applied sciences

MDPI

Article

Machine Translation in Low-Resource Languages by an Adversarial Neural Network

Mengtao Sun [1,*], Hao Wang [2], Mark Pasquine [3] and Ibrahim A. Hameed [1]

[1] Department of ICT and Natural Sciences, Norwegian University of Science and Technology, 6009 Ålesund, Norway; ibib@ntnu.no
[2] Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway; hawa@ntnu.no
[3] Department of International Business, Norwegian University of Science and Technology, 6009 Ålesund, Norway; mapa@ntnu.no
* Correspondence: mengtao.sun@ntnu.no

Abstract: Existing Sequence-to-Sequence (Seq2Seq) Neural Machine Translation (NMT) shows strong capability with High-Resource Languages (HRLs). However, this approach poses serious challenges when processing Low-Resource Languages (LRLs), because the model expression is limited by the training scale of parallel sentence pairs. This study utilizes adversary and transfer learning techniques to mitigate the lack of sentence pairs in LRL corpora. We propose a new Low resource, Adversarial, Cross-lingual (LAC) model for NMT. In terms of the adversary technique, LAC model consists of a generator and discriminator. The generator is a Seq2Seq model that produces the translations from source to target languages, while the discriminator measures the gap between machine and human translations. In addition, we introduce transfer learning on LAC model to help capture the features in rare resources because some languages share the same subject-verb-object grammatical structure. Rather than using the entire pretrained LAC model, we separately utilize the pretrained generator and discriminator. The pretrained discriminator exhibited better performance in all experiments. Experimental results demonstrate that the LAC model achieves higher Bilingual Evaluation Understudy (BLEU) scores and has good potential to augment LRL translations.

Keywords: machine learning; adversarial machine learning; imbalanced datasets; transfer learning

check for updates

Citation: Sun, M.; Wang, H.; Pasquine, M.; A. Hameed, I. Machine Translation in Low-Resource Languages by an Adversarial Neural Network. *Appl. Sci.* **2021**, *11*, 10860. https://doi.org/10.3390/app112210860

Academic Editor: Valentino Santucci

Received: 6 October 2021
Accepted: 12 November 2021
Published: 17 November 2021

1. Introduction

Traditional Neural Machine Translation (NMT) models directly learn and fit the correspondence between source and target language pairs through deep neural networks. This approach is based on a sequence-to-sequence (Seq2Seq) architecture which is comprised of encoder and decoder networks. At present, the most popular NMT models such as RNNsearch [1] and Transformer [2] have designs based on the Seq2Seq model architecture. RNNsearch has achieved remarkable translative scores due to its ability to supplement a human-like attention mechanism between the encoder and decoder. RNNsearch achieved several state-of-the-art records up to 2018 and is still widely used in machine translation today. In 2017, a novel architecture known as Transformer was introduced and outperformed existing models in different natural language processing tasks. Recently, researchers have developed a new embedding method based on Transformer, i.e., Bidirectional Encoder Representations from Transformers (BERT) [3]. However, the aforementioned approaches require a large amount of parallel bilingual data for training. For It is laborious for Low-Resource Languages (LRL) to build an adequate corpus for training satisfactory models.

Ruder [4] systematically summarized the necessity of working on LRL information processing. In addition to linguistic diversity, models developed for LRLs can generally help strengthen the featurization, cope with overfitting problems, and facilitate useful applications. For this purpose, there has been much research focusing on LRLs. Zoph et al. [5]

analyzed the relevance in translations by exploiting the pretrained model through the transfer encoder and decoder, but the performances of LRLs were unstable when using different High-Resource Language (HRL) models. To cope with the instability, Maimaiti et al. [6] presented a multi-round transfer learning approach, which alleviated the unpredictability of cross-lingual and generative training to some extent. Moreover, Cheng [7] utilized a pivot language to bridge the language pairs and train a joint network of NMT, i.e., A→B, B→C. Ren et al. [8] introduced a triangle architecture where a small language was an intermediate variable in the translation process between rich languages, dividing the translation process into two translation processes. Their models use the rich bilingual pairs in an HRL corpus to improve the performance of LRL translation.

This study presents research on adversarial learning, which achieves a higher performance in image generation [9]. It incorporates rival losses during training and can yield more explicit images. Recently, this has also been applied to NLP tasks. However, no study has investigated how adversarial learning applies to and influences LRL translation. We seek better feature extraction in the small-scale training of sentence pairs to obtain more accurate translations in complex systems. Moreover, we also take advantage of transfer learning in our proposed model to further improve NMT performance.

There are some challenges to consider when attempting to implement this strategy. First, it is problematic to utilize adversary and Seq2Seq together, as the performances of both techniques need to be analyzed and evaluated. Second, it is challenging to improve translation scores in cross-lingual transfer learning [5,6]. Third, it is challenging to develop a new method combining a pretrained model. Therefore, the proposed system should be developed as an end-to-end differentiable model.

This study proposes a novel Low resource, Adversarial, and Cross-lingual Neural Machine Translation (LAC) model for NMT. The proposed model focuses mainly on LRLs and is expected to overcome the limitations of Seq2Seq, leverage the capabilities of multilingual NMT, and produce high-quality translations. To be more specific, the contributions of this study are summarized as follows:

- A novel translation model, LAC, is designed. Compared to Seq2Seq, this model takes advantage of the adversary technique, reduces the required size of the corpus, and significantly enhances the experimental results on LRLs;
- The LAC model is designed to be end-to-end differentiable and transferable. A pretrained discriminator demonstrated a stronger ability for feature extraction and achieved a higher accuracy in terms of Bilingual Evaluation Understudy (BLEU) scores compared to a non-transferred LAC system;
- The effectiveness of the generator and discriminator in the LAC model is investigated. From the exploratory experiments, the results are analyzed in an interpretable manner.

2. Related Work

2.1. Adversarial Neural Networks

Despite wide usage in image generation, adversarial learning was only proposed for NMT in 2018. Wu et al. [10] utilized the adversary technique to strengthen the Seq2Seq-NMT, namely an Adversarial Neural Machine Translation, which outperformed traditional architectures. Cao et al. [11] also pointed out that the adversary technique supplemented the rival losses to enhance the feature selection from a sequence. The text limitation is that token samples are discrete and undifferentiable, making it inoperable to backpropagate the errors from the discriminator D to the generator G. As a result, G parameters cannot be updated. Recent studies focused on solving the undifferentiability problem by using a lingual adversary technique to address this problem. SeqGAN [12] focused on the differentiation problem using a policy gradient algorithm. Inspired by reinforcement learning, SeqGAN bypasses the generator differentiation problem by directly performing a gradient policy update. A decisional error gradient (instead of an error gradient) was conveyed to train the generator G. Wu et al. [10] used the same strategy to address the gradient problem in a generator. Their model successfully applied adversarial learning to an NMT and achieved better translation scores.

Nevertheless, with reinforcement learning, tuning the parameters requires many experiments in different language models. Lee et al. [13] introduced alternative methods to make the input of D continuous from discrete samplings, e.g., using the hidden states of a generator before activation [14] or substituting the activation function of a generator such as Gumbel-softmax [15]. In this way, the output of G will be the tokens' distributions rather than the tokens' samplings. Press et al. [16] successfully adopted this approach in adversarial text generation systems, which share some similarities with NMT systems. In this work, we use the method mentioned in [14,16], using the hidden states of a generator before activation. An A-NMT uses a pre-trained NMT model as the generator in the most primitive state. However, warm starting seems to reduce generalization in deep neural networks [17]. In addition, it cannot be well adopted in transfer learning of LRL corpora. In the proposed LAC model, the discriminator and generator are designed to facilitate training from scratch. For other related adversarial models, Yi et al. [18] proposed adversarial transfer learning to alleviate the low resource conditions of an acoustic model. Dai et al. [19] put forward a novel metric-based GAN, which used the distance-criteria to distinguish between real and fake samples. Dong et al. [20] presented a semi-supervised adversarial training process for cross-lingual text classification, where the labeled data from one language could be applied to a completely different language classification. We also refer to various solutions for imbalance datasets. Alam [21] proposed a new model specified for imbalanced datasets of credit card default prediction. Khushi utilize the testing results of 20+ class imbalance models with three types of classifiers to detect the best imbalance techniques for medical datasets [22]. Some works explore the risk factors in machine learning models that influence the class identification in an imbalanced dataset [23–25].

2.2. Low Resource Languages Machine Translation

Existing methods of low resource languages machine translation are based on lingual features and transfer learning. For lingual features, Li et al. [26] utilized subword segmentation in Tibetan neural machine translation. The structure of Tibetan words consists of two levels. First, Tibetan words consist of a sequence of syllables, and then a syllable consists of a sequence of characters. According to this special word structure, they proposed two methods for Tibetan to extract the lingual features for machine translation. Tran et al. [27] proposed a new method for word segmentation in Vietnam-Chinese machine translation. They improved the word tokens for isolated Chinese and Vietnamese pairs, made the word boundaries of two languages more symmetric, and achieved 1-1 alignments. As a result, the performance improved by using the embeddings of new word tokens. Choi et al. [28] pointed out that Korean and Japanese share the same grammatical structure for transfer learning. They built an unsupervised machine translation system based on the similarity of the two languages. Nguyen et al. [29] performed Zero-shot reading comprehension by cross-lingual transfer learning. They analyzed the influences of grammatical structure on the model performance and concluded that similar grammatical sentences could improve the effectiveness in cross-lingual transfer learning.

3. Adversarial Model

3.1. GAN

The seminal paper on adversarial training by Goodfellow et al. proposed a Generative Adversarial Network (GAN) in 2014 [9]. The new adversarial model first produces an over expected explicit image without human intervention. Here, we briefly review the three types of GANs originally proposed for adversarial training.

3.1.1. Basic GAN

We denote the randomly initialized Gaussian distribution as Yz, real distribution as Yr, and model distribution as Yg. The goal is to learn the mapping from Yz to Yg and make the distance between Yr and Yg as close as possible, i.e., $x \in Yz$ with distribution $x \sim p_{Yz}(x)$ will be mapped into the domain $\hat{x} \in Yg$ with distribution $\hat{x} \sim p_{Yg}(\hat{x})$, $\hat{x} = G(x)$. The objective function is expressed as:

$$\min_{\mathrm{G}} \max_{\mathrm{D}} \mathcal{L}(\mathrm{G},\mathrm{D}) = \underbrace{\mathbb{E}_{\mathrm{x}\sim\mathrm{p}_{\mathrm{Yr}}}[\log \mathrm{D}(\mathrm{x})]}_{\mathrm{Lr}} + \underbrace{\mathbb{E}_{\mathrm{G(x)}\sim\mathrm{p}_{\mathrm{Yg}}}[1 - \log \mathrm{D}(\mathrm{G}(\mathrm{x}))]}_{\mathrm{Lg}} \quad (1)$$

The inputs of D are two types of data, $\{x\}$ and $\{\hat{x}\}$, in turn. The inputs of G are $\{x\}$. Here, D determines the gradients of G. In the most common training, we maximize D in k times, minimize G one time every epoch, and $k = 10$ is the default. Lr and Lg are marked in Figure 1a.

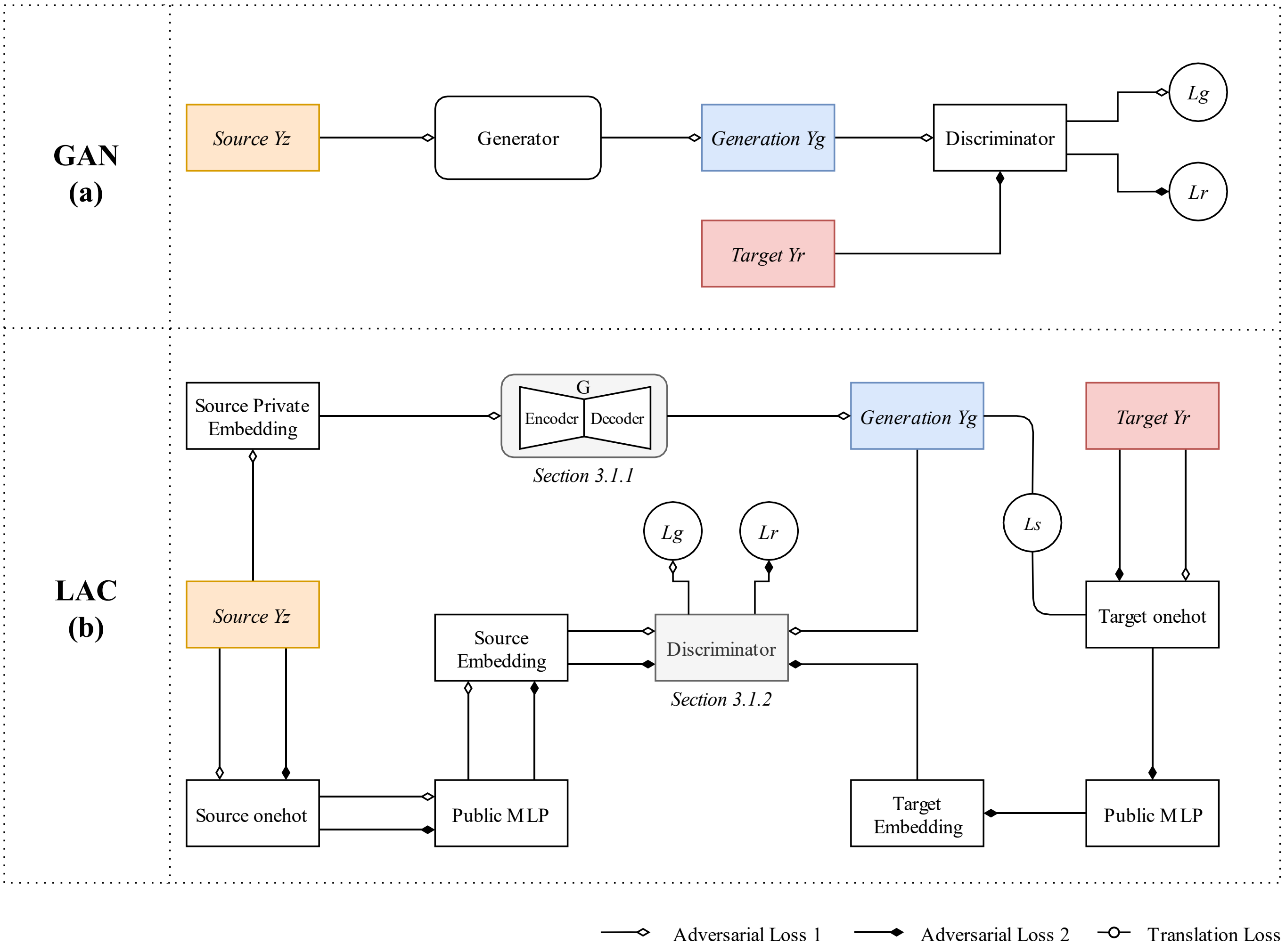


Figure 1. Comparison of the GAN and LAC models. (**a**): GAN: for image generation, the Source, Generation, and Target are randomly initialized noises, generated images, and real images, respectively. (**b**): LAC: the Source, Generation, and Target are the source language, generated translation, and human translation, respectively. Lg, Lr are the adversarial losses, Ls is the translation loss. Batches run along the White (♢) and Black (♦) routes in turn.

GANs have successfully generated images, yielding realistic images that can even fool the human eye. Nevertheless, this type of structure depends heavily on data distributions. It is not stable and often difficult to train without distribution overlaps between generated and real images. Arjovsky et al. [30] proposed the Wasserstein GAN (WGAN) to address these challenges.

3.1.2. WGAN

The loss functions in a GAN are approximated to calculate the Jensen–Shannon (JS) divergence of two distributions. This can easily become locally saturated, leading to the problem of gradient vanishing. Therefore, Arjovsky et al. [30] proposed the Wasserstein distance, substituting the JS divergence with continuity and differentiability. The objective function of a WGAN is expressed as:

$$\min_{\mathrm{G}} \max_{\mathrm{D}\in|\mathrm{f(D)}|_{\mathrm{L}}\leq 1} \mathcal{L}(\mathrm{G},\mathrm{D}) = \underbrace{\mathbb{E}_{\mathrm{x}\sim \mathrm{p}_{\mathrm{Y_r}}}[\mathrm{D(x)}]}_{\mathrm{Lr}} - \underbrace{\mathbb{E}_{\mathrm{G(x)}\sim \mathrm{p}_{\mathrm{Y_g}}}[\mathrm{D(G(x))}]}_{\mathrm{Lg}} \tag{2}$$

where $|f|_L \leq 1$ is a 1-Lipschitz constraint. In a WGAN, the 1-lipschitz constraint is implemented by clipping a compact space $[-c,\ c]$ on the parameters of the discriminator.

In a WGAN, the optimization of $\max_{\mathrm{D}\in|\mathrm{f(D)}|_{\mathrm{L}}\leq 1} \mathcal{L}(\mathrm{G},\mathrm{D})$ is equal to the Wasserstein distance of $(G(x), x)$. In other words, it uses a neural network to approach the Wasserstein distance.

Formally:

$$\text{Wasserstain distance} = \max_{\mathrm{D}\in|\mathrm{f(D)}|_{\mathrm{L}}\leq 1} \mathcal{L}(\mathrm{G},\mathrm{D})$$

i.e., $\max_{\mathrm{D}\in|\mathrm{f(D)}|_{\mathrm{L}}\leq 1} \mathcal{L}(\mathrm{G},\mathrm{D})$ measures the difference between $x \sim p_{Yr}$ and $G(x) \sim p_{Yg}$.

3.1.3. WGAN-GP

Weight clipping is purely used to meet the 1-Lipschitz condition. In later training, most of the WGAN weights normally become plus or minus c, which is not satisfactory in some cases. Gulrajani et al. introduced an improved WGAN with a gradient penalty (WGAN-GP) instead of weight clipping [31]. The WGAN-GP penalizes the gradient norm of the discriminator by using the following objective function:

$$\min_{\mathrm{G}} \max_{\mathrm{D}} \mathcal{L}(\mathrm{G},\mathrm{D}) = \underbrace{\mathbb{E}_{\mathrm{x}\sim \mathrm{p}_{\mathrm{Y_r}}}[\mathrm{D(x)}]}_{\mathrm{Lr}} - \underbrace{\mathbb{E}_{\mathrm{G(x)}\sim \mathrm{p}_{\mathrm{Y_g}}}[\mathrm{D(G(x))}]}_{\mathrm{Lg}} + \underbrace{\lambda\, \mathbb{E}_{\tilde{\mathrm{x}}\sim \mathrm{p}_{\tilde{\mathrm{x}}}}\left[\left(\|\nabla_{\tilde{\mathrm{x}}} \mathrm{D}(\tilde{\mathrm{x}})\|_2 - 1\right)^2\right]}_{\text{Gradient Penalty}} \tag{3}$$

where λ is the penalty coefficient. $\mathrm{p}_{\tilde{\mathrm{x}}}$ is the sampling distribution that uniformly samples along straight lines between pairs of points sampled from the data distribution $\mathrm{p}_{\mathrm{Y_r}}$ and generator distribution $\mathrm{p}_{\mathrm{Y_g}}$. This method performs better than the standard WGAN and achieves stable training on various GAN architectures.

3.2. LAC

As depicted in Figure 1a, the entire GAN system is composed of a discriminator D and generator G, which play minimax games with each other. Two adversarial losses are used to optimize the parameters of G and D in turn. G yields fake samples to confuse the discriminator D and adjusts its parameters according to the recognition in terms of D. In contrast, the goal of the discriminator D is to identify the fake samples generated by G as accurately as possible and adjust its parameters accordingly. Adversarial training and GAN are different concepts. A GAN is used for unsupervised learning, which can generate explicit images without human intervention. Our proposed LAC model is classified as supervised learning. We incorporated the rival losses of a GAN for machine translation because they were helpful for LRL translation.

The LAC model comprises a generator G and a discriminator D, as shown in Figure 1b. The source language and human translations are embedded by a public Multi-Layer Perceptron (MLP). An MLP is a class of feed-forward neural networks. It can be comprised of different layers, and its purpose is to map the one-hot representation of a token into context embedding, which aligns with the work done by Mikolov et al. [32]. Here, we utilize a 1-layer feed-forward neural network for simplicity. The public feed-forward network is used for the source and target languages. To avoid underrepresenting, we set the hidden units to 5000. We define the distribution of the source language Yz, human translation Yr, and generated translation Yg. The inputs to the discriminator are (Yz, Yg) and (Yz, Yr) in turn, yielding two types of adversarial losses Lg and Lr, respectively, as shown in Figure 1b. The distribution of (Yz, Yg) and (Yz, Yr) is as close as possible, based on WGAN-GP. That is, embedding $u \in Yz$ with distribution $u \sim p_{Yz}(u)$ will be mapped into the domain $\hat{v} \in Yg$ with distribution $\hat{v} \sim p_{Yg}(\hat{v})$, $\hat{v} = G(u)$.

The distribution of Yg and Yr is also as close as possible. That is, $\hat{v} \in Yg$ approaches $v \in Yr$ with distribution $v \sim p_{Yr}(v)$ as close as possible. We constrain v, $\hat{v}$, and u in the same dimension. The adversarial losses Lg, Lr are generated from D to measure the Wasserstein distance of $(Yz, Yg) - (Yz, Yr)$. The translation consistency loss Ls measures the distance of $(Yg) - (Yr)$. The objective function is expressed as:

$$\begin{aligned}\min_{\mathrm{G}} \max_{\mathrm{D}} \mathcal{L}(\mathrm{G}, \mathrm{D}) = {} & \underbrace{\mathbb{E}_{\mathrm{u} \sim \mathrm{p}_{\mathrm{Yz}},\, \mathrm{v} \sim \mathrm{p}_{\mathrm{Yr}}(\mathrm{v})}[\mathrm{D}(\mathrm{u}, \mathrm{v})]}_{\mathrm{Lr}} \underbrace{- \mathbb{E}_{\mathrm{u} \sim \mathrm{p}_{\mathrm{Yz}}, \mathrm{G}(\mathrm{u}) \sim \mathrm{p}_{\mathrm{Yg}}(\mathrm{G}(\mathrm{u}))}[\mathrm{D}(\mathrm{u}, \mathrm{G}(\mathrm{u}))]}_{\text{Adversarial loss (Lg)}} \\ & \underbrace{+ \lambda\, \mathbb{E}_{\mathrm{u} \sim \mathrm{p}_{\mathrm{Yz}}, \underset{\mathrm{v}}{\sim} \sim \mathrm{p}_{\tilde{\mathrm{v}}}}\left[\left(\left\| \nabla_{\underset{\mathrm{v}}{\sim}} \mathrm{D}\left(\mathrm{u}, \underset{\mathrm{v}}{\sim}\right) \right\|_2 - 1\right)^2\right]}_{\text{Gradient Penalty}} \underbrace{- \mu\, \mathbb{E}_{\mathrm{v} \sim \mathrm{p}_{\mathrm{Yr}}(\mathrm{v}), \mathrm{G}(\mathrm{u}) \sim \mathrm{p}_{\mathrm{Yg}}(\mathrm{G}(\mathrm{u}))}[\mathrm{v} \log \mathrm{G}(\mathrm{u})]}_{\text{Translation Loss (Ls)}}\end{aligned} \tag{4}$$

where λ is the penalty coefficient. Distribution $\mathrm{p}_{\tilde{\mathrm{v}}}$ is the linear interpolation between distributions Yg and Yr in terms of WGAN-GP. Coefficient μ controls the translation weight. $\mathrm{v} \log \mathrm{G}(\mathrm{u})$ is the cross-entropy of the real and generated translations. We found that cross entropy greatly outperformed Mean Absolute Error and Mean Square Error in machine translation. In a word, Equation (4) consists of adversarial rival loss of WGAN-GP and cross-entropy loss between machine translation and ground truth.

4. LAC Configuration

4.1. Generator

Traditional Seq2Seq NMT models consist of an encoder and decoder, two components of a recurrent neural network. A Gated Recurrent Unit (GRU) [33] is a typical recurrent neural network proposed to solve long-term memory problems and gradients in backpropagation. Compared with Long-Short Term Memory, GRU can greatly improve training efficiency. Therefore, current researchers are more inclined to use GRU.

RNNsearch was proposed in 2014 and is an attention mechanism that makes the decoder conditionally focus on the fraction of hidden states of the encoder. This generally enhances the translation performance. We utilized the RNNsearch as a generator, comprised of a GRU encoder, attention mechanism, and GRU decoder. According to WGAN-GP, we adopted an extra fully-connected layer after data passes through the RNNsearch to produce a logit as output. We also adopted "teacher forcing" to train the LAC model, i.e., using human translation v_{t-1} to calculate generation $\hat{v}_t$.

To recap briefly, given source u and the human translation v_{t-1} in last time step, the generated translation $\hat{\mathrm{v}}_{\mathrm{t}}$ is:

$$\underset{\mathrm{V_t}}{\hat{}} = \mathrm{FC}(\mathrm{h_t}; \mathrm{d}) \tag{5}$$

$$\mathrm{h_t} = \mathrm{RNNsearch}(\mathrm{h_{t-1}}, \mathrm{v_{t-1}}, \mathrm{c_t}) \tag{6}$$

where $\mathrm{h_t}$ is the hidden state from the decoder at time t, and $\mathrm{c_t}$ is the context embedding from the encoder and attention mechanism. d is the number of neurons, which is in accordance with the vocabulary scale in human translation.

From Equation (4), we minimize the generator loss as follows:

$$\mathrm{G\_loss} = -\mathbb{E}_{\mathrm{u} \sim \mathrm{p}_{\mathrm{Yz}}, \mathrm{G}(\mathrm{u}) \sim \mathrm{p}_{\mathrm{Yg}}(\mathrm{G}(\mathrm{u}))}[\mathrm{D}(\mathrm{u}, \mathrm{G}(\mathrm{u}))] - \mu\, \mathbb{E}_{\mathrm{v} \sim \mathrm{p}_{\mathrm{Yr}}(\mathrm{v}), \mathrm{G}(\mathrm{u}) \sim \mathrm{p}_{\mathrm{Yg}}(\mathrm{G}(\mathrm{u}))}[\mathrm{v} \log \mathrm{G}(\mathrm{u})] \tag{7}$$

4.2. Discriminator

Given source u, human translation v, generated translation $\hat{v}$, pairs (u, v) and $(u, \hat{v})$ are separately fed into the discriminator to yield a translative matching degree. Ideally, the output will be greater in (u, v) and smaller in $(u, \hat{v})$. A residual convolutional neural network (CNN) [34] was designed to classify the input pairs based on their hierarchical properties, as shown in Figure 2.

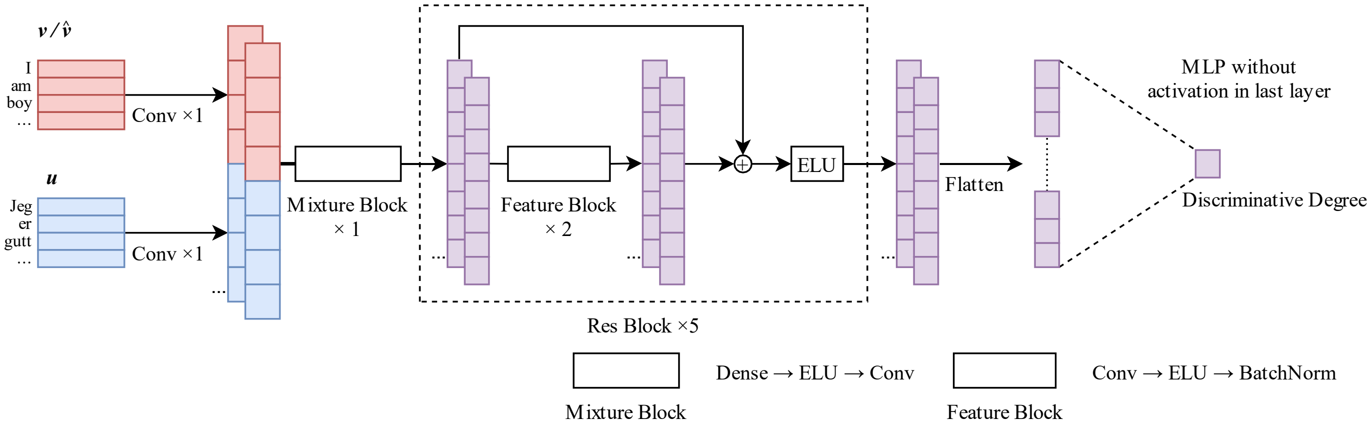


Figure 2. Structure of Discriminator D. Red and blue represent the $(u, \hat{v})$ and (u, v) pairs, respectively, and purple denotes the mixture hidden states.

The discriminator consists of three types of blocks: *Mixture*, *Res*, and *Feature*. For a *Mixture Block*, two types of embeddings in the input pair separately pass a private convolutional layer, and then are concatenated. This block includes dense exponential linear units (ELU) [35] and a convolutional layer in sequence to fuse their embeddings thoroughly. An ELU activation function tends to converge errors to zero faster and produce more accurate results in real tasks than the rectified linear unit (RELU) [36]. For the *Res Block*, the residual connection converges faster under the premise of the same number of layers. After removing a few layers, the performance of the residual network will not be significantly affected [37].

Moreover, Balduzzi et al. [38] pointed out that the residual network could solve the problem of the shattering gradient. Inside the *Res Block*, the *Feature Blocks* contain 1D Convolution, ELU, and a batch normalization layer in line. The hidden state goes into an MLP after being flattened. Here, MLP is a 3-layer feedforward network consisting 256 neurons in the first and second layer with ELU activation, 1 neuron in the third layer without activation. It is noteworthy that the activation function is removed in the last layer of the MLP, based on WGAN-GP. The blocks and layers are depicted in Figure 2.

From Equation (4), we minimize the discriminator loss as follows:

$$\mathrm{D\_loss} = -\mathbb{E}_{\mathrm{u}\sim \mathrm{p}_{\mathrm{Yz}},\, \mathrm{v}\sim \mathrm{p}_{\mathrm{Yr}}(\mathrm{v})}[\mathrm{D}(\mathrm{u},\mathrm{v})] + \mathbb{E}_{\mathrm{u}\sim \mathrm{p}_{\mathrm{Yz}}, \mathrm{G}(\mathrm{u})\sim \mathrm{p}_{\mathrm{Yg}}(\mathrm{G}(\mathrm{u}))}[\mathrm{D}(\mathrm{u},\mathrm{G}(\mathrm{u}))] - \lambda\, \mathbb{E}_{\mathrm{u}\sim \mathrm{p}_{\mathrm{Yz}}, \underset{\mathrm{v}}{\sim}\sim \mathrm{p}_{\underset{\mathrm{v}}{\sim}}}\left[\left(\|\nabla_{\underset{\mathrm{v}}{\sim}} \mathrm{D}\left(\mathrm{u}, \underset{\mathrm{v}}{\sim}\right)\|_2 - 1\right)^2\right] \quad (8)$$

5. Experiments

This section describes the corpora across different source languages translated to English and the baseline methods applied for comparison. We also detail the hyperparameter configuration of the proposed model.

5.1. Dataset

The Tatoeba Dataset comprises short and clean parallel language pairs from 81 languages for the English translation and has been widely used for rare language NMT research [39,40]. LRL is a comparable concept that HRL reflects according to:

(1) The dataset only comprises limited bilingual sentence pairs.
(2) The languages do not have a good pretrained model, or the relative studies are insufficient.

As shown in Table 1, by the number of sentence pairs used in this work, 7 types of translations are selected: tur-eng, aze-eng, ind-eng, tgl-eng, dan-eng, nob-eng and kor-eng. Among them, the following 5 datasets are very low resources: aze-eng, ind-eng, tgl-eng, nob-eng and kor-eng. Here, tur and aze are cognate, and they have similar grammatical structures. dan and nob are cognate, and they have similar grammatical structures. ind, tgl and kor are isolated languages, and they have quite different grammatical structures.

Table 1. Attributions of Translation Corpora.

Language Codes	Full Names	Avg Sentence Length	Train	Val	Test
tur-eng	Turkish-English	8.05	7.0 k	2.0 k	2.0 k
aze-eng	Azerbaijani-English	7.01	2.2 k	0.4 k	0.4 k
ind-eng	Indonesian-English	8.36	2.2 k	0.4 k	0.4 k
tgl-eng	Tagalog-English	8.34	2.2 k	0.4 k	0.4 k
dan-eng	Danish-English	8.94	7.0 k	2.0 k	2.0 k
nob-eng	Norwegian-English	9.14	2.2 k	0.4 k	0.4 k
kor-eng	Korean-English	7.27	2.2 k	0.4 k	0.4 k

To help the source language better align with the target language, the data is processed as follows. Two special tags, "<start>" and "<end>", are inserted at the beginning and end of sentences to signal the start and termination of the translation system, respectively. The words are changed to lowercase and stop words and stop punctuations are removed. All the languages are processed in the same way. We set the max length of a sentence to 9 words, based on an average sentence length. Examples of words before and after preprocessing are shown in Table 2.

Table 2. Examples of Tatoeba Corpus before and after preprocessing.

Language Codes	Before		After	
	Source	Target	Source	Target
tur-eng	Tom şirketin %30'unun sahibi.	Tom owns 30% of the company.	<start> tom şirketin 30 unun sahibi . <end>	<start> tom owns 30 of the company . <end>
aze-eng	Ağzınızı açın!	Open your mouth!	<start> ağzınızı açın ! <end>	<start> open your mouth ! <end>
ind-eng	Aku membayar $200 untuk pajak.	I paid $200 in taxes.	<start> aku membayar 200 untuk pajak . <end>	<start> i paid 200 in taxes . <end>
tgl-eng	"Terima kasih." "Sama-sama."	"Thank you." "You're welcome."	<start> terima kasih. sama sama . <end>	<start> thank you. You re welcome . <end>
dan-eng	Vores lærer sagde at vand koger ved 100 °C.	Our teacher said that water boils at 100 °C.	<start> vores lærer sagde at vand koger ved 100 °C . <end>	<start> our teacher said that water boils at 100 °C . <end>
nob-eng	Du hater virkelig ekskona di, gjør du ikke?	You really do hate your ex-wife, don't you?	<start> du hater virkelig ekskona di, gjør du ikke ? <end>	<start> you really do hate your ex wife, don t you ? <end>
kor-eng	게임은2:30에 시작해.	The game starts 2:30.	<start> 게임은2 30 에 시작해 . <end>	<start> the game starts 2 30 . <end>

5.2. Parameters

We set source embeddings, target embeddings, and source private embeddings as 128 dimensions for the LAC model. The vocabulary list was limited to 5 K words for each source and 4 K words for the target (English). The generator contained 768 units in the GRU layer. The structure of the discriminator shown in Figure 2 has 128 units in each CNN layer in terms of embeddings. The loss here is calculated on a 128 batch size. If the batch size is too small, the randomness will be higher in training. We used the Adam optimizer with a learning rate of 0.001 for the training from scratch in the generator and discriminator. The learning rate was set to 0.0001 when transfer learning.

5.3. Metrics

BLEU scores are often used as the fundamental metric for the evaluation of NMT systems. Ref. [41] analyzed previous criteria and argued that current BLEU methods could not adequately judge translations with a low presence of outliers. Instead, Character *n*-gram F-score (ChrF) [42] was more powerful in efficacy. We used word-level BLEU as our testing metric because it provided some useful confidence conclusions on translation results. We also used F3 values of *n*-gram (ChrF3) to monitor the training progress, where the result was the macro-averaged value of $n = 2$ to $n = 6$.

5.4. Baseline Models

Our baselines include two stages. First, we verified the effectiveness of our proposed LAC model by comparing it with four types of Seq2Seq based neural networks. Second, we compared the LAC model in non-transfer training with a transfer pre-trained Generator, Discriminator, and both. In the deep learning era, traditional machine learning methods are getting weaker at present [43,44]. Therefore, we perform several latest studies on machine translation as baselines. The baseline models are:

RNNsearch: This method is based on word-level sequences. We applied a bidirectional GRU for the encoder, and the attention structure in [1] with another bidirectional GRU for the decoder.

RNNsearch + Unknown (UNK) Replace: As mentioned in [45], using a very large target vocabulary without increasing the training complexity can become difficult. A good solution is replacing the low frequent vocabulary with a special unified UNK token. In low-resource translation, from Turkish to English, this can determine the influence of a low frequent vocabulary on a sentence pair.

BERT: BERT is a pretrained text representative model. More details can be found from the research [2] and [3]. Zhu et al. [46] incorporated BERT into Transformer for NMT. In this study, BERT was directly employed as the encoder to replace a bidirectional GRU (bi-GRU) encoder.

ALBERT: BERT is primarily reliant on large graphic and tensor processing memory. To address this problem, a lite BERT (ALBERT) was proposed as a substitution. With lower complexity, this model shows stronger results in several benchmarks [47].

6. Results

This section discusses the main results of our proposed LAC model for the machine translation task across different LRLs. The proposed model achieved the best results compared with several typical models. We also probe the effectiveness and transferability of the LAC model using explanatory experiments.

6.1. Main Results

6.1.1. Comparison of Baseline Models

A comparison of baseline models was applied to a Turkish-English dataset, as shown in Table 3.

Table 3. Comparison of baseline models.

First Proposed	Details	BLEU
RNNsearch. 2015 [1]	GRU_encoder + Att. + GRU_decoder	33.6
RNNsearch + UNK Replace. 2015 [45]	RNNsearch + UNK Replace	32.8
BERT. 2019. [3] 2020. [46]	BERT_encoder + RNNsearch	34.7
ALBERT. 2020 [47]	ALBERT_encoder + RNNsearch	35.8
LAC-RNNsearch	Adversary (RNNsearch, D)	**37.9**

In our experiment, the traditional RNNsearch model obtained a 33.6 BLEU score in Turkish-English Translation dataset. RNNsearch with UNK Replace cannot help to generalize and obtain better features when lacking sentence pairs, resulting in a decreased BLEU score of 0.8. BERT and its variants show more powerful capabilities and achieved a higher results. Compared to RNNsearch, BERT and ALBERT obtained 1.1 and 2.2 increases in BLEU scores. We incorporated RNNsearch with adversary and conducted the training from scratch. The BLEU score improved by 4.3 with less training data and outperformed the pretrained BERT and ALBERT models.

6.1.2. Comparison of Languages (aze/ind/tgl/kor/nob-eng)

We selected LRLs for our experiments comprised of limited sentence pairs only. The results on the aze/ind/tgl/kor/nob-eng datasets are shown in Table 4.

Table 4. Comparison of low-resource Corpora.

Language Codes	RNNsearch	LAC-RNNsearch
aze-eng	20.4	**20.7**
ind-eng	17.7	**19.3**
tgl-eng	22.0	**22.8**
kor-eng	17.6	**17.7**
nob-eng	14.4	**15.3**

The pretrained embeddings are not available in low resource corpora, so that all the language models were trained from scratch. The RNNsearch was used as the baseline, and the proposed LAC model demonstrated an average enhancement compared with these results. We can see that LAC model has an increment of 0.3 in aze-eng, 1.6 in ind-eng, 0.8 in tgl-eng, 0.1 in kor-eng and 0.9 in nob-eng.

6.2. *Transfer Learning*

We transfer tur-eng as HRL to aze-eng model, and transfer dan-eng as HRL to nob-eng model. Because the two HRLs has the same grammatical structure as their related LRLs. The transferability of the LAC model was tested with a separated transfer generator, separated discriminator, and both the generator and discriminator, as seen in Tables 5 and 6, respectively. The BLEU scores indicate a positive impact when a pre-trained discriminator was used.

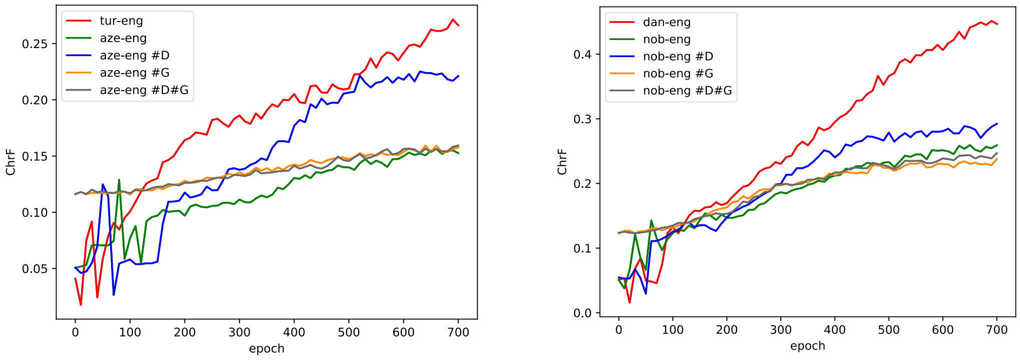
Table 5. Transfer learning of LAC from tur-eng to aze-eng.

aze–eng	BLEU	ChrF3
Non-transfer	20.7	19.4
Transfer G	18.5	16.6
Transfer D	**21.2**	**23.9**
Transfer G and D	18.8	17.1

Table 6. Transfer learning of LAC from dan-eng to nob-eng.

nob–eng	BLEU	ChrF3
None-Transfer	15.3	26.9
Transfer G	15.6	24.7
Transfer D	**15.8**	**29.2**
Transfer G and D	15.5	25.5

The ChrF3 scores from pretrained components in different training steps are shown in Figure 3, which demonstrate that our proposed LAC model can consistently improve translations when the training steps are increased. #D denotes the transfer discriminator, #G denotes the transfer generator, and #D #G denotes both.



(**a**) (**b**)

Figure 3. ChrF3 scores during the training: (**a**) transfer tur-eng to aze-eng, and (**b**) transfer tur-eng to aze-eng. The red line (—) is the reference language. The green (—), blue (—), gold (—), and gray (—) lines represent the non-transfer, transfer Generator, Discriminator, and both, respectively.

Figure 3a shows the change of ChrF3 with increasing steps. Overall, the translation performance of the tur-eng model is better than for the aze-eng model. #G and #D #G demonstrated better performance early compared with training from scratch. They continued to improve slowly but were surpassed by the pretrained discriminator in a later stage. The BLEU score of the discriminator surpassed those of the non-transfer, #G, and #D #G after approximately 300 epochs and then maintained the lead position.

Our hypothesis is also proven in the dan-eng to nob-eng transfer learning experiment. Applying a pretrained discriminator in other languages achieved a higher ChrF3 score than using other pretrained components, as shown in Figure 3b. #G and #D#G had no positive or negative influences on the training progress compared with non-transfer training.

6.3. Case Study

Four translations of different models in Azerbaijan-English and Norwegian-English were generated, provided in Tables 7 and 8. We observed that the proposed LAC model improved and generated better translations, while the RNNsearch remained in a fixed pattern. Because the dataset is very limited, RNNsearch translation tended to be shorter, sentences were not as diverse, and it usually reduplicated common words. In four modes of the LAC model, #D produced the most informative translation. As a result, the generator will receive more useful information and produce more human-like translations by transfer discriminator.

Table 7. Azerbaijan-English.

Epoch	100	200	300	700
Source	Sizin mənə nə edəcəyimi deməyə haqqınız yoxdur.			
Ground Truth	You have no right to tell me what to do.			
RNNsearch	.	i .	i think i think i think i think	i think that what do you know tom
LAC	i m the know i m the know	i m a good to go to go	i don t know you re not your	you have three children are you want to
LAC #G	door tom	door tom	is have you ?	is have you ?
LAC #D #G	door tom	door tom	is have you the .	is have you ?
LAC #D	i m a m a m a m	you re you re you re you re	you have to be your problem to do	you have no right to tell me what
Source	Sən Avstriyanın harasında böyümüsən?			
Ground Truth	Where in Austria did you grow up?			
RNNsearch	.	i .	tom is the dog .	tom is monday .
LAC	i m a know the know the know	i m not a good to go to	you can t want you have a good	where in austria ?
LAC #G	do you house the tom	do you .	do you re ?	do my to tom
LAC #D #G	do you like tom	do you very the tom	do you re ?	do my to tom
LAC #D	i m a m a m a m	you ?	where are you in japan ?	where in austria did you grow up ?
Source	Mən maşında idim.			
Ground Truth	I was in the car.			
RNNsearch	.	i .	i have a good .	i ate the library .
LAC	i m a m a m a m	i m a good .	tom has a good .	i think of the cat .
LAC #G	is briefly .	i happy is .	i m s t .	i m s t .
LAC #D #G	is briefly .	i happy is .	i m s t .	i m s t .
LAC #D	i m a was .	i m a i m a i m	i was in the driver .	i was in the car .
Source	Əminəm ki, Tom sənə nifrət etmir.			
Ground Truth	I'm sure Tom doesn't hate you.			
RNNsearch	.	i .	i m not a good .	i m sure tom will you like it
LAC	i m the know i m the know	i m a good to go to go	i don t know you re not your	i m sure tom doesn t hate you
LAC #G	i will the he	i will the you very the he	i m s the you her you re	i m s the to be you her
LAC #D #G	i will the he	i will the the the the the the	i m s the you her the you	i m s the you the to be
LAC #D	i m a m a m a m	i m a tom s a tom is	i m here .	i m sure tom doesn t hate you

Table 8. Norwegian-English.

Epoch	100 (10)	200 (20)	300 (30)	700 (70)
Source	Jeg betraktet Tom som en venn.			
Ground Truth	I regarded Tom as a friend.			
RNNsearch	i .	i .	i ve been a lot of the truth	i ve been to be a friend .
LAC	i m a lot of a lot of	i m a lot of the tom is	i m sure tom is a friend in	i wonder tom will have a friend .
LAC #G	i if tom to tom	i was into . .	i was tom .	i m and tom a lot it tom
LAC #D #G	i that	i ve tom . .	i was t to to .	i m tom a friend .
LAC #D	i m to i m to i m	i m a lot of the tom s	i tom tom a friend .	i assumed tom was a friend .
Source	Er det noe du ikke forteller oss?			
Ground Truth	Is there something you're not telling us?			
RNNsearch	i .	i .	i ve been to be a lot of	what is it s someone know that you
LAC	i m a lot of a lot of	are you have a lot of this is	is there is you re not his life	is there something you re not telling me
LAC #G	i do about you the .	are s you you the .	are s you t you the .	are s you t you that ?
LAC #D #G	i do people you the .	are s you the .	are s you the .	are s you that you you that you
LAC #D	tom is a lot .	are you have to do you are you	is there s something to do not to	is there something you re not telling me
Source	Jeg skulle ønske det var mer jeg kunne ha gjort.			
Ground Truth	I wish there was more I could've done.			
RNNsearch	i .	i .	i wish i wish i wish i wish	i wish i wish i wish i wish
LAC	i ve	i m not to be a lot of	i wish there s more than i was	i wish there was more than i was
LAC #G	do wish the .	i wish . .	i wish i do i could the .	i wish there do i could the .
LAC #D #G	i wish the .	i wish . .	i wish i had the .	i wish there will more do more more
LAC #D	i m i m i m i m	i m a lot of i ve been	i wish it was more than i was	i wish there were more than i could
Source	Tom skal gjøre det i morgen.			
Ground Truth	Tom will be doing that tomorrow.			
RNNsearch	i .	i .	i m a lot of the lot of	you re supposed to be happy to be
LAC	tom s a lot of a lot of	tom is a lot .	tom will be happy to be the truth	tom will do that .
LAC #G	he t with really	tom your with i that with i that	tom a lot know find	tom will really tomorrow .
LAC #D #G	he t with he t	tom . .	tom a lot know a lot it ,	tom will it tomorrow
LAC #D	tom is the lot .	tom s a lot of the room .	tom is do that .	tom will do that tomorrow .

6.4. Ablation Study

The contribution of different components was observed in the LAC model. Ablation experiments were performed on the tur-eng dataset and the results are displayed in Table 9. When we substituted the encoder-decoder of the generator using RNN, there was a 2.4%

decrease. The performance decreased by 2.9% if the attention mechanism was removed. From these results, we found that the generator played an important role in the LAC model.

Table 9. Ablation Study of LAC in Turkish-English Sentence Pairs.

Model	BLEU
LAC	**37.9**
G_RNN	35.5
G_No Attention	35.0
D_ReLu Nonlinearity	36.5
D_ res block × 1	35.6

The importance of the discriminator was also demonstrated. The activation function caused a 1.4% decline with the RELU replacement. Moreover, the results demonstrated that the LAC model with a single res block layer, i.e., reducing the ability of discriminator, the result has a 2.3% drop in BLEU score.

6.5. Wasserstein Distance

The Wasserstein distance of training progress in each step is shown in Figure 4.

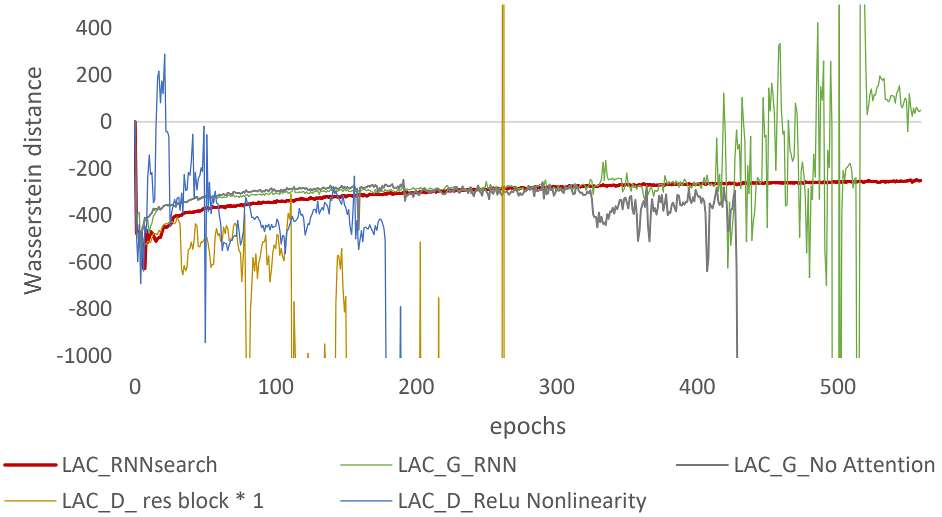


Figure 4. Wasserstein distance during the training on the validation dataset, plotted by every epoch.

The critic ability is reduced in the ablation study against the discriminator (i.e., the gold and blue lines). The Wasserstein distance cannot be accurately measured as the curve is diverging. That is, the discriminator is unable to detect generated and real translation. As a result, the rival loss cannot be used to improve the translation.

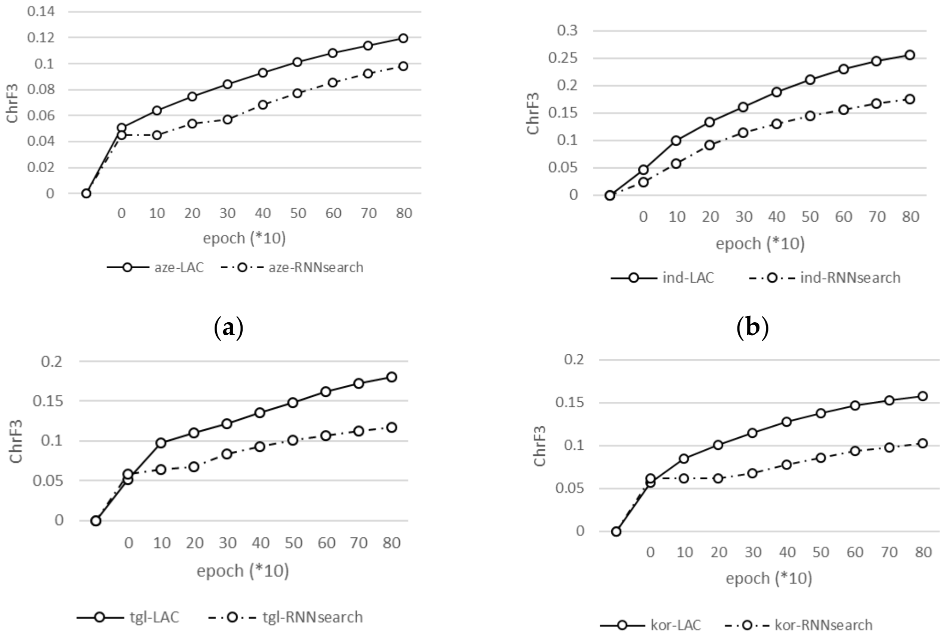
The ablation study reduces the translative ability against the generator (i.e., the green and gray lines). The curve started to diverge after 300 epochs because the generator was well fitted and could produce some translations. That is, the generated and real translations were not separable from the discriminator. Additionally, due to the weak translative ability, the generator could only produce the most common words.

The optimization showed smooth and incremental converging progress in our proposed LAC model (the red line). The LAC model achieved the best BLEU score based on translative and critical abilities, as shown in Sections 6.1 and 6.2.

From Figure 4, the LAC model incorporates the knowledge from adversarial systems and human translation. It got the better translation features and produced the best translation score in low-resource languages machine translations.

6.6. Steps of Message Passing

The ChrF3 curve of the LAC during different step numbers was plotted to demonstrate the influence of epochs during the update process and the performances with and without an adversary. Figure 5 illustrated the ChrF3 scores for four LRLs when the step number was increased.



(**a**) (**b**)

(**c**) (**d**)

Figure 5. ChrF3 training curves for four languages: (**a**) aze-eng, (**b**) ind-eng, (**c**) tgl-eng, and (**d**) kor-eng. LAC-RNNsearch and RNNsearch are denoted by the solid and dotted lines, respectively.

The results indicate that the number of update steps is crucial to the performance of the LAC model, which increased on all four datasets. The ChrF3 of the LAC model not only outperformed RNNsearch in the testing, but it was larger at each step for all four LRLs. The generator learned an extra rival loss from the discriminator, aggregating the global information from the machine and human translations in each step. LAC model can therefore capture more valuable information through the adversary.

7. Conclusions

This study proposed a new machine translation model based on an adversarial mechanism, named LAC. The results of the LAC are significantly stronger than those of the traditional machine translation models without an adversarial mechanism. LAC does not have an over-complex structure, but it shows better evaluations compared with the latest models. Furthermore, the higher performance are widely shown in multiple languages, indicating that the adversary can effectively improve the model capabilities.

Typically, transfer learning is not suitable for machine translation even though the languages have similar grammar. In this experiment, we analyzed the transferability of LAC inter similar languages. First, we used both pretrained discriminator and generator from a relative and higher resource language. Then we used a separate pretrained discriminator and generator. We found that using a separate pretrained discriminator shows better performance. Similarly, in case studies, a separate pretrained discriminator produced more fluent and correct sentences. It manifested that the LAC model has the potential in cross-lingual transfer learning compared with traditional models.

We analyzed the impact of different components in the LAC network by ablation experiments. In conclusion, no matter the ability of discriminator or generator is reduced, the translation results eventually became worse. Furthermore, from the Wasserstein distance curves (i.e., convergence curves) of ablation experiments, we found that reducing the capabilities of the discriminator or generator will eventually make the LAC model non-convergent. It showed that the original LAC model incorporates the useful adversarial features from discriminator and generator. The performance of the LAC model is the result of the interaction of discriminator and generator.

Finally, we tested the translation performance of the model in different iteration steps, and we found that the LAC model was better than the translation system without an adversarial mechanism during iteration. It shows that the adversarial mechanism can improve the model's ability in any step, and the improvement is stable.

In summary, experimental results showed that LAC has good potential in LRL translations. For future works, we will explore how to improve and leverage the discriminator and generator so that the translation performance can be further improved. In addition, we will work on how to reduce the computational costs of adversarial training.

Author Contributions: Methodology, M.S.; software, M.S.; validation, M.S.; formal analysis, M.S.; investigation, M.S.; data curation, M.S.; writing—original draft preparation, M.S.; writing—review and editing, H.W., M.P. and I.A.H.; visualization, M.S.; supervision, H.W., M.P. and I.A.H.; project administration, H.W., M.P. and I.A.H.; funding acquisition, H.W., M.P. and I.A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Norwegian University of Science and Technology.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7 May 2015.
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems Conference, Long Beach, CA, USA, 4 December 2017; pp. 5998–6008.
3. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2 June 2019.
4. Ruder, S.; Vulić, I.; Søgaard, A. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.* **2019**, *65*, 569–631. [CrossRef]
5. Zoph, B.; Yuret, D.; May, J.; Knight, K. Transfer learning for low-resource neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1 November 2016. [CrossRef]
6. Maimaiti, M.; Liu, Y.; Luan, H.; Sun, M. Multi-Round Transfer Learning for Low-Resource NMT Using Multiple High-Resource Languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2019**, *18*, 4. [CrossRef]
7. Yong, C. Joint Training for Neural Machine Translation. Ph.D. Thesis, IIIS Department, Tsinghua University, Beijing, China, 2014.
8. Ren, S.; Chen, W.; Liu, S.; Li, M.; Zhou, M.; Ma, S. Triangular architecture for rare language translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15 July 2018. [CrossRef]
9. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
10. Wu, L.; Xia, Y.; Tian, F.; Zhao, L.; Qin, T.; Lai, J.; Liu, T.Y. Adversarial neural machine translation. In Proceedings of the Asian Conference on Machine Learning (ACML 2018), Beijing, China, 14 November 2018; pp. 534–549.
11. Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Liu, S. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October 2018; pp. 182–192. [CrossRef]
12. Yu, L.; Zhang, W.; Wang, J.; Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4 February 2017. [CrossRef]

13. Lee, H.; Yu, T. ICASSP 2018 tutorial: Generative adversarial network; Its applications to signal processing; Natural language processing. In Proceedings of the ICASSP 2018, Calgary, AB, Canada, 15 April 2018. [CrossRef]
14. Zhang, Y.; Gan, Z.; Fan, K.; Chen, Z.; Henao, R.; Shen, D.; Carin, L. Adversarial feature matching for text generation. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6 August 2017; Volume 70, pp. 4006–4015.
15. Nie, W.; Narodytska, N.; Patel, A. Relgan: Relational generative adversarial networks for text generation. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April 2018.
16. Press, O.; Bar, A.; Bogin, B.; Berant, J.; Wolf, L. Language generation with recurrent generative adversarial networks without pre-training. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6 August 2017.
17. Ash, J.T.; Adams, R.P. On the difficulty of warm-starting neural network training. *arXiv* **2019**, arXiv:1910.08475.
18. Yi, J.; Tao, J.; Wen, Z.; Bai, Y. Language-Adversarial Transfer Learning for Low-Resource Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2019**, *27*, 621–630. [CrossRef]
19. Dai, G.; Xie, J.; Fang, Y. Metric-based generative adversarial network. In Proceedings of the 25th ACM International Conference on Multimedia (MM '17), Mountain View, CA, USA, 23–27 October 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 672–680. [CrossRef]
20. Dong, X.; Zhu, Y.; Zhang, Y.; Fu, Z.; Xu, D.; Yang, S.; de Melo, G. Leveraging adversarial training in self-learning for cross-lingual text classification. In Proceedings of the 43rd International ACM SIGIR Conference on Research, Development in Information Retrieval, Virtual Conference, China, 25–30 July 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1541–1544. [CrossRef]
21. Alam, T.M.; Shaukat, K.; Hameed, I.A.; Luo, S.; Sarwar, M.U.; Shabbir, S.; Li, J.; Khushi, M. An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access* **2020**, *8*, 201173–201198. [CrossRef]
22. Khushi, M.; Shaukat, K.; Alam, T.M.; Hameed, I.A.; Uddin, S.; Luo, S.; Yang, X.; Reyes, M.C. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access* **2021**, *9*, 109960–109975. [CrossRef]
23. Alam, T.M.; Shaukat, K.; Mahboob, H.; Sarwar, M.U.; Iqbal, F.; Nasir, A.; Hameed, I.A.; Luo, S. A Machine Learning Approach for Identification of Malignant Mesothelioma Etiological Factors in an Imbalanced Dataset. *Comput. J.* **2021**, bxab015. [CrossRef]
24. Latif, M.Z.; Shaukat, K.; Luo, S.; Hameed, I.A.; Iqbal, F.; Alam, T.M. Risk factors identification of malignant mesothelioma: A data mining based approach. In Proceedings of the 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), Istanbul, Turkey, 12–13 June 2020; pp. 1–6.
25. Yang, X.; Khushi, M.; Shaukat, K. Biomarker CA125 Feature engineering and class imbalance learning improves ovarian cancer prediction. In Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 16–18 December 2020; pp. 1–6.
26. Li, Y.; Jiang, J.; Yangji, J.; Ma, N. Finding better subwords for Tibetan neural machine translation. In Proceedings of the Transactions on Asian and Low-Resource Language Information Processing, Gold Coast, Australia, 16–18 December 2021; Volume 20, pp. 1–11.
27. Tran, P.; Dinh, D.; Nguyen, L.H. Word re-segmentation in Chinese-Vietnamese machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2016**, *16*, 1–22. [CrossRef]
28. Choi, Y.-S.; Park, Y.-H.; Yun, S.; Kim, S.-H.; Lee, K.-J. Factors Behind the Effectiveness of an Unsupervised Neural Machine Translation System between Korean and Japanese. *Appl. Sci.* **2021**, *11*, 7662. [CrossRef]
29. Nguyen, T.Q.; Chiang, D. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. In Proceedings of the EMNLP, Hong Kong, China, 3 November 2019.
30. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6 August 2017; pp. 214–223.
31. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of Wasserstein GANs. In Proceedings of the Advances in Neural Information Processing Systems Conference, Long Beach, CA, USA, 4 December 2017; pp. 5767–5777.
32. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems Conference, Lake Tahoe, NV, USA, 5 December 2013.
33. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar, 25–29 October 2014; pp. 1724–1734. [CrossRef]
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
35. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by Exponential Linear Units (ELUs). In Proceedings of the International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico, 2 May 2016.
36. Kim, J.; Won, M.; Serra, X.; Liem, C.C. Transfer learning of artist group factors to musical genre classification. In Proceedings of the Web Conference 2018 (WWW '18), Lyon, France, 23 April 2018; pp. 1929–1934. [CrossRef]
37. Veit, A.; Wilber, M.J.; Belongie, S. Residual networks behave like ensembles of relatively shallow networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5 December 2016; pp. 550–558.
38. Balduzzi, D.; Frean, M.; Leary, L.; Lewis, J.P.; Ma, K.W.D.; McWilliams, B. The shattered gradients problem: If resnets are the answer, then what is the question? In Proceedings of the 34th International Conference on Machine Learning (ICML'17), Sydney, Australia, 6 August 2017; Volume 70, pp. 342–350.

39. Fu, Z.; Xian, Y.; Geng, S.; Ge, Y.; Wang, Y.; Dong, X.; Wang, G.; de Melo, G. ABSent: Cross-lingual sentence representation mapping with bidirectional GANs. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020), New York, NY, USA, 7 February 2017. [CrossRef]
40. Tran, C.; Tang, Y.; Li, X.; Gu, J. Cross-lingual retrieval for iterative self-supervised training. In Proceedings of the Advances in Neural Information Processing Systems Conference (NIPS 2020), Virtual Conference, 6 December 2020.
41. Mathur, N.; Baldwin, T.; Cohn, T. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Virtual Conference, 5 July 2020. [CrossRef]
42. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation (EMNLP 2015 Workshop), Lisbon, Portugal, 17 September 2015; pp. 392–395. [CrossRef]
43. Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Xu, M. A Survey on Machine Learning Techniques for Cyber Security in the Last Decade. *IEEE Access* **2020**, *8*, 222310–222354. [CrossRef]
44. Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Chen, S.; Liu, D.; Li, J. Performance Comparison and Current Challenges of Using Machine Learning Techniques in Cybersecurity. *Energies* **2020**, *13*, 2509. [CrossRef]
45. Jean, S.; Cho, K.; Memisevic, R.; Bengio, Y. On using very large target vocabulary for neural machine translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015), Beijing, China, 26 July 2015. [CrossRef]
46. Zhu, J.; Xia, Y.; Wu, L.; He, D.; Qin, T.; Zhou, W.; Li, H.; Liu, T.Y. Incorporating BERT into neural machine translation. In Proceedings of the International Conference on Learning Representations (ICLR 2020), Virtual Conference, 30 April 2020.
47. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A lite BERT for self-supervised learning of language representations. In Proceedings of the International Conference on Learning Representations (ICLR 2020), Virtual Conference, 30 April 2020.