7	71.8	74.0	73.1	69.3	64.5	63.7	74.5	73.2	70.3	70.2	72.3	66.9	**71.6**
+RA +SA +PGD	80.8	74.5	77.3	73.6	75.8	74.9	70.0	64.8	65.7	76.3	74.9	71.6	71.4	74.5	68.5	**73.0**
Improvement(%)	4.0	6.4	6.6	6.4	4.3	6.1	4.6	5.2	8.1	6.0	6.7	7.3	3.9	5.5	5.5	**5.8**

Table 5: Accuracy (in %) on XNLI experiments trained using 20k vs. 80k augmentation data from EA, RA, SA, with and without PGD.

Approach	*en*	*de*	*es*	*zh*	*fr*	*ru*	*ar*	*sw*	*ur*	*bg*	*el*	*th*	*tr*	*vi*	*hi*	**avg**
XLM-R (20k)	77.7	71.7	72.6	69.5	72.7	70.2	67.7	60.7	61.0	72.0	70.2	67.4	69.0	71..0	64.9	69.1
+EA (20k)	77.4	69.1	71.9	67.5	71.6	69.3	65.5	61.0	61.5	71.1	69.2	67.1	67.1	68.8	63.9	68.1
+EA (80k)	77.8	70.3	73.1	69.2	72.9	70.3	67.5	61.6	63.5	72.1	70.1	68.1	68.7	69.5	65.1	**69.3**
+RA (20k)	78.4	71.0	73.1	67.3	73.0	70.2	67.1	61.5	61.1	71.9	70.3	65.5	67.5	69.5	64.7	**68.8**
+RA (80k)	77.5	70.8	73.3	68.1	72.2	70.3	66.8	60.7	60.3	72.5	70.5	66.0	67.6	69.3	63.3	68.6
+SA (20k)	78.2	70.6	72.8	67.3	72.6	70.3	66.5	61.4	60.4	71.8	69.6	66.9	67.6	69.5	64.0	68.6
+SA (80k)	79.5	72.0	74.4	69.6	74.1	71.9	67.5	63.6	62.7	73.6	71.9	69.0	69.2	71.0	66.1	**70.4**
+PGD	78.9	71.8	74.5	70.2	73.5	71.1	67.3	60.7	62.0	72.9	71.3	68.7	69.2	71.3	64.9	69.9
+EA +PGD (20k)	77.6	70.9	73.9	69.8	73.0	71.1	67.1	62.4	63.8	73.0	71.3	68.9	69.1	71.2	65.8	69.9
+EA +PGD (80k)	77.9	71.9	74.4	71.1	73.5	71.5	68.8	63.3	64.4	74.1	68.3	69.5	68.9	70.4	66.9	**70.3**
+RA +PGD (20k)	78.9	72.5	74.7	71.1	74.5	72.0	68.6	63.1	63.6	73.3	72	69.0	69.9	71.7	65.9	**70.7**
+RA +PGD (80k)	78.4	71.9	74.9	71.0	73.7	71.9	68.7	62.6	64.0	73.4	72.1	68.9	69.9	71.9	66.4	70.4
+SA +PGD (20k)	79.3	73.3	74.0	69.4	73.3	71.0	67.6	62.7	62.4	73.7	71.7	68.3	69.28	71.1	65.6	70.2
+SA +PGD (80k)	80.4	73.4	75.7	71.8	74.0	73.1	69.3	64.5	63.7	74.5	73.2	70.3	70.2	72.3	66.9	**71.5**

that EA with PGD requires up to 80k augmented instances, i.e., 3 times the size of the original training data, to outperform XLM-R with PGD, whereas only 20k augmented examples suffice for RA with PGD to beat XLM-R with PGD.

Case Studies. To better illustrate the principles of our data augmentation technique, we provide several examples. Table 6 shows two examples of the three data augmentation processes on XNLI. For the first example, the original label is **contradiction**, so **entailment** and **neutral** serve as requested labels to generate new training text. Next, our Teacher model attempts to rectify these labels. Although our generative model treats *Vrenna and I fought him in a fight, but he had just gotten us* as neutral to S_1 (*Vrenna and I both fought him and he nearly took us*), the Teacher model changes the label to **entailment**. For the second example, both the generative and Teacher model are unable to conclude that *The rice ripens in the summer* is contradictory with the premise. From the two EA outputs, we can observe *him* is randomly deleted in Example (1) and *the* and *rice* is swapped in Example (2), which loses some information, whereas RA Seq2Seq generated examples maintain all crucial information despite the reordering.

4 Related Work

Data Augmentation. Data augmentation is a promising technique, especially when dealing with scarce data, imbalanced data, or semi-supervised learning problems. Back-translation (Sennrich et al., 2015) has been considered as a technique to obtain alternative examples preserving the original semantics, by translating an existing example in language L_A into another language L_B and then translating it back into L_A to obtain an augmented example. Yu et al. (2018) and Xie et al. (2020) applied it to question answering and semi-supervised monolingual training scenarios. However, this requires high-quality translation engines that often do not exist in the settings in which one wishes to apply cross-lingual systems.

Wei and Zou (2019) instead combined synonym replacement, random insertion, random swapping, and random deletion in a method named EDA. Since insertion and deletion may affect the semantics of the utterance, some studies opt to control

Table 6: Examples of XNLI data augmentation. V: Version (O: Original). RL: Requested Label. L: Final (possibly rectified) label.

	V	RL	L	Text
(1)	O	–	**contradiction**	$\mathbf{S_1}$: Vrenna and I both fought him and he nearly took us. $\mathbf{S_2}$: Neither Vrenna nor myself have ever fought him.
	EA	–	**contradiction**	$\mathbf{S_1}$: Vrenna and I both fought him and took nearly he us. $\mathbf{S_2}$: Neither Vrenna nor myself have ever fought.
	RA	–	**contradiction**	$\mathbf{S_1}$: Vrenna and I both him fought and he us nearly took. $\mathbf{S_2}$: Neither me nor Vrenna have him ever fought.
	SA	**entailment**	**entailment**	$\mathbf{S_2}$: It was the guy that nearly took the couple of us.
	SA	**neutral**	**entailment**	$\mathbf{S_2}$: Vrenna and I fought him in a fight, but he had just gotten us.
(2)	O	–	**contradiction**	$\mathbf{S_1}$: In summer the rice forms a green velvety blanket, then turns golden in autumn when it ripens and is harvested. $\mathbf{S_2}$: The rice is golden and harvestable in the summer, but turns green in autumn.
	EA	–	**contradiction**	$\mathbf{S_1}$: Harvested summer the rice forms a green velvety blanket then turns golden in autumn when is ripens and it in. $\mathbf{S_2}$: The the is golden and harvestable in rice summer, but turns green in autumn.
	RA	–	**contradiction**	$\mathbf{S_1}$: In summer forms the rice a green velvety blanket, turns then in autumn golden when it ripens and harvested is. $\mathbf{S_2}$: The rice is golden and harvestable in the summer, but turns in autumn green.
	SA	**entailment**	**entailment**	$\mathbf{S_2}$: The rice turns golden in autumn when it ripens.
	SA	**neutral**	**entailment**	$\mathbf{S_2}$: The rice ripens in the summer and then turns golden in the autumn.

the selection of words to be replaced with indicators such as TF-IDF scores (Xie et al., 2020). Fadaee et al. (2017) use contextualized word embeddings to replace the target word. Kobayashi (2018) proposed a bi-directional language-model-based augmentation method, and Wu et al. (2019) further improved its results by switching to BERT. Another major category is text generation based augmentation. Anaby-Tavor et al. (2020) proposed a language model based data augmentation method, shown to improve classifier performance on a variety of English datasets. It relies on GPT-2 (Radford et al., 2018) to generate a single new sequence in each instance.

Our work, in contrast, presents a novel augmentation scheme designed to cope with the special challenges of sentence pair classification, where a Seq2Seq Transformer enables augmentation based on a paired input sentence. Our method also introduces a Teacher model to rectify labels. Apart from this, we expand the idea of language model based augmentation to cross-lingual settings and leverage noisy instances with adversarial training.

Adversarial Training. Many approaches for improving the robustness of a machine learning system against adversarial perturbations (Szegedy et al., 2014) have been advanced. Goodfellow et al. (2015) proposed a fast gradient method based on linear perturbation of non-linear models. Later, Madry et al. (2017) presented PGD-based adversarial training through multiple projected gradient ascent steps to adversarially maximize the loss. In NLP, Belinkov and Bisk (2017) exploited structure-invariant word manipulation and robust training on noisy texts for improved robustness. Iyyer et al. (2018) proposed syntactically controlled paraphrase networks with back-translated data and used them to generate adversarial examples. Adversarial training also plays a role in improving a neural model's generalization. For instance, Cheng et al. (2019) used adversarial source examples to improve a translation model. Dong et al. (2020) exploit FGM-based adversarial training in self-learning for improved cross-lingual text classification. In our setting, we count on adversarial training in the word embedding space and show that PGD-based adversarial training remains effective when the adversarial perturbation is applied to noisy augmented examples.

5 Conclusion

While multilingual pretrained model have enabled better cross-lingual learning, we still often encounter data scarcity issues due to the high cost of collecting data, which weakens the generalization ability of the multilingual model.

To address this, this paper proposes a novel data augmentation strategy with label rectification to build synthetic examples, outperforming even models trained with larger amounts of ground-truth data. We show that we can best learn from such noisy instances with adversarial training, which enables the classifier to transfer more information from the source language to other languages and to become more robust. Remarkably, with this, our models trained without any target language training data at all are able to outperform models trained fully on in-language training data. Moreover, the amount of augmented data from our Seq2Seq-based reorder augmentation used in training is much less than that required by the state-of-the-art EDA method in order to achieve comparable performance. Finally, in our series of follow-up experiments comparing different training regimens and variants, one notable finding is that our overall augmented approach can even outperform non-augmented supervision with twice as many ground truth labels. Overall, this suggests our combination of data augmentation with adversarial training as a valuable way of learning substantially more accurate and more robust models without any target-language training data.

Broader Impact

Research on cross-lingual NLP is often motivated by a desire to provide state-of-the-art advances to linguistic communities that have been underserved. Such advances may enable better access to information as well as to products and services. However, there is a risk that such technological advances may not always be desired by the relevant communities and may indeed also cause harm to them (Bird, 2020). Moreover, cross-lingual systems in particular may exhibit biases with regard to the source language used for training and the general cultural assumptions reflected in such data. In light of this, special care needs to be taken to analyze potential outcomes and risks before deploying cross-lingual systems in real-world applications.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *AAAI*, pages 7383–7390.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xin Dong and Gerard de Melo. 2019. A robust self-learning framework for cross-lingual text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6306–6310, Hong Kong, China. Association for Computational Linguistics.

Xin Dong, Yaxin Zhu, Yupeng Zhang, Zuohui Fu, Dongkuan Xu, Sen Yang, and Gerard de Melo. 2020. Leveraging adversarial training in self-learning for cross-lingual text classification. In *Proceedings of* *the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1541–1544, New York, NY, USA. Association for Computing Machinery.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Zuohui Fu, Yikun Xian, Shijie Geng, Yingqiang Ge, Yuting Wang, Xin Dong, Guang Wang, and Gerard de Melo. 2020. ABSent: Cross-lingual sentence representation mapping with bidirectional GANs. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*. AAAI Press.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 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)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. *ICLR*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of EMNLP-IJCNLP 2019*, pages 6382–6388, Hong Kong, China.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Xunjie Zhu, Tingfeng Li, and Gerard de Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637, Melbourne, Australia. Association for Computational Linguistics.