Data Augmentation for Low-Resource Languages NMT Guided by Constrained Sampling

A Preprint

Mieradilijiang Maimaiti, Yang Liu*,[†] Huanbo Luan, Maosong Sun
Institute for Artificial Intelligence, State Key Laboratory of Intelligent Technology and Systems
Beijing National Research Center for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China.
meadljmm15@mails.tsinghua.edu.cn;{liuyang.china,luanhuanbo}@gmail.com;sms@tsinghua.edu.cn

August 3, 2021

Abstract

Data augmentation is a ubiquitous approach for several text generation tasks. Intuitively, in the machine translation (MT) paradigm, especially in low-resource languages (LRLs) scenario, many data augmentation methods have appeared. The most commonly used methods are building pseudo corpus by randomly sampling, omitting, or replacing some words in the text. However, previous approaches hardly guarantee the quality of augmented data. In this work, we try to augment the corpus by introducing a constrained sampling method. Additionally, we also build the evaluation framework to select higher quality data after augmentation. Namely, we use the discriminator submodel to mitigate syntactic and semantic errors to some extent. Experimental results show that our augmentation method consistently outperforms all the previous SOTA methods on both small and large scale corpora in 8 language pairs from 4 corpora by $2.38 \sim 4.18$ BLEU points.

Keywords Artificial Intelligence · Natural Language Processing (NLP) · Low-Resource Languages · Neural Machine Translation (NMT) · Data Augmentation · Constrained Sampling.

1 Introduction

In recent years, the end-to-end framework has been a common architecture in the neural network. Neural machine translation (NMT), which employs neural networks to model the translation process of natural languages with end-to-end manner, has attracted the most attention from the community [1, 2, 3]. Excelling in learning representations and capturing long-distance dependencies by exploiting gating [4, 5] and attention mechanisms [2, 6], NMT has shown significant superiority over conventional statistical machine translation (SMT) [7] in many natural language pairs [8]. Therefore, NMT has recently become an appealing approach for real-world machine translation (MT) systems.

However, NMT demands large amounts of parallel corpora, and therefore it usually learns poorly on low-resource languages (LRLs). Large-scale parallel corpora in high-resource languages (HRLs) are easy to obtain. The high agglutinations of LRLs bring some challenges for training NMT model on LRLs [9]. In machine translation approaches, low-resource languages frequently cause a data scarcity problem [10]. Therefore, we focus on this issue in our proposed model. Data augmentation (DA) is a vital method to improve the accuracy of deep learning approaches by building extra samples. Meanwhile, these augmented data may have lower quality than the original real training data. The model quality is a grave issue of NMT, and it roughly relies on the accessibility of large amounts of parallel data, which is frequently hard to achieve. Besides, it is hard to remarkably outperform SMT since neural networks tend to learn defectively on LRLs.

[*] Yang Liu is the corresponding author.

[†] Yang Liu is also with the Beijing Academy of Artificial Intelligence, Beijing Advanced Innovation Center for Language Resources, Beijing 100084, China.

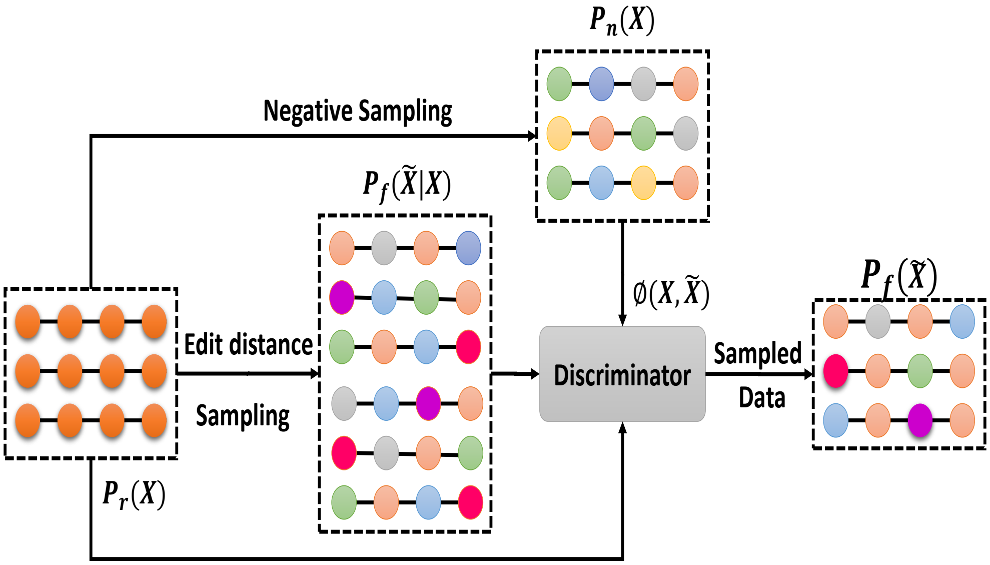

Figure 1: Illustration of constrained sampling for data augmentation. Where $P_r(\mathbf{x})$ represents real data distribution and $P_f(\tilde{\mathbf{x}}|x)$ denotes achieved fake data distribution using edit distance sampling, $P_n(\mathbf{x})$ stands for our negative sample is obtained by negative sampling, $\phi(\mathbf{x}, \tilde{\mathbf{x}})$ is our discriminator sub-model to evaluate the quality of augmented data and is trained on only $P_r(\mathbf{x})$ and $P_n(\mathbf{x})$ rather than also using edit distance sampled data all at once. The $P_f(\tilde{\mathbf{x}})$ refers to our final generated data selected by discriminator $\phi(\mathbf{x}, \tilde{\mathbf{x}})$.

Generally, one of the most challenging tasks is the translation from LRLs (Azerbaijani, Hindi, Uyghur, Uzbek) and morphologically rich languages (Arabic, Turkish) to HRLs. DA has been widely using in several fields. In computer vision, some transformations like resizing, rotating, cropping, and flipping are major data augmentation techniques [11]. Furthermore, similar conversion methods [12] have also been broadly leveraged in natural language processing (NLP) tasks. Yet, due to text characteristics, common random transformation usually brings trivial changes and even makes some semantic or syntactic mistakes. Contextual augmentation [13] can be seen as a novel method, which exploits the language model (LM) to replace some words with another. The drawback of this method is that several unseen samples have high variations and they require too much sampling time. Additionally, since the minor revising of sequences can result in extreme changes in their semantics, it is problematic to seek simple correspondence for some tough NLP tasks like MT. Because of such complexity, the previous literature in DA for NMT is rather insufficient. [14] swaps the words in the sequence randomly. [15] drops some words randomly in a sentence to help NMT training via a learning encoder. [12] shows that data nosing is an effective technique for NMT by replacing the words with a placeholder. [16] also gains improvements in NMT by alternating the words in the target with rare words and revising the corresponding words in the source. Besides, the most recent work [17] proposes a method that leverage prior knowledge by taking advantage of the bi-directional language model to replace word tokens.

In this work, we derive a quite straightforward but effective constrained sampling method for data augmentation in NMT. We take DA as a sampling method and introduce different data augmentation methods for MT in LRLs, and we believe the constrained sampling method which exploits edit distance calculation is more efficient than other approaches that use random sampling words among the original text. Additionally, we also devise the evaluation sub-model to select the higher quality data after generation, and expect such sub-model to ignore sequences with semantic or syntactic errors to some extent. Precisely, as depicted in Figure 1, our method can be decomposed into three steps. Firstly, to train the discriminator sub-model, we need to use both positive and negative samples. Thus we build some negative samples from real data leveraging the negative sampling method (e.g. like word omitting methods). Here we take advantage of the word omitting techniques as a negative sampling method to generate some negative data. Secondly, we train our evaluation sub-model using both real data and generated negative data in the first step. Thirdly, we augment some samples using the edit distance sampling method on real data distribution and ignore the low-quality augmented sequences by the discriminator sub-model. Our contributions are as follows:

- We introduce a novel sampling method for data augmentation in low-resource languages for machine translation task.

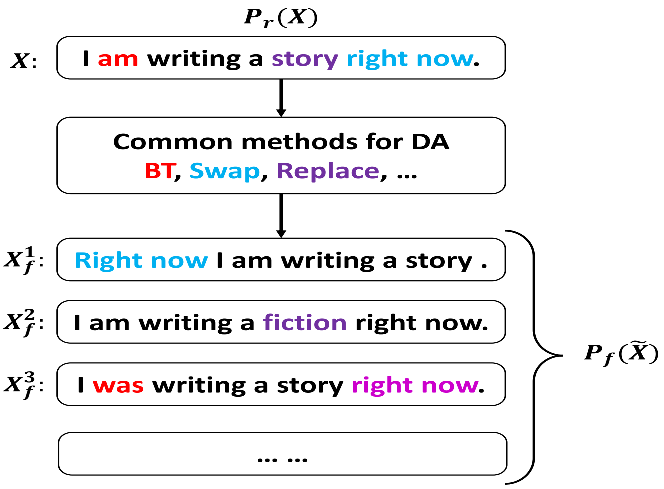
$P_r(X)$
X: I am writing a story right now.
Common methods for DA
BT, Swap, Replace, ...
X_f^1: Right now I am writing a story .
X_f^2: I am writing a fiction right now.
X_f^3: I was writing a story right now.
... ...
$P_f(\widetilde{X})$

Figure 2: The commonly used methods of DA for NMT, where $P_r(X)$ denotes the real data distribution, $P_f(\tilde{\mathbf{x}})$ represents the fake data distribution, and "X_f^1", "X_f^2" and "X_f^3" refer to augmented data generated from real data using common approaches such as swapping, replacing and BT.

- We take the constrained sampling method as an augmenter rather than employing the random replacement during sampling from real data.
- We propose the model transparent approach, which can be fit into different MT architectures. Our method is not only language independent but also extendable for other NLP tasks.
- We design a discriminator sub-model as an evaluator to mitigate the errors in augmented sequences instead of leveraging commonly used LM.

2 Background

2.1 Neural Machine Translation

We can regard X as a source language sentence and Y as a target language sentence. Given a source sentence $\mathbf{x} = x_1, \dots, x_i, \dots, x_I$ and a target sentence $\mathbf{y} = y_1, \dots, y_j, \dots, y_J$, standard NMT models [1, 2, 3] usually factorize the sentence-level translation probability as a product of word-level probabilities:

$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \prod_{j=1}^{J} P(y_j|\mathbf{x}, \mathbf{y}_{<j};\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ is a set of model parameters, and $\mathbf{y}_{<j}$ is a partial translation.

Let $\langle \mathbf{X}, \mathbf{Y} \rangle = \{\langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$ be a training corpus. The log-likelihood of the parallel training data is maximized by the standard training objective:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N} \log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)};\boldsymbol{\theta}) \right\}. \tag{2}$$

The translation decision rule for unseen source sentence $\mathbf{x}$ learned model parameters $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ P(\mathbf{y}|\mathbf{x};\hat{\boldsymbol{\theta}}) \right\}. \tag{3}$$

Meanwhile, calculating the highest probability $\hat{\mathbf{y}} = \hat{y}_1, \dots, \hat{y}_j, \dots, \hat{y}_J$ of the target sentence can be separated at the word level:

$$\hat{y}_j = \underset{y}{\operatorname{argmax}} \left\{ P(y|\mathbf{x}, \hat{\mathbf{y}}_{<j};\hat{\boldsymbol{\theta}}) \right\}. \tag{4}$$

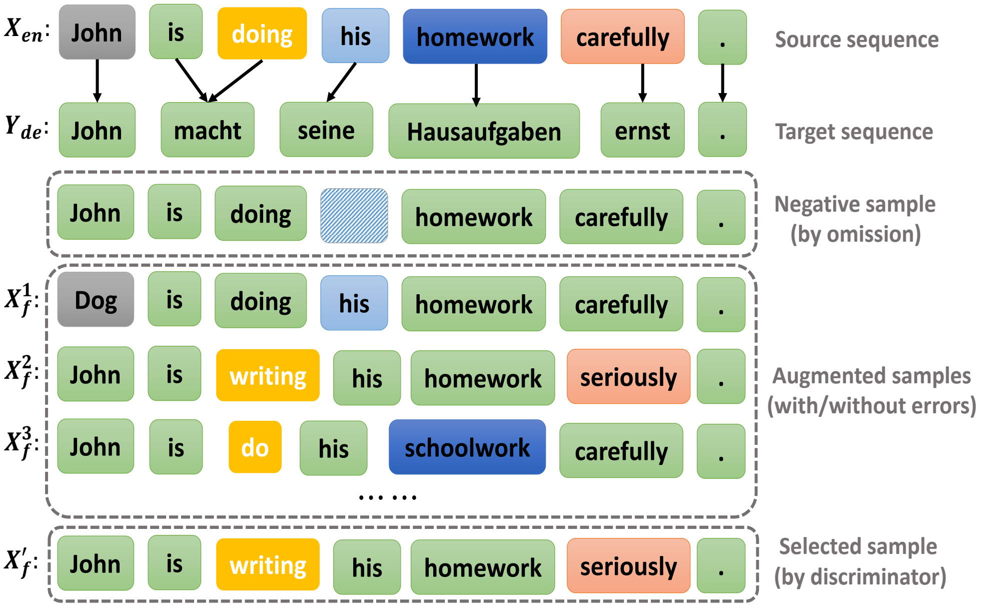

Figure 3: The source augmentation process with Constrained Sampling method in En - De task. The generated sample sets $\{x_f^1, x_f^2, x_f^3, \dots\}$ represent common DA tricks which hard to avoid semantic (x_f^1) or syntactic (x_f^3) errors. While our method prefer to reduce errors using discriminator model after augmentation.

2.2 Data Augmentation

The DA method has been widely used in various tasks, such as computer vision [11], dialogue generation [18], machine translation [16] and other NLP tasks [12]. Therefore, DA is still a ubiquitous and pervasive method in NMT, which samples some fake data distribution $P_f(\tilde{\mathbf{x}})$ using some common methods (see Figure 2) based on the real data distribution $P_r(\mathbf{x})$. For instance, the existing DA approaches for NMT mainly include swapping two words randomly [19], using BT (back-translation) [20] and replacing words [16] with different words among a given sequence. Intuitively, for a real sequence $X = w_1, w_2, \dots, w_m$, which is composed of m words the aforementioned methods encourage the augmentation model to generate the fake sequence $X_f = \tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_n$ by leveraging random manner. However, it is quite hard to guarantee the quality of produced sequences frequently.

3 Methodology

Let $\mathbf{x} = x_1, \dots, x_i, \dots, x_I$ be a source sentence with I words and $\mathbf{y} = y_1, \dots, y_j, \dots, y_J$ be a target sentence with J words. We use $D_r = \{\langle \mathbf{x}^{(m)}, \mathbf{y}^{(m)} \rangle\}_{m=1}^{M}$ be the original training data with M sentences. As illustrated in Figure 3, we formulate the DA task as follows: given the $P_r(\mathbf{x})$, the task is to train the augmentation model based on the $P_r(\mathbf{x})$. We expect our model to generate $P_f(\tilde{\mathbf{x}})$, where $P_f(\tilde{\mathbf{x}})$ is suited to $P_r(\mathbf{x})$ well.

3.1 Constrained Sampling

We are highly inspired by [21] to exploit the constrained sampling method for NMT in LRLs. Our augmentation steps are the same as [22] but we substitute the words with different sampling manner rather than random sampling. Let $P_r(\mathbf{x})$ be the real data distribution. We use $\tilde{\mathbf{x}}$ to denote an artificial source sentence augmented from the original sentence $\mathbf{x}$. We define $P_f(\tilde{\mathbf{x}}|\mathbf{x})$ as the distribution of generating $\tilde{\mathbf{x}}$ given $\mathbf{x}$. As a result, the distribution of augmented data can be represented as

$$P_f(\tilde{\mathbf{x}}) = \mathbb{E}_{\mathbf{x} \sim P_r(\mathbf{x})} \Big[P_f(\tilde{\mathbf{x}}|\mathbf{x}) \Big]. \tag{5}$$

From $\mathbf{x}$ to $\tilde{\mathbf{x}}$, there is a unique sequence of replacing operators. For example, let $\mathbf{x}$ be "I like the book" and $\tilde{\mathbf{x}}$ be "I like a movie". The edit distance is 2, the positions of words to be replaced are 3 and 4, and the replacing words are "a" and "movie". More formally, we use d to denote the edit distance between $\mathbf{x}$ and $\tilde{\mathbf{x}}$, $\mathbf{p} = \{p_1, \dots, p_d\}$, and

$\mathbf{w} = \{w_1, \ldots, w_d\}$ be the list of replacing words. In the above example, $d = 2$, $p_1 = 3$, $p_2 = 4$, $w_1 =$ "a", and $w_2 =$ "movie". According to the original real data distribution $P_r(\mathbf{x})$, we define the augmentation distribution as

$$\begin{aligned} & P_f(\tilde{\mathbf{x}}|\mathbf{x}) \\ = \; & P(d, \mathbf{p}, \mathbf{w}|\mathbf{x}) && (6) \\ = \; & P(d|\mathbf{x})P(\mathbf{p}|\mathbf{x}, d)P(\mathbf{w}|\mathbf{x}, d, \mathbf{p}) && (7) \\ = \; & P(d|\mathbf{x})\prod_{i=1}^{d} P(p_i|\mathbf{x}, d)P(w_i|\mathbf{x}, d, \mathbf{p}). && (8) \end{aligned}$$

More precisely, the constrained sampling method is composed as follows:

1. To sample an edit distance d based on real data sample $\mathbf{x}$, then, following [22], for some input sentences, we define the edit distance sub-model as

$$P(d|\mathbf{x}) = \frac{\exp\{-d/\tau\}c(d, I)}{\sum_{d'=0}^{I} \exp\{-d'/\tau\}c(d', I)}, \tag{9}$$

   where τ denotes the temperature hyper-parameter and it restricts the search space surrounding the original sentence. We can infer that the larger τ obtains more samples with long edit distances. Also, the $c(\cdot)$ is given by

$$c(d, I) = \binom{I}{d} \times (|\mathcal{V}| - 1)^d, \tag{10}$$

   where $c(d, I)$ stands for the variety of sentences whose edit distance to a sentence of length I is $d(d \in \{0, 1, 2, 3, \ldots, I\})$ and note that $\mathcal{V}$ is the vocabulary.

2. To replace the words, we need to select the positions in terms of the sampled edit distance d. We define the position sub-model as

$$P(p|\mathbf{x}, d) = \frac{d}{I}, \tag{11}$$

   in terms of the previous sampling approach in which we achieve the position set $\{p_1, p_2, p_3, \ldots, p_d\}$. This method roughly guarantees the edit distance d between the new sentence and the initial sentence.

3. Our model verifies a new word for the replacement, at each sampled position p_j. From X_{j-1} to $X_j(j = 1, 2, 3, \ldots, d)$, at each step, we can sample the new word w_j from the distribution $P(w|X_{j-1}, p = p_j)$, and then we switch the old word at the position p_j of X_{j-1} to achieve X_j. Finally, the replacing sub-model is defined as

$$P(w_i|\mathbf{x}, d, \mathbf{p}) = P(w_i|x_{i-1}, p_i). \tag{12}$$

   The random sampling paradigm samples a new word w_j in terms of a uniform distribution over vocabulary $\mathcal{V}$. Additionally, the constrained sampling scenario [23] samples w_j to maximize the score of the LM of the sequence X_j. The aforementioned commonly used sampling methods are available for modeling the $P(w_j X_{j-1}, p = p_j)$, In the experiment, we compare our method with random sampling and a highly similar sampling approach called "Hamming Distance Sampling" [24]) which is used in MT on different three corpora with several languages.

3.2 Evaluation Sub-model

In this work, to increase the quality of augmented fake data, we train the discriminator ϕ on both the original real data D_r and negative instances D_n, which are generated by using negative sampling. As shown in Algorithm 1, $\phi(\mathbf{x}, \tilde{\mathbf{x}})$ plays the role of a filter after augmentation using constrained sampling. To reduce the errors among generated data distribution $P_f(\tilde{\mathbf{x}})$, we design the evaluation sub-model by taking advantage of $\phi(\mathbf{x}, \tilde{\mathbf{x}})$. Discriminator focuses on separating $P_r(\mathbf{x})$ and $P_f(\tilde{\mathbf{x}})$ which is similar to that of GANs. We are also encouraged by Least-Square GAN [25], and we set the loss function as below:

Algorithm 1 Constrained Sampling for Data Augmentation

```
Input: Original data $D_r = \{\langle \mathbf{x}^{(m)}, \mathbf{y}^{(m)} \rangle\}_{m=1}^{M}$.
Output: Augmented data $D_a$.
Use negative sampling to build a set of negative samples $D_n = \{\langle \bar{\mathbf{x}}^{(m)}, \mathbf{y}^{(m)} \rangle\}_{m=1}^{M}$.
Train discriminator $\phi(\mathbf{x}, \tilde{\mathbf{x}})$ using $D_r$ and $D_n$.
Sample $\tilde{\mathbf{x}}$ using $P_f(\tilde{\mathbf{x}}|\mathbf{x})$ from $D_r$ to build a coarse augmented dataset $D_c$.
Use $\phi(\mathbf{x}, \tilde{\mathbf{x}})$ to filter $D_c$ to obtain the final augmented dataset $D_a = \{\langle \tilde{\mathbf{x}}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$.
Return the new training data $D_r \cup D_a$ with $M + N$ sentence pairs.    ▷ $N = 5$ default value
```

$$\begin{aligned}\mathcal{L}_\phi = &\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim P_r(\mathbf{x})}[(\phi(\mathbf{x}) - 1)^2] \\ &+ \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim P_n(\mathbf{x})}[(\phi(\mathbf{x}))^2],\end{aligned} \tag{13}$$

where loss function $\mathcal{L}_\phi$ makes the discriminator ϕ allocate higher rewards to real data $P_r(\mathbf{x})$ than augmented fake data $P_f(\tilde{\mathbf{x}})$. In this way the ϕ can deliver more rewards and helps the augmentation model to select the generated data $P_f(\tilde{\mathbf{x}})$ with higher quality. Note that it is easy to extend to augment the target side and both sides.

4 Experiments

4.1 Setup

4.1.1 Data preparation

The source side of the language pairs used in our main experiment are Azerbaijan (Az), Hindi (Hi), Uyghur (Ug), Uzbek (Uz), and Turkish (Tr) and were obtained from Tanzil corpora[1]. Furthermore, the language pair English (En) - German (De) is achieved from WMT14[2] while the other two language pairs De - En and Vietnamese (Vi) - En are attained from IWSLT14 and IWSLT15[3], respectively. All the training corpora are publicly available and the specifications of the corpora are listed in Table 1. We set the target side to English, and the source sides include morphological rich LRLs. Moreover, for the pair of En - De , we combine the *newstest2012* and *newstest2013* as the development set, and take *newstest2014* as the test set. Besides, for the task of De - En, we split the original data into two different corpus sizes "160K" and "7K" and allocate them for the training set and the validation set separately. Meanwhile, we merge *dev2020, dev2010, tst2010, tst2011, tst2012* as the test set. Additionally, for the task of Vi -En, we take the *tst2012* and *tst2013* as the validation set and the test set, respectively. The process of building the vocabulary are different among all language pairs. Precisely, we build the source and target vocabulary by 32K BPE [4] method [26] without jointly training for Tanzil corpus and with jointly training for WMT14 and IWSLT15 corpora. Moreover, we build the source and target vocabulary for IWSLT14 by 10K BPE. In addition, to train the baselines BT and Copy we exploit the monolingual corpora in En and De, from WMT17[5]. As shown in Table 2, the original real target sentence is augmented by leveraging the constrained sampling approach from the target side of Ug - En.

We use the pre-processing *script* to clean up the data by removing the duplicated sentences, removing the blank lines in the corpus, and cleaning the mismatched length sentences. We leverage the `tokenizer` toolkit [7] [6] for word tokenization and do not use any UNK-replacement techniques. In addition, we employ an open-source toolkit FairSeq[7] for the NMT system, together with Transformer architecture (base model), to train and evaluate the baselines Trans, Bt, and Copy. Meanwhile, for other baselines Swap, Drop, Blank, Smooth, Switch, and Sca, we also take advantage of the FairSeq toolkit. Among them the baselines Switch[8] and Sca[9] use their own codes which were also implemented with FairSeq. Likewise, we ran all the experiments for 50 epochs with FairSeq and 100K iter steps with

[1] http://opus.nlpl.eu/Tanzil.php
[2] https://nlp.stanford.edu/projects/nmt/data/
[3] https://wit3.fbk.eu
[4] https://github.com/rsennrich/subword-nmt
[5] http://www.statmt.org/wmt17/
[6] https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer/tokenizer.perl
[7] https://github.com/pytorch/fairseq
[8] https://github.com/cindyxinyiwang/fairseq
[9] https://github.com/teslacool/SCA

Table 1: Characteristics of our corpora. While “Dev." and “Test" indicate the validation set and test set, “Vocab.” and “Token.” denote vocabulary and all tokens, respectively. The “Avglen” represents the average length of sentences.

Language Pairs	Train	Dev.	Test	Source			Target		
				Vocab.	Token.	Avglen.	Vocab.	Token.	Avglen.
Az → En	21.2K	1.0K	1.0K	24.6K	3.1M	14.50	18.9K	4.0M	18.77
Hi → En	182.0K	1.0K	1.0K	13.3K	7.3M	39.97	20.1K	5.0M	27.09
Ug → En	81.1K	1.0K	1.0K	18.3K	2.1M	25.61	19.5K	2.2M	26.45
Uz → En	134.6K	1.0K	1.0K	25.2K	2.2M	15.97	20.1K	2.5M	18.12
Tr → En	141.9K	1.0K	1.0K	86.0K	0.8M	5.85	54.4K	1.0M	6.86
En → De	4.5M	6.0K	2.7K	817.0K	116.1M	26.0	1.7M	109.7M	24.5
De → En	160K	7.3K	6.8K	113.5K	3.1M	19.35	53.3K	3.3M	20.44
Vi → En	140.5K	1.6K	1.3K	25.3K	3.5M	24.64	48.64K	2.9M	20.10

Table 2: The augmentation samples by using constrained sampling method. In this table, we give augmentation samples of target side in the language pair Ug - En. The “Target$_{ori}$” denotes an original target side before being augmented.

Method	Augmentation Sample
Target$_{ori}$	it was **generally anticipated** that fortuyn could have achieved a **major breakthrough** in the general election.
Our work	it was generally anticipated that fortuyn could have achieved a major turning-point in the general election. it was generally expected that fortuyn could have achieved a major breakthrough in the general election. it was generally anticipated that fortuyn could have achieved a important breakthrough in the general election. it was usually anticipated that fortuyn could have achieved a major breakthrough in the general election.

Table 3: Hyper-parameter settings While the “Sample size" represents the augmentation sample size for sampling step.

Parameter	Value
Temperature (τ)	0.85
Sample Size	5
Word Embedding	620
Hidden State	512
Vocabulary Size	30K
Batch Size	80
Sequence Length	50
Beam Size	4
Dropout	0.1
Learning Rate	1.0

THUMT on 4 GPUs (TITAN X) using default parameters. The other main hyper-parameters used in our experiment are shown in Table 3. We use the case-sensitive BLEU[10] [27] score and TER[11] [28] score to evaluate the translation performance. Besides, we use the pre-tanned language model GPT-2 [29] to calculate the language perplexity. Except for the language perplexity, we also use the ROUGE[12] [30] score to further evaluate and compare the quality of our result to other baseline systems.

4.1.2 Baselines

It stands to reason to compare our proposed method with highly similar approaches. Therefore, we select analogous approaches as follows:

[10] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[11] https://github.com/jhclark/tercom

[12] https://github.com/pltrdy/rouge

- TRANS (Transformer): is the SOTA architecture in NMT system [3]. It is one of the commonly used and well known architectures in the NMT scenario. No matter in the building of both the HRLs or LRLs systems, the majority of tech engineers and researchers prefer to exploit this architecture.
- BT (Back Translation): uses the monolingual data to augment the parallel corpus in MT [20]. This training method needs the pre-trained MT model to leverage the large amounts of monolingual data to generate the pseudo corpus, and finally combining them with the original corpus to expand the initial size of the corpus. However, it is hard to guarantee that it can obtain higher quality data.
- COPY (Copy Translation): creates the text from monolingual data in the target side, so that each source sentence is identical to the target sentence [31]. Copied data mixed with the parallel data to train MT normally. Specifically, they train a NMT system to both translate source language text and copy target-language text, thereby exploiting monolingual corpora in the target language.
- SWAP (Swapping): replaces the words in neighboring positions within a window size k randomly [14, 19]. Clearly, the core idea is building the fake data by taking advantage of randomly swapping the words in a sentence. As well as, just merely shuffle the sentence between each words. The model consists of a slightly modified attention encoder-decoder model that can be trained on monolingual corpora alone using a combination of denoising and BT.
- DROP (Dropping): drops the word tokens arbitrarily from the sequence [15]. The key idea is randomly dropping the word among the current sentence to build the pseudo corpus. It may make the perplexity higher and higher after augmentation. Namely, they provide the deep averaging network, and it can be improved by applying a novel dropout-inspired regularizer: for each training instance, randomly drop some of the tokens' embeddings before computing the average.
- BLANK (Blanking): substitutes the word tokens with placeholder token "_" without considering the quality of the generated data [12]. However, the blank noising can be seen as a way to avoid over-fitting on specific contexts.
- SMOOTH (Smoothing): changes the word tokens with a sample that is achieved from unigram distribution over vocabulary [12]. They consider the maximum likelihood estimate of n-gram probabilities estimated using the pseudo counts of the augmented data.
- SWITCH (Switch-out): samples the data by leveraging the hamming distance sampling method in MT [24]. Precisely, they formulate the design of a data augmentation policy with desirable properties as an optimization problem, and derive a generic analytic solution. They also randomly replace words in both the source sentence and the target sentence with other random words from their corresponding vocabularies.
- SCA (Soft Context Augmentation): data augmentation for NMT replaces randomly chosen word with a soft distributional representation [32] to leverage the bi-lingual language model. More accurately, they replace the one-hot representation of a word by a distribution (provided by a language model) over the vocabulary, i.e., replacing the embedding of this word by a weighted combination of multiple semantically similar words.

4.2 Main Result

4.2.1 Comparison on LRLs

As shown in Table 4, we compare our method with all the baselines on MRLs and LRLs from the Tanzil corpus. The performance of BT is not consistently improved on many languages. Clearly, BT does not produce a better result than Trans in Hi - En. To contrast, the COPY also achieves even worse results than Trans and BT. It gains weak improvements in Az - En, Hi - En and Uz - En, the main reason may be because of the key idea in COPY (see Session 4.1.2), and there were big distinctions between source and target pairs, as well as they share fewer words between them. Namely, it copies the monolingual data in the target as the same size as the source and combine them with original data to train NMT. However Ug - En and Tr - En obtain better results than the previous baselines. Intuitively, we believe that the main reason might be that they have many shared words between Ug and En, Tr and En. Another reason might be that we convert Ug into a unified transliteration type with En by following the idea of [33]. Thus, COPY achieves better results than BT and Trans in Ug - En and Tr - En. Obviously, other baselines constantly outperform over Trans, BT, and COPY. Our method was more efficient no matter what the source and targets were. The sampling with both the source and target sides consistently better than the previous baselines.

4.2.2 Comparison on WMT and IWSLT

To further validate the effectiveness of our approach, we also compare with whole the baseline systems on other different language pairs from WMT14, IWSLT14, and IWSLT15, respectively. As given in Table 5, it is apparent that the BT and

Table 4: The comparison with BLEU score between baseline systems on LRLs from Tanzil corpora. While "Augment$_{source}$, Augment$_{target}$ and Augment$_{source+target}$ " represent augment source side, target side, and both sides, respectively. The "$\star$" and $\star\star$" denote significantly better than best baselines with $p < 0.01$ and with $p < 0.05$.

Method	Tanzil Corpora				
	Az - En	Hi - En	Ug - En	Uz - En	Tr - En
Trans (Vaswani et al., 2017) [3]	21.03	20.15	19.19	17.76	22.72
Bt (Sennrich et al., 2016) [20]	21.32	19.20	20.01	18.72	24.95
Copy (Currey et al., 2017) [31]	20.34	19.80	20.35	17.48	23.82
Swap (Artetxe et al., 2017) [14]	20.32	21.33	21.76	19.21	25.08
Drop (Iyyer et al., 2015) [15]	21.19	20.78	21.72	19.30	25.77
Blank (Xie et al.,2017) [12]	23.23	20.40	21.67	19.39	25.44
Smooth (Xie et al.,2017) [12]	25.88	21.87	22.24	19.48	25.85
Switch (Wang et al., 2018) [24]	**26.36**	**22.61**	**23.16**	19.65	25.54
Sca (Zhu et al., 2019) [32]	25.32	22.16	22.90	**19.77**	**25.92**
Our work					
Augment$_{source}$	27.37	23.53	22.94	**21.22***	26.17
Augment$_{target}$	26.76	22.74	23.43	19.76	26.04
Augment$_{source+target}$	**27.59***	**23.68***	**23.67****	20.14	**26.66***

Table 5: The comparison with BLEU score between baselines on three languages from WMT and IWSLT. The $\star\star$" denotes significantly better than best baselines with $p < 0.05$.

Method	WMT14	IWSLT14	IWSLT15
	En - De	De - En	Vi - En
Trans (Vaswani et al., 2017) [3]	27.24	33.53	25.32
Bt (Sennrich et al., 2016) [20]	27.30	33.69	26.34
Copy (Currey et al., 2017) [31]	27.27	34.62	26.45
Swap (Artetxe et al., 2017) [14]	27.19	33.98	26.98
Drop (Iyyer et al., 2015) [15]	27.22	34.68	27.35
Blank (Xie et al.,2017) [12]	27.32	34.83	27.80
Smooth (Xie et al.,2017) [12]	27.48	34.85	**29.31**
Switch (Wang et al., 2018) [24]	27.39	34.75	28.58
SCA (Zhu et al., 2019) [32]	**27.51**	**34.89**	29.23
Our work			
Augment$_{source}$	27.57	34.93	**29.88****
Augment$_{target}$	27.63	34.98	29.61
Augment$_{source+target}$	**27.94** **	**35.14****	29.79

Table 6: The effectiveness of hyper-parameter temperature τ for augmentation methods, while "τ" denotes the hyper parameter temperature, we augment the source side with different τ.

Temperature (τ)	Tanzil Corpus		IWSLT Corpus	
	Ug - En	Tr - En	De - En	Vi - En
0.75	22.39	26.00	34.76	29.80
0.80	22.90	26.16	34.92	29.85
0.85	**22.94**	**26.17**	**34.93**	**29.88**
0.90	22.31	26.07	34.90	29.78
0.95	22.29	25.92	34.85	29.61

Copy obtain fewer improvements than other baselines with word transmission. There may be have two main reasons. Firstly, the WMT14 has a larger corpus size than other language pairs, and the data sampling method does not work. Secondly, the language pair in IWSLT15 has less shared vocabulary. After mixing the copied target side with the source side it may degrade the quality of data (especially in Vi - En task). However, other random sampling methods also obtained a better performance, yet were not as obvious as our method was. The sampling of both the source and target also enhances the generalization skill of the NMT model.

Table 7: The performance of our model with and without exploiting discriminator sub-model. The “Dis." discriminator sub-model (ϕ). While “$\times$" stands for without using discriminator sub-model and “$\surd$" denotes leveraging the discriminator sub-model. The $\star\star$” denotes significantly better with $p < 0.05$.

Discriminator	Tanzil Corpus		IWSLT Corpus	
	Ug - En	Tr - En	De - En	Vi - En
$\times$	23.25	26.16	34.85	29.61
$\surd$	**23.67****	**26.66****	**35.14****	**29.79****

Table 8: The comparison with TER score (the smaller the better) between different methods for LRLs from Tanzil corpora. The “$\star$” and $\star\star$” denote significantly better than best baselines with $p < 0.01$ and with $p < 0.05$.

Method	Translation Error Rate (TER) ↓				
	Az - En	Hi - En	Ug - En	Uz - En	Tr - En
Trans (Vaswani et al., 2017) [3]	66.72	66.73	69.51	76.96	68.74
Bt (Sennrich et al., 2016) [20]	67.17	69.07	71.13	75.54	68.88
Copy (Currey et al., 2017) [31]	68.79	69.80	80.29	79.00	68.89
Swap (Artetxe et al., 2017) [14]	67.38	70.35	65.88	70.57	66.07
Drop (Iyyer et al., 2015) [15]	67.74	72.49	68.25	71.85	65.13
Blank (Xie et al.,2017) [12]	64.60	68.88	68.18	69.71	65.61
Smooth (Xie et al.,2017) [12]	60.64	67.51	66.85	70.95	64.39
Switch (Wang et al., 2018) [24]	**60.38**	63.59	**64.68**	**69.39**	65.60
Sca (Zhu et al., 2019) [32]	60.90	**62.32**	65.98	71.47	**64.15**
Our work					
Augment$_{source}$	**58.27****	**60.14****	**61.47***	**66.89***	**64.04**

4.2.3 Effect of Temperature τ

As we have mentioned in Session 3.2, the temperature hyper-parameter τ manages the search space neighboring the real data distribution when the edit distance is calculated (see Eq.(9)). To explore the impact of τ on the performance of the augmentation model, we test the sampling method with four language pairs from the Tanzil and IWSLT corpora via fixing other hyper parameters. As shown in Table 6, the augmentation skill of our method with $\tau = 0.85$ is better than other values on four language pairs from two corpora. Beside, the results achieved from $\tau = 0.80$ are also highly close to the value of $\tau = 0.85$. Clearly, the smaller ($\tau = 0.75$) or the bigger value ($\tau = 0.95$) are unable to bring beneficial results.

4.2.4 Ablation Study

Our proposed model consists of two parts: augmentation sub-model and evaluation sub-model. We investigate the impacts of τ on the performance of the augmentation model. Meanwhile we also explore the effectiveness of the evaluation sub-model to the entire architecture. As given in Table 7, the discriminator model plays a vital role in selection of augmented data after generation. Clearly, when we exploit the discriminator sub-model as our data selector, the model performance is better than without using the discriminator sub-model. We believe that if we could increase the performance of our discriminator, we would achieve even better results than the current version and we can enhance the quality of augmented data homogeneously.

4.2.5 Comparison with TER Score

Generally, we have mentioned many augmentation approaches for NMT in Session 4.1.2. To investigate the influence of these methods for the quality of augmented data, we compare their performance using the Translation Error Rate (TER) score, which is another well known evaluation standard in the MT community. As Table 8 shows, the commonly used word transformation methods with random sampling, such as shuffling (Swap), Drop, and Bt hurt the model quality. The majority of them ignored the evaluation of the augmented data, or selected the sequence just using LM. Our augmentation method and discriminator sub-model is more effective than others. Additionally, it is easy to observe from the Table 8 that, the stronger baselines Switch and Sca obtain remarkably better results than previous baselines, but not too steady on different language pairs. In contrast, our model achieves consistently better results than all the pipelines on whole LRLs from Tanzil corpora.

Table 9: The comparison with language perplexity (the smaller the better) between different methods for LRLs from Tanzil corpora.

Method	Language Perplexity ↓				
	Az - En	Hi - En	Ug - En	Uz - En	Tr - En
Trans (Vaswani et al., 2017) [3]	23.28	22.62	22.33	23.98	23.64
BT (Sennrich et al., 2016) [20]	22.78	21.49	20.23	22.19	23.47
Copy (Currey et al., 2017) [31]	22.88	21.51	21.30	23.70	23.60
Swap (Artetxe et al., 2017) [14]	21.16	18.19	17.34	19.34	23.26
Drop (Iyyer et al., 2015) [15]	**20.73**	21.30	17.81	**18.58**	23.40
Blank (Xie et al.,2017) [12]	23.14	21.94	17.35	19.48	23.48
Smooth (Xie et al.,2017) [12]	22.24	17.25	18.52	20.31	23.45
Switch (Wang et al., 2018) [24]	21.57	**17.17**	**16.81**	18.94	23.19
SCA (Zhu et al., 2019) [32]	21.20	16.42	17.34	19.42	**22.86**
Our work					
Augment$_{source}$	**18.47**	**16.31**	**16.66**	**18.27**	**21.98**

Table 10: The comparison with ROUGE score (Rouge-1) between different methods for LRLs from Tanzil corpora. While the "P." , "R." and "F." represent the evaluation indicators Precision, Recall and F-measure,respectively. The "$\star$" and $\star\star$" denote significantly better than best baselines with $p < 0.01$ and with $p < 0.05$.

Method	ROUGE score ↑ (Rouge - 1)											
	Az - En			Ug - En			Uz - En			Tr - En		
	P.	R.	F.	P.	R.	F.	P.	R.	F.	P.	R.	F.
Trans [3]	52.55	52.04	51.51	51.39	47.00	48.25	47.49	47.41	45.93	51.90	53.22	51.50
Bt [20]	54.44	53.54	53.23	49.56	52.02	49.86	50.71	48.68	48.16	51.94	53.49	51.64
Copy [31]	51.65	51.93	51.01	49.60	49.83	48.92	46.81	47.09	45.41	53.34	53.23	52.21
Swap [14]	55.32	51.89	52.74	56.95	52.97	54.06	54.86	47.49	49.13	54.29	54.83	53.62
Drop [15]	56.32	53.34	53.88	55.04	53.30	53.35	54.27	47.16	48.44	54.76	54.98	53.95
Blank [12]	54.42	54.51	53.81	56.81	54.04	54.51	**56.87**	47.12	49.77	54.60	54.98	53.85
Smooth [12]	57.61	56.83	56.56	56.59	53.31	54.07	54.60	46.70	48.60	54.43	54.50	53.62
Switch [24]	**58.34**	57.32	**57.12**	58.20	54.58	55.49	55.12	49.22	**50.40**	**54.82**	54.82	**53.96**
Sca [32]	57.19	56.09	55.96	**58.41**	54.57	**55.53**	53.69	47.75	48.86	54.63	54.66	53.81
Our work												
Augment$_{source}$	**59.80****	56.28	**57.21**	**60.99***	53.03	**55.67**	**57.23****	50.68	**52.16***	**54.83**	55.07	**54.00**

4.2.6 Comparison with Perplexities

To consider the adequacy and the fluency of the translation result, and to further validate the effectiveness of our proposed model we compare to other baseline systems by using language perplexity. As shown in Table 9, our method achieves better performance between whole baseline systems comparison with language perplexity on LRLs from Tanzil corpora. In contrast, the baseline Drop achieves better improvements in Az-En and Uz-En, the similar baseline Switch also gains better result in Hi-En and Ug-En, and the strong baseline Sca obtains better performance in Tr-En. No previous system consistently achieved a better result on whole LRLs; however, our method constantly outperforms other baselines on all low-resource language pairs.

4.2.7 Comparison with ROUGE Score

To further validate the quality of the generated data, we compare the generalization skills of our mode with all pipelines. We also employ the ROUGE score (the greater the better) that is commonly used to compare model performance, as well as the adequacy and fluency of the generated data. As shown in Table 10, the majority of techniques that are used to build the fake data also achieve better results (Rouge-1 score), but our method significantly outperformed all the baselines on LRLs from Tanzil corpora. As depicted in Figure 4, we also further compare the effectiveness of these approaches for the performances of generated pseudo data using Rouge-2 score on the language pair of Ug - En and our model obtained better results than others. As illustrated in the line chart in Figure 5, the stronger baselines Switch and Sca also obtained more effective results than previous baseline systems. However, they ignore the quality and the evaluation after generating the data. Thus our proposed model gains more effective results than others on Ug - En with Rouge-L score.

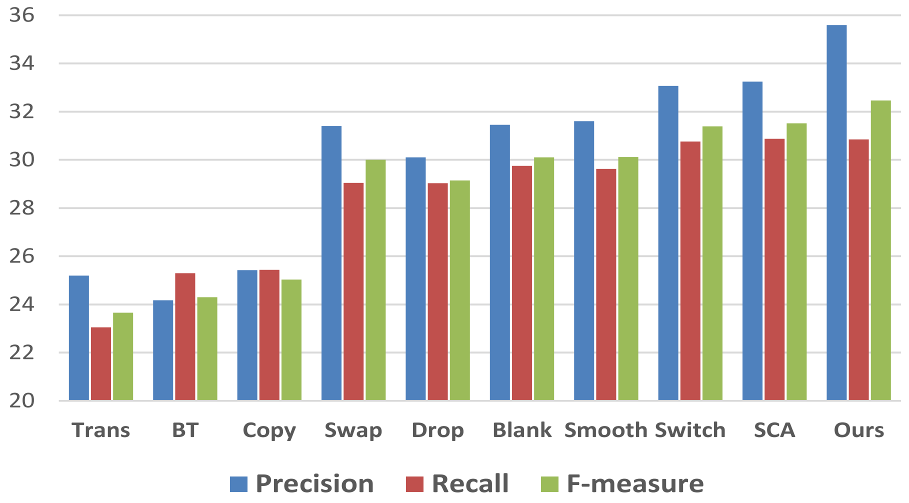

Figure 4: The comparison with ROUGE score (Rouge-2) between different methods for the language pair of Ug - En. While “Ours" represents the results of source augmented method Augment$_{source}$.

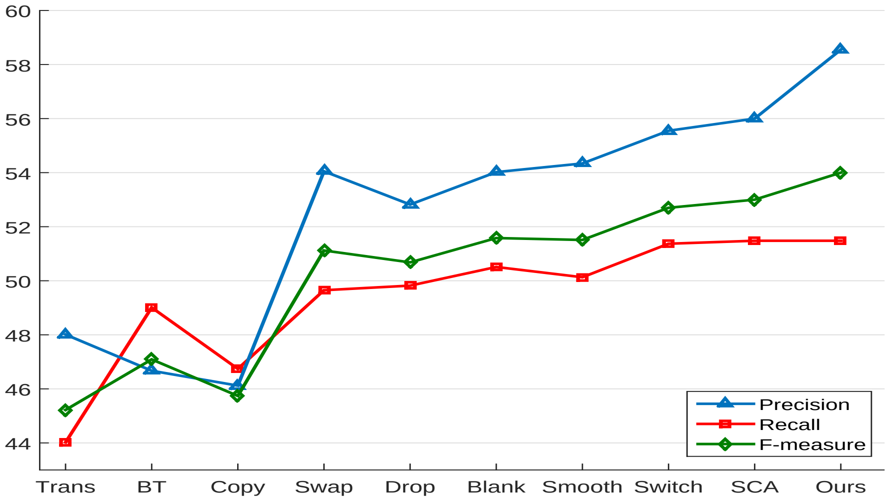

Figure 5: The comparison with ROUGE score (Rouge-L) between different methods for the language pair of Ug - En. While “Ours" represents the results of source augmented method Augment$_{source}$.

4.2.8 Case Study

Table 11 indicates the translation examples from various systems. It is effortless to find that most of the baseline methods suffer from the syntactic errors like “the religious learning men forbids them” from SWAP, while BT generates replicated statement like “forbid from eating illegal illegal things”. The baseline BLANK also produces some confusing expressions like ‘why do not the religious learned men forbid them” and also dropped the subject (“priests”) of the source sentence. In addition, the baseline COPY achieves irrelevant words at the end of the translation result, while the SCA achieves better results than all the previous techniques. In contrast with the above-mentioned approaches (all baselines), our method achieves good results that are analogous to the reference by sampling a higher quality sequence.

5 Related Work

As a key potential strategy for overcoming the grave problem posed by the shortage of large-scale parallel corpora, NMT has gained increasing attention in the MT community [26, 3, 34, 35, 36]. Several methods have been presented that aim to handling the issue of deficiency of the parallel training corpora in NMT for LRLs. Most of the current literature can be classified into meta-learning [37, 38], transfer learning [39, 33, 40], domain adaptation [41, 42], pivot learning [43], and

Table 11: Translation example of Tr - En between various DA methods. The "Ref." represents standard reference.

Method	Translation result
Source	din adamları ve alimleri onları günah olan sözleri söylemekten ve haram yemekten menetselerdi ya
Reference	why do not their priests and monks **forbid** them from speaking evil and devouring the **forbidden**
TRANS [3]	*wherefore is it that the divines and priests forbid them not of their speaking* ***of sin*** *and their devouring of the forbidden*
BT [20]	*why do not the religious learned* ***men*** *forbid them from* ***uttering sinful*** *words and from eating illegal* ***illegal*** *things*
COPY [31]	*why do the masters and the rabbis not* ***forbid*** *them to utter sin and* ***consume*** *the unlawful*
SWAP [14]	*why do not the rabbis and the religious* ***learning men forbids*** *them from* ***uttering sinful*** *words and from eating illegal things*
DROP [15]	*why do not the priests and religious learned men* ***forbid*** *them from saying sinful words and from eating illegal things*
BLANK [12]	*why do not the* ***religious*** *learned men forbid them from uttering sinful words and from having illegal things*
SMOOTH [12]	*why do not the* ***rabbis*** *and the religious forbid them from* ***talking*** *sin words and from eating illegal things*
SWITCH [24]	*why do not the rabbis and the religious learned men forbid them from* ***saying*** *sinful words and from eating illegal things*
SCA [32]	*why do not the rabbis and the religious learned men forbid them from* ***saying*** *sinful words and forbid from the devouring*
Our work	
Aug$_{src}$	*why do not* ***their*** *rabbis and* ***monk forbid*** *them from talking wicked words and consuming* ***forbidden*** *things*

zero-shot learning [10, 44, 45]. Spontaneously, NMT systems have developed rapidly in recent years. [46] enhances the performance of NMT, leveraging word-level domain context in the multi-domain community. Meanwhile, [47] presents the iterative dual training method for domain adaptation tasks in NMT. Moreover, [48] commits to distinguishing and exploiting different word-level domain contexts for multi-domain NMT, and improves the NMT model generalization skill by exploiting multi-task learning to jointly model NMT and monolingual attention-based domain classification tasks. In the past two years, many researchers proposed new approaches [16] in the NMT community for LRLs. [33] proposes the multi-round transfer learning for LRLs in NMT. [37] presents meta-learning for low-resource NMT and achieves a remarkably better result. [49] discusses some pitfalls to be aware of when training low-resource NMT systems and recent techniques that have shown to be especially helpful in low-resource settings.

Additionally, there are also a variety of techniques for data augmentation in the NMT scenario. To our knowledge, these approaches can be categorized into three types. The first type is based on BT that leverages monolingual data to augment the training corpus guided by the trained MT model. It is an effective method, but usually helpless to errors. The second type is word transformation with or without language model. For instance, swapping [14], replacing [16], switching [17], omitting [19], and context knowledge [32] methods. The third type is based on various sampling methods. [24] also proposes an effective method that uses hamming distance sampling by following the idea of [22], and takes the DA as an optimization problem. Besides, [18] also takes advantage of the edit distance sampling method by further considering the instability of augmentation in the dialogue generation. This method also can be seen as another revised version of [22]. The techniques belong to the previous two types, more or less, and have ignored the selection of higher quality sequence after generation. The methods in the last type (e.g. hamming distance sampling) replaces the words based on uniform distribution. In contrast, we consider both source and target augmentation, while also design evaluation sub-model to pick out sequences with better quality instead of only using LM. Furthermore, we believe our augmentation method samples better sentences before being transmitted to the evaluation sub-model since we use bidirectional LM to replace words which are the position to be selected by learning rather than random selection.

6 Conclusions and Future Work

In this paper, we introduce rather unambiguous and effective DA methods for NMT in LRLs, that sample data from original real data distribution by exploiting the edit distance sampling. We also design the evaluation model to help the augmentation model to reserve high-quality sentences from the generated pseudo data. Our method is model transparent and language-independent. Thus we can incorporate such a sampling method into different MT systems on several languages. Experimental results show that our method significantly outperformed previous approaches in the literature and it has further validated the effectiveness of our model. In the future, apart from the MT tasks, we aim to use this method for other NLP tasks, such as information retrieval, speech recognition, summarization and dialogue generation.

In the future, we will use this method for other text generation tasks, such as information retrieval, speech recognition, summarization, dialogue generation, chat-bot, question answering, and Knowledge graph apart from the machine translation tasks. Besides, we would like to examine the effect of other factors to augmentation, as well as the effectiveness of various hyper-parameters of augmentation sub-model, and we will also investigate the influences of evaluation sub-model with different tricks.

7 Acknowledgements

We would like to thank all anonymous reviewers for their valuable comments and suggestions during both the major and minor revision for this work. This work was supported by the National Key R&D Program of China (No. 2017YFB0202204), National Natural Science Foundation of China (No. 61925601, No. 61772302) and Sogou Inc. Besides, the authors also would like to thank Mike Madden, MBA candidate at Johns Hopkins for patiently proofreading this work before submission.

References

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[2] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[5] Kyunghyun Cho, B. V. Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[6] Y. Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Q. Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, Taku Kudo, H. Kazawa, K. Stevens, G. Kurian, Nishant Patil, W. Wang, C. Young, J. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, G. S. Corrado, Macduff Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144, 2016.

[7] Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL*, 2003.

[8] Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. Is neural machine translation ready for deployment? a case study on 30 Translation directions. arXiv:1610.01108v2, 2016.

[9] Clara Vania and Adam Lopez. From characters to words to in between: Do we capture morphology? In *ACL*, 2017.

[10] Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. A teacher-student framework for zero-resource neural machine translation. In *ACL*, 2017.

[11] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, V. Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019.

[12] Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. Data noising as smoothing in neural network language models. *ArXiv*, abs/1703.02573, 2017.

[13] Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. Conditional bert contextual augmentation. *ArXiv*, abs/1812.06705, 2018.

[14] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *ArXiv*, abs/1710.11041, 2017.

[15] Mohit Iyyer, Varun Manjunatha, Jordan L. Boyd-Graber, and Hal Daumé. Deep unordered composition rivals syntactic methods for text classification. In *ACL*, 2015.

[16] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In *ACL*, 2017.

[17] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *NAACL-HLT*, 2018.

[18] Pei Ke, Fei Huang, Minlie Huang, and Xiaoyan Zhu. Araml: A stable adversarial training framework for text generation. In *EMNLP*, 2019.

[19] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *ArXiv*, abs/1711.00043, 2017.

[20] Rico Sennrich, Barry Haddow, and Alexandra" Birch. Improving neural machine translation models with monolingual data. In *ACL*, 2016.

[21] Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *AAAI*, 2018.

[22] Mohammad Norouzi, Samy Bengio, Zhifeng Chen, Navdeep Jaitly, Mike Schuster, Yonghui Wu, and Dale Schuurmans. Reward augmented maximum likelihood for neural structured prediction. In *NIPS*, 2016.

[23] Jinyue Su, Jiacheng Xu, Xipeng Qiu, and Xuanjing Huang. Incorporating discriminator in sentence generation: a gibbs sampling method. In *AAAI*, 2018.

[24] Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. Switchout: an efficient data augmentation algorithm for neural machine translation. In *EMNLP*, 2018.

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhixiang Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *ICCV*, 2017.

[26] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2016.

[27] Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[28] Matthew Snover, Bonnie J. Dorr, Richard Schwartz, and Linnea Micciulla. A study of translation edit rate with targeted human annotation. 2006.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[30] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[31] Anna Currey and Antonio Valerio Miceli Barone. Copied monolingual data improves low-resource neural machine translation. In *WMT*, 2017.

[32] Jinhua Zhu, Fei Gao, Lijun Wu, Yingce Xia, Tao Qin, Wengang Zhou, Xueqi Cheng, and Tie-Yan Liu. Soft contextual data augmentation for neural machine translation. In *ACL*, 2019.

[33] Mieradilijiang Maimaiti and Maosong Sun Yang Liu, Huanbo Luan. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *TALLIP*, 18, 2019.

[34] Xiangwen Zhang, Jinsong Su, Yue Qin, Y. Liu, R. Ji, and Hongji Wang. Asynchronous bidirectional decoding for neural machine translation. In *AAAI*, 2018.

[35] Jinsong Su, Xuehua Zhang, Qian Lin, Yue Qin, Junfeng Yao, and Yang Liu. Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding. *Artif. Intell.*, 277, 2019.

[36] B. Zhang, Deyi Xiong, John Xie, and Jinsong Su. Neural machine translation with gru-gated attention model. *IEEE transactions on neural networks and learning systems*, 2020.

[37] Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor O. K. Li. Meta-learning for low-resource neural machine translation. In *EMNLP*, 2018.

[38] Rumeng Li, Xun Wang, and Hong Yu. A meta learning method leveraging multiple domain data for low resource machine translation. In *AAAI 2020*, 2020.

[39] Barret Zoph, Deniz Yuret, Jonathan May, and K. Knight. Transfer learning for low-resource neural machine translation. *ArXiv*, abs/1604.02201, 2016.

[40] M. Maimaiti and Xiaohui Zou. Discussion on bilingual cognition in international exchange activities. In *IFIP TC12 ICIS*, 2018.

[41] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of simple domain adaptation methods for neural machine translation. *CoRR*, abs/1701.03214, 2017.

[42] A. Imankulova, Raj Dabre, A. Fujita, and K. Imamura. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. *ArXiv*, abs/1907.03060, 2019.

[43] Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. Joint training for pivot-based neural machine translation. In *IJCAI*, 2017.

[44] L. Sestorain, Massimiliano Ciaramita, C. Buck, and Thomas Hofmann. Zero-shot dual machine translation. *ArXiv*, abs/1805.10338, 2018.

[45] Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In *AAAI*, 2020.

[46] Jiali Zeng, Jinsong Su, H. Wen, Yang Liu, J. Xie, Yongjing Yin, and J. Zhao. Multi-domain neural machine translation with word-level domain context discrimination. In *EMNLP*, 2018.

[47] Jiali Zeng, Y. Liu, Jinsong Su, Yubing Ge, Yaojie Lu, Yongjing Yin, and Jiebo Luo. Iterative dual domain adaptation for neural machine translation. In *EMNLP/IJCNLP*, 2019.

[48] Jinsong Su, Jiali Zeng, John Xie, H. Wen, Yongjing Yin, and Y. Liu. Exploring discriminative word-level domain contexts for multi-domain neural machine translation. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[49] Rico Sennrich and Biao Zhang. Revisiting low-resource neural machine translation: A case study. In *ACL*, 2019.