Alternated Training with Synthetic and Authentic Data for Neural Machine Translation

Rui Jiao¹, Zonghan Yang¹, Maosong Sun¹, Yang Liu∗¹,2,3,4,5

¹Department of Computer Science and Technology, Tsinghua University, Beijing, China
²Institute for AI Industry Research, Tsinghua University, Beijing, China
³Institute for Artificial Intelligence, Tsinghua University, Beijing, China
⁴Beijing National Research Center for Information Science and Technology
⁵Beijing Academy of Artificial Intelligence

Abstract

While synthetic bilingual corpora have demonstrated their effectiveness in low-resource neural machine translation (NMT), adding more synthetic data often deteriorates translation performance. In this work, we propose alternated training with synthetic and authentic data for NMT. The basic idea is to alternate synthetic and authentic corpora iteratively during training. Compared with previous work, we introduce authentic data as guidance to prevent the training of NMT models from being disturbed by noisy synthetic data. Experiments on Chinese-English and German-English translation tasks show that our approach improves the performance over several strong baselines. We visualize the BLEU landscape to further investigate the role of authentic and synthetic data during alternated training. From the visualization, we find that authentic data helps to direct the NMT model parameters towards points with higher BLEU scores and leads to consistent translation performance improvement.

1 Introduction

While recent years have witnessed the rapid development of Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017), it heavily relies on large-scale, high-quality bilingual corpora. Due to the expense and scarcity of authentic corpora, synthetic data has played a significant role in boosting translation quality (He et al., 2016; Sennrich et al., 2016a; Zhang and Zong, 2016; Cheng et al., 2016; Fadaee et al., 2017).

Existing approaches to synthesizing data in NMT focus on leveraging monolingual data in the training process. Among them, back-translation (BT) (Sennrich et al., 2016a) has been widely used to generate synthetic bilingual corpora by using a trained target-to-source NMT model to translate large-scale target-side monolingual corpora. Such synthetic data can be used to improve source-to-target NMT models. Despite the effectiveness of back-translation, the synthetic data inevitably contains noise and erroneous translations. As a matter of fact, it has been widely observed that while BT is capable of benefiting NMT models by using relatively small-scale synthetic data, further increasing the quantity often deteriorates translation performance (Edunov et al., 2018; Wu et al., 2019; Caswell et al., 2019).

This problem has attracted increasing attention in the NMT community (Edunov et al., 2018; Wang et al., 2019). One direction to alleviate the problem is to add noise or a special tag on the source side of synthetic data, which enables NMT models to distinguish between authentic and synthetic data (Edunov et al., 2018; Caswell et al., 2019). Another direction is to filter or evaluate the synthetic data by calculating confidence over corpora, making NMT models better exploit synthetic data (Imamura et al., 2018; Wang et al., 2019). While these methods have outperformed the conventional BT approach, NMT models still suffer from a performance degradation as the size of synthetic data keeps increasing. Hence, how to better take advantage of limited authentic data and abundant synthetic data still remains a grand challenge.

In this work, we propose alternated training with synthetic and authentic data for neural machine translation. The basic idea is to alternate synthetic and authentic corpora iteratively during training. Compared with previous work, we introduce authentic data as guidance to prevent the training of NMT models from being disturbed by noisy synthetic data. Our approach is inspired by the characterization of synthetic and authentic corpora as two types of different approximations for the dis-
tribution of infinite authentic data. We visualize the BLEU landscape to further investigate the role of authentic and synthetic data during alternated training. We find that the authentic data helps to direct NMT model parameters towards the points with higher BLEU scores. Experiments on Chinese-English translation tasks show that our approach improves the performance over strong baselines.

2 Alternated Training

Let \( x \) be a source sentence and \( y \) be a target sentence. We use \( P(y|x; \theta) \) to denote an NMT model parameterized by \( \theta \). Let \( D_a = \{ (x_n, y_n) \}_{n=1}^{N} \) be an authentic parallel corpus containing \( N \) sentence pairs. Traditional NMT aims to obtain \( \hat{\theta}_a \) that maximizes the log-likelihood on \( D_a \):

\[
\hat{\theta}_a = \operatorname{argmax}_\theta \left\{ \frac{1}{N} \sum_{n=1}^{N} \log P(y_n|x_n; \theta) \right\}. \tag{1}
\]

Back-translation generates additional synthetic parallel data from the monolingual corpus. Let \( D_m = \{ y_m \}_{m=1}^{M} \) be a monolingual corpus containing \( M \) target-side sentences. Back-translation first trains a target-to-source model \( \hat{\theta}_{BT} \) on \( D_a \):

\[
\hat{\theta}_{BT} = \operatorname{argmax}_\theta \left\{ \frac{1}{N} \sum_{n=1}^{N} \log P(x_n|y_n; \theta) \right\}, \tag{2}
\]

which is then used to translate each sentence in the target-side monolingual corpus \( D_m \):

\[
\hat{x}_m = \operatorname{argmax}_x \left\{ P(x|y_m; \hat{\theta}_{BT}) \right\}, \tag{3}
\]

where \( m = 1, \ldots, M \). The synthetic corpus \( D_s \) is generated by pairing the translations \( \{ \hat{x}_m \}_{m=1}^{M} \) with \( D_m \), i.e. \( D_s = \{ (\hat{x}_m, y_m) \}_{m=1}^{M} \). The required source-to-target model is finally trained on the combination of authentic and synthetic data:

\[
\hat{\theta}_s = \operatorname{argmax}_\theta \left\{ \frac{1}{N+M} \left( \sum_{n=1}^{N} \log P(y_n|x_n; \theta) + \sum_{m=1}^{M} \log P(y_m|\hat{x}_m; \theta) \right) \right\}. \tag{4}
\]

Suppose that there exists infinite authentic parallel data, which can be characterized as distribution \( p(x, y) \). Synthesizing the large-scale corpus \( D_s \) is to better approach the authentic parallel data distribution. Furthermore, the finite corpora \( D_a \) and \( D_s \cup D_a \) can be viewed as different empirical approximations of \( p(x, y) \):

\[
p_a(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta_{(x_n, y_n) \in D_a}(x, y), \tag{5}
\]

\[
p_s(x, y) = \frac{1}{N+M} \left( \sum_{n=1}^{N} \delta_{(x_n, y_n) \in D_a}(x, y) + \sum_{m=1}^{M} \delta_{(\hat{x}_m, y_m) \in D_s}(x, y) \right), \tag{6}
\]

where \( \delta \) represents the Dirac distribution. On the one hand, \( D_a \) is considered to be of higher quality as \( \lim_{N \to \infty} p_a(x, y) = p(x, y) \) exactly recovers the authentic data distribution. On the other hand, although \( D_s \) contains certain noise (as \( \lim_{M \to \infty} p_s(x, y) \neq p(x, y) \)), it provides more diversified data samples that enable the NMT model to reconstruct the global distribution. As the two corpora are complementary to each other, we introduce authentic data periodically during the training process with synthetic data. Intuitively, alternated training using authentic corpora helps to rectify the deviation of training direction affected by the noisy synthetic data and enhances model performance.

Our proposed alternated training approach is shown in Algorithm 1. Starting with random initialization, each alternation cycle during training consists of two steps. For the \( t \)-th cycle, the first step is to finetune the model \( \hat{\theta}_a^{(t)} \) with Eq. (4) on \( D_s \cup D_a \) until convergence\(^1 \) to obtain \( \hat{\theta}_s^{(t+1)} \), which

---

---

\(^1\)We also attempted to train S/A-steps for certain iterations. Empirically, the proposed convergence-based method performed better.
is referred as S-Step (line 4). The second step is to alter the training data back to $D_a$ and finetune $\hat{\theta}^{(t+1)}_a$ with Eq. (1) until convergence to obtain $\hat{\theta}^{(t+1)}_a$, which is referred as A-Step (line 5). We alternate the training process until convergence. It is noted that back-translation is equivalent to a single S-Step performed in our approach.

3 Experiments

3.1 Setup

We evaluated our training strategy on Chinese-English and German-English translation tasks. We reported the tokenized BLEU score as calculated by multi-bleu.perl.

For the Chinese-English task, we extracted 1.25M parallel sentence pairs from LDC as our authentic bilingual corpus and 10M English-side sentences from WMT17 Chinese-English training set as our monolingual corpus for back-translation. NIST06 was used as the validation set. We use NIST02, 03, 04, 05 and 08 datasets as test sets. For the German-English task, we selected the dataset of IWSLT14 German-English task, which contains 16k parallel sentence pairs for training. We further extracted 4.5M English-side sentences from WMT14 German-English training set as monolingual dataset. We segmented Chinese sentences by THULAC (Sun et al., 2016) and tokenized English and German sentences by Moses (Koehn et al., 2007). The vocabulary was built by Byte Pair Encoding (BPE) (Sennrich et al., 2016b) with 32k merge operations. We used Transformer (Vaswani et al., 2017) implemented in THUMT (Tan et al., 2020) with standard hyperparameters as a base model. We used Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ with the maximum learning rate $= 7 \times 10^{-4}$.

We applied early-stopping to verify convergence of each single S/A-step. If the validation BLEU failed to exceed the highest score during the certain S/A-step after 10K training iterations, we consider the model converged and alternated the training set. For the whole training process, we set the maximum training iterations as 250k for Chinese-English task and 150k for German-English task.

3.2 Results

Figure 1 shows the comparison among several approaches in different scales of training sets on the Chinese-English task. The leftmost point is trained on the authentic data, and other points are trained on the combination of authentic and synthetic corpora. The X-axis shows the synthetic data scale ranging from 1.25M (the size of authentic data) to 10M (the full size of the monolingual corpus). The Y-axis shows the BLEU scores of the combined test set. We find that the performance of BT rises firstly but then decreases as more synthetic data is added, which confirms the findings of Wu et al. (2019). In contrast, our approach achieves consistent improvement with the enlargement of the synthetic data scale.

Table 1 shows the detailed translation performance on the Chinese-English task when the synthetic data scale is set to 10M. It can be seen that our alternated training strategy outperforms conventional back-translation and tagged back-translation on all test sets. We find that during training, the S-Steps account for about 73% of the total training time, and the A-Steps account for 27%. This finding suggests that our training procedure composes mainly of S-Steps, and moderate A-Steps are efficient to guide the NMT model towards better points, which lead to the improvement of BLEU performance.

Table 2 shows the results of the German-English task. Similar to the Chinese-English task, we vary the synthetic data scale from 1M to 4.5M for experiments. We find that the performance degradation also occurs while utilizing large-scale synthetic data, and alternated training approach alleviate the problem and perform better than corresponding baselines.

![Figure 1: Comparison with several baselines in different data scale. Our alternated approach outperforms the conventional back-translation method and improves the performance of Tagged BT. Moreover, with the enlargement of the synthetic data scale, the BLEU score rises steadily by alternated training.](image-url)
Table 1: BLEU scores on the NIST Chinese-English task with 10M additional synthetic corpus. “Base” means only authentic data is used. “BT” corresponds to the back-translation method (Sennrich et al., 2016a). “BT-tagged” corresponds to the tagged BT technique proposed by Caswell et al. (2019). “AlterBT” means alternated training on authentic data and synthetic data using “BT” in each alternation. “AlterBT-tagged” means alternated training on authentic data and synthetic data using “BT-tagged” in each alternation. “+” means significantly better than BT (p < 0.01). “†” means significantly better than BT-tagged (p < 0.01).

<table>
<thead>
<tr>
<th>Data</th>
<th>NIST06</th>
<th>NIST02</th>
<th>NIST03</th>
<th>NIST04</th>
<th>NIST05</th>
<th>NIST08</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>45.94</td>
<td>45.82</td>
<td>45.35</td>
<td>46.88</td>
<td>45.43</td>
<td>36.98</td>
<td>44.40</td>
</tr>
<tr>
<td>BT</td>
<td>43.89</td>
<td>44.79</td>
<td>44.40</td>
<td>46.24</td>
<td>45.45</td>
<td>36.45</td>
<td>43.57</td>
</tr>
<tr>
<td>BT-tagged</td>
<td>46.79</td>
<td>47.11</td>
<td>46.49</td>
<td>47.73</td>
<td>47.17</td>
<td>38.41</td>
<td>45.47</td>
</tr>
<tr>
<td>AlterBT</td>
<td>49.07†</td>
<td>48.77†</td>
<td>48.36†</td>
<td>49.51†</td>
<td>49.94†</td>
<td>40.95†</td>
<td>47.68†</td>
</tr>
<tr>
<td>AlterBT-tagged</td>
<td>49.40†</td>
<td>49.04†</td>
<td>48.37†</td>
<td>49.10†</td>
<td>49.64†</td>
<td>40.56†</td>
<td>47.49†</td>
</tr>
</tbody>
</table>

Table 2: BLEU scores on the IWSLT14 German-English task with 1M and 4.5M additional synthetic corpus. “+” means significantly better than BT (p < 0.01). “†” means significantly better than BT-tagged (p < 0.01).

<table>
<thead>
<tr>
<th>Scale</th>
<th>1M</th>
<th>4.5M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>34.16</td>
<td>34.16</td>
</tr>
<tr>
<td>BT</td>
<td>37.36</td>
<td>36.30</td>
</tr>
<tr>
<td>BT-tagged</td>
<td>37.65</td>
<td>37.42</td>
</tr>
<tr>
<td>AlterBT</td>
<td>38.20†</td>
<td>38.53†</td>
</tr>
<tr>
<td>AlterBT-tagged</td>
<td>37.98†</td>
<td>39.19†</td>
</tr>
</tbody>
</table>

3.3 BLEU Landscape Visualization

To validate the assumption that the authentic data helps to rectify the deviation in synthetic data and redirect the NMT model parameters to a better optimization path, we further investigate the BLEU landscape to compare our method with the BT approach during the same training steps.

The visualization of the BLEU landscape is shown in Figure 2. Checkpoints during alternated training are projected onto the 2D plane defined by \( \hat{\theta}_s^{(t)} \), \( \hat{\theta}_a^{(t)} \) and \( \hat{\theta}_s^{(t+1)} \). Our projection method considers both the model parameters and their translation performance (See Appendix A for details). For the conventional BT approach, the model parameters are stuck in an inefficient optimization path (highlighted in blue dashed lines). In our approach, we find that authentic data effectively guides the model towards a better direction for A-Step (highlighted in red solid lines). For S-Step (highlighted in red dashed lines), although training with synthetic data deteriorates the BLEU performance, it pushes the model away from the original route, and enables authentic data to further redirect the model into a better point with a higher BLEU score.

4 Related Work

Our work is based on back-translation (BT), an approach to leverage monolingual data by an additional target-to-source system. BT was proved to be effective in neural machine translation (NMT) systems (Sennrich et al., 2016a). Despite its effectiveness, BT is limited by the accuracy of synthetic data. Noise and translation errors hinder the boosting of model performance (Hoang et al., 2018). The
negative results become more evident when more synthetic data is mixed into training data (Caswell et al., 2019; Wu et al., 2019).

Considerable studies have focused on the accuracy problem in synthetic data and further extended back-translation. Imamura et al. (2018) demonstrate that generating source sentences via sampling increases the diversity of synthetic data and benefits the BT system. Edunov et al. (2018) further propose a noisy beam search method to generate more diversified source-side data. Caswell et al. (2019) add a reserved token to synthetic source-side sentences in order to help NMT model distinguish between authentic and synthetic data. Another perspective aims at measuring the translation quality of synthetic data. Imamura et al. (2018) filter sentence pairs with low likelihood or low confidence. Wang et al. (2019) use uncertainty-based confidence to measure words and sentences in synthetic data. Different from the aforementioned works, our approach introduces neither data modification (e.g. noising or tagging) nor additional models for evaluation. We alternate training set on the original authentic and synthetic data.

The work relatively close to ours is Iterative Back-Translation (Hoang et al., 2018), which refines forward and backward model via back-translation data, and regenerates more accurate synthetic data from monolingual data. Our work differs from Iterative BT in that we do not require source-side monolingual corpora or repeatedly fine-tune the backward model.

5 Conclusion

In this work, we propose alternated training with synthetic and authentic data for neural machine translation. Experiments have shown the supremacy of our approach. Visualization of the BLEU landscape indicates that alternated training guides the NMT model towards better points.

Acknowledgments

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). We thank all anonymous reviewers for their valuable comments and suggestions on this work.

References


A Method for Visualization

We first define the projection plane $S$ by parameters $\hat{\theta}_s^{(l)}$, $\hat{\theta}_a^{(l)}$ and $\hat{\theta}_s^{(l+1)}$. Selecting $\hat{\theta}^* = \hat{\theta}_s^{(l)}$ as the basic point and $\delta = \hat{\theta}_a^{(l)} - \hat{\theta}_s^{(l)}$, $\eta = \hat{\theta}_s^{(l+1)} - \hat{\theta}_s^{(l)}$ as two basis vectors, we plot the function $f(x, y) = \text{BLEU}(D_{DE\text{V}}; \hat{\theta}^* + x\delta + y\eta)$ in the 2D surface. We calculate the BLEU scores for all NMT models defined by grid points on the projection plane, and construct the BLEU contours via linear interpolation in MATPLOTLIB (Hunter, 2007).

We project the model checkpoints onto the 2D plane $S$ to represent the parameter geometry and their translation performance. As the 2D contour plane consists of several regions corresponded with different BLEU ranges, we formulate the visualization task into the following problem:

$$\begin{equation}
(x_i, y_i) = \underset{(x, y) \in S_i}{\text{argmin}} \left\| \hat{\theta}_i - \left( \hat{\theta}^* + x\delta + y\eta \right) \right\|_F^2,
\end{equation}$$

where $S_i$ denotes the BLEU region that the performance of $\hat{\theta}_i$ lies in. It is noted that according to the Pythagorean theorem, for $\forall (\hat{x}, \hat{y}) \in S_i$,

$$\begin{equation}
\left\| \hat{\theta}_i - \left( \hat{\theta}^* + \hat{x}\delta + \hat{y}\eta \right) \right\|_F^2 = \left\| \hat{\theta}_i - \left( \hat{\theta}^* + \hat{x}\delta + \hat{y}\eta \right) \right\|_F^2 + \left\| \left( \hat{\theta}^* + \hat{x}\delta + \hat{y}\eta \right) - \left( \hat{\theta}^* + x\delta + y\eta \right) \right\|_F^2,
\end{equation}$$

where

$$\begin{equation}
(\hat{x}_i, \hat{y}_i) = \underset{(x, y) \in S}{\text{argmin}} \left\| \hat{\theta}_i - \left( \hat{\theta}^* + x\delta + y\eta \right) \right\|_F^2.
\end{equation}$$

As $(x_i, y_i) \in S_i$, we can substitute $(\hat{x}, \hat{y})$ in Eq. (8) with $(x_i, y_i)$. Notice that the first term on the right-hand side of Eq. (8) is independent of $(x_i, y_i)$, the minimizer $(x_i, y_i)$ thus satisfies the following conditions:

$$\begin{equation}
\begin{aligned}
(x_i, y_i) &= \underset{(\hat{x}, \hat{y}) \in S_i}{\text{argmin}} \left\| \left( \hat{\theta}^* + \hat{x}\delta + \hat{y}\eta \right) - \left( \hat{\theta}^* + \hat{x}\delta + \hat{y}\eta \right) \right\|_F^2,
\end{aligned}
\end{equation}$$

with $(\hat{x}_i, \hat{y}_i)$ satisfying Eq. (9).

According to Eq. (8), our projection method can be divided into two steps. The first step is to calculate $(\hat{x}_i, \hat{y}_i)$ in Eq. (9), which minimizes the first term of Eq. (8). By the least square method, we obtain the analytic solution to $(\hat{x}_i, \hat{y}_i)$ as follows:

$$\begin{equation}
\begin{cases}
x_i = \frac{V_B - UC}{B^2 - AC},
y_i = \frac{UB - VA}{B^2 - AC},
\end{cases}
\end{equation}$$

where

$$\begin{align}
A &= \langle \delta, \delta \rangle, \\
B &= \langle \delta, \eta \rangle, \\
C &= \langle \eta, \eta \rangle, \\
U &= \langle \hat{\theta}_i - \hat{\theta}^*, \delta \rangle, \\
V &= \langle \hat{\theta}_i - \hat{\theta}^*, \eta \rangle.
\end{align}$$

The second step is to solve $(x_i, y_i)$ in Eq. (10), which minimizes the second term of Eq. (8). Specialy, we have $(x_i, y_i) = (\hat{x}_i, \hat{y}_i)$ if $(\hat{x}_i, \hat{y}_i) \in S_i$. Otherwise, as the BLEU region $S_i$ is enclosed by polygon boundaries with limited edges, we simply calculate the distance between $(\hat{x}_i, \hat{y}_i)$ and each edge and select the minimum one. The boundary point minimizing the distance is then determined as $(x_i, y_i)$. We cast the projection point from $(\hat{x}_i, \hat{y}_i)$ to $(x_i, y_i)$ in order to restore the origin BLEU performance of $\hat{\theta}_i$. [1]