공학석사 학위논문

Building a Neural Machine Translation System Using Only Synthetic Parallel Data

합성 병렬데이터를 활용한 인공신경망 기계번역 시스템 구축

2017 년 8 월

서울대학교 대학원

전기 · 정보공학부

박재홍

Abstract

Building a Neural Machine Translation System Using Only Synthetic Parallel Data

Jaehong Park
Electrical and Computer Engineering
The Graduate School
Seoul National University

Recent works have shown that synthetic parallel data automatically generated by translation models can be effective for various neural machine translation (NMT) issues. In this study, we build NMT systems using only synthetic parallel data. We also present a novel synthetic parallel corpus as an efficient alternative to real parallel data. The proposed pseudo parallel data are distinct from those of previous works in that ground truth and synthetic examples are mixed on both sides of sentence pairs. Experiments on Czech–German and French–German translations demonstrate the efficacy of the proposed pseudo parallel corpus in empirical NMT applications, which not only shows enhanced results for bidirectional translation tasks, but also substantial improvement with the aid of a ground truth parallel corpus.

Keywords : neural machine translation, synthetic parallel data, pseudo parallel data, phrase-based statistical machine translation
Student Number : 2015-20929

Table of Contents

List of Tables

List of Figures

I. Introduction

Neural machine translation (NMT) employing the encoder-decoder architecture [1, 2, 3] has shown promising results in recent years. Combined with the attention mechanism, NMT has reported state-of-the-art translation quality for several language pairs [4, 5, 6].

Given the data-driven nature of NMT, the limited number of source-to-target bilingual sentence pairs have been one of the major obstacles in building competitive NMT systems. Recently, pseudo parallel data, which refer to the synthetic bilingual sentence pairs automatically generated by existing translation models, have reported promising results regarding the data scarcity in NMT. Many studies have found that combining pseudo parallel data with a real bilingual parallel corpus significantly enhances the quality of NMT models [5, 7, 8]. In addition, synthesized parallel data have played a vital role in resolving many NMT issues, such as domain adaptation [5], zero-resource NMT [9], and the rare word problem [10].

Inspired by their efficacy, we attempt to build NMT models using only synthetic parallel data. To the best of our knowledge, building NMT systems with only synthetic data has yet to be studied. Through our research, we explore the availability of pseudo parallel data as an efficient alternative to the real-world parallel corpus. The active usage of synthetic parallel data in NMT has particular significance in low-resource language pairs where real parallel data are very limited or not established. Even in recent approaches, such as zero-shot

NMT [11] and the pivot-based method [12] where direct source-to-target bilingual data are not required, the direct parallel corpus substantially improves translation quality where pseudo parallel data can also be employed.

Existing synthetic parallel data, however, have several drawbacks as a reliable alternative to the real-world parallel corpus. One weakness is that sentences from the real corpus only exist on a single side of pseudo sentence pairs while the other side is composed only of synthetic sentences. For instance, given a translation task, existing synthetic parallel corpora can be classified into two groups: source-originated and target-originated. As illustrated in Figure 1, each of the source- and target-originated synthetic parallel data is constructed by automatically translating a source-side or target-side monolingual corpus. The bias of synthetic examples in sentence pairs, however, may lead an imbalance in the quality of learned NMT models when the given pseudo parallel corpus is applied to bidirectional translation tasks (e.g. French → German and German → French). In addition, the reliability of the synthetic parallel data is heavily influenced by a single translation model where the synthetic examples originate. Low-quality synthetic sentences generated by the model would prevent NMT models from learning solid parameters during the training process.

To overcome these shortcomings, we propose a new type of synthetic parallel corpus called $\text{PSEUDO}_{\text{mix}}$. In contrast to previous approaches, $\text{PSEUDO}_{\text{mix}}$ includes both synthetic and real sentences on either side of training sentence pairs. In practice, it can be readily

built by mixing source- and target-originated pseudo parallel corpora for a given translation task. Experiments on several language pairs show that the proposed $\text{PSEUDO}_{\text{mix}}$ has useful properties that make it a reliable candidate for real-world parallel data. Specifically, we make the following contributions:

i) Our work provides a thorough investigation on exploiting synthetic parallel data in low-resource NMT scenarios.

ii) The proposed synthetic parallel data $\text{PSEUDO}_{\text{mix}}$ shows enhanced translation quality compared to existing source- and target-originated pseudo parallel corpora in bidirectional translation tasks.

iii) When fine-tuned using ground truth parallel data, a model trained with $\text{PSEUDO}_{\text{mix}}$ outperforms other fine-tuned models trained with source-originated and target-originated synthetic parallel data, indicating substantial improvement in translation quality.

II. Background:

Neural Machine Translation

Given a source sentence $x = (x_1, \cdots, x_{T_x})$ and its corresponding target sentence $y = (y_1, \cdots, y_{T_y})$, the NMT aims to model the conditional probability $p(y|x)$ with a single large neural network. To parameterize the conditional distribution, recent studies on NMT employ the encoder–decoder framework [1, 2, 3]. Thereafter, the attention mechanism [13, 14] has been introduced to address the performance drop that occurs in long source sentences [15].

In this study, we use the attentional NMT architecture proposed by Bahdanau et al. [13]. In their work, the encoder reads the source sentence one symbol at a time and generates a sequence of source representations $h = (h_1, \cdots, h_{T_x})$. Specifically, the encoder is a bidirectional recurrent neural network (BiRNN) with gated recurrent units (GRU) [2]. The BiRNN consists of two recurrent neural networks (RNN): forward RNN $\vec{f}$ and backward RNN $\overleftarrow{f}$. The forward RNN $\vec{f}$ reads the source sentence sequentially from the first element x_1 to the last element x_{T_x}, computing a sequence of forward hidden states $(\vec{h}_1, \cdots, \vec{h}_{T_x})$. The computation of the forward hidden state $\vec{h}_i$ is given as follows:

$$\vec{h}_i = \begin{cases} (1 - \vec{z}_i) \circ \vec{h}_{i-1} + \vec{z}_i \circ \underline{\vec{h}_i} & , \quad if \;\; i > 0 \\ 0 & , \quad if \;\; i = 0 \end{cases}$$

where

$$\underline{\vec{h}_i} = \tanh\left(\vec{W}\bar{E}x_i + \vec{U}\left[\vec{r}_i \circ \vec{h}_{i-1}\right]\right)$$

$$\vec{z}_i = \sigma(\vec{W}_z\bar{E}x_i + \vec{U}_z\vec{h}_{i-1})$$

$$\vec{r}_i = \sigma(\vec{W}_r\bar{E}x_i + \vec{U}_r\vec{h}_{i-1})$$

Each source symbol x_i is denoted as a $1-of-K$ coded vector, i.e., $x_i \in \mathbb{R}^{K_x}$ where K_x is the vocabulary size of the source language. $\bar{E} \in \mathbb{R}^{m \times K_x}$ is the embedding matrix for the source language where m is the embedding dimensionality. $\vec{W}, \vec{W}_z, \vec{W}_r \in \mathbb{R}^{n \times m}$, $\vec{U}, \vec{U}_z, \vec{U}_r \in \mathbb{R}^{n \times n}$ are learnable encoder weight matrices where n is the number of hidden units. The backward RNN reads the source sentence in the reverse order (from x_{T_x} to x_1) and calculates a sequence of backward hidden states $(\overleftarrow{h}_1, \cdots, \overleftarrow{h}_{T_x})$ in the same manner. The final source representation h_i for the source element x_i is obtained by concatenating the forward hidden state $\vec{h}_i$ and the backward hidden state $\overleftarrow{h}_i$.

$$h_i = \begin{bmatrix} \vec{h}_i \\ \overleftarrow{h}_i \end{bmatrix}$$

Through the concatenation of the forward and backward hidden states, the source representation h_i effectively summarizes the information surrounding the source element x_i.

The decoder, which is another recurrent neural network with GRU, predicts the target sentence one symbol at a time. The hidden state s_t of the decoder RNN is computed as follows:

$$s_t = \begin{cases} (1 - z_t) \circ s_{t-1} + z_t \circ \bar{s}_t, & if \ t > 0 \\ \tanh(W_s \overleftarrow{h}_1) & , \quad if \ t = 0 \end{cases}$$

where

$$\bar{s}_t = \tanh(WEy_{t-1} + U[r_t \circ s_{t-1}] + Cc_t)$$
$$z_t = \sigma(W_z Ey_{t-1} + U_z s_{t-1} + C_z c_t)$$
$$r_t = \sigma(W_r Ey_{t-1} + U_r s_{t-1} + C_r c_t)$$

Each target symbol y_t is also denoted as a $1-of-K$ coded vector, i.e., $y_t \in \mathbb{R}^{K_y}$ where K_y is the vocabulary size of the target language. $E \in \mathbb{R}^{m \times K_y}$ is the embedding matrix for the target language where m is again the embedding dimensionality. $W, W_z, W_r \in \mathbb{R}^{n \times m}$, $U, U_z, U_r \in \mathbb{R}^{n \times n}$ and $C, C_z, C_r \in \mathbb{R}^{n \times 2n}$ are learnable decoder weight matrices where n is again the number of hidden units. The weight matrix $W_s \in \mathbb{R}^{n \times n}$ is also learnable and used to convert the last backward encoder hidden state $\overleftarrow{h}_1$ into the decoder initial hidden state s_0. The context vector c_t is used to determine the relevant part of the source sentence to predict y_t. It is computed as the weighted average of source representations $h_1, \cdots, h_m$. Each weight α_{ti} for h_i implies the probability of the target symbol y_t being aligned to the source symbol x_i:

$$c_t = \sum_{i=1}^{T_x} \alpha_{ti} h_i$$

where

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{T_x} \exp(e_{tk})}$$

$$e_{ti} = \ v_a^{\mathrm{T}} \tanh(W_a s_{t-1} + U_a h_i)$$

$v_a \in \mathbb{R}^{n'}$, $W_a \in \mathbb{R}^{n' \times n}$ and $U_a \in \mathbb{R}^{n' \times 2n}$ are learnable attention weight matrices. Note that the alignment model used to compute e_{ti} is essentially a feedforward neural network with a single hidden layer. In the attentional NMT architecture, the conditional distribution of the target symbol y_t is modeled as a function of the previously predicted output y_{t-1}, the hidden state of the decoder network s_t, and the context vector c_t.

$$p(y_t | y_{<t}, x) \propto \exp\{g(y_{t-1}, s_t, c_t)\}$$

where $y_{<t} = (y_1, \cdots, y_{t-1})$. In detail, the function g is defined as

$$p(y_t | y_{<t}, x) = p(y_t | s_t, y_{t-1}, c_t) \propto \exp\{y_t^{\mathrm{T}} W_o \beta_t\}$$

where $W_o \in \mathbb{R}^{K_y \times l}$ is a learnable projection matrix. β_t is computed by

$$\beta_t = \left[\max\left\{\overline{\beta}_{t,2j-1}, \overline{\beta}_{t,2j}\right\}\right]_{j=1,\cdots,l}^{\mathrm{T}}$$

where $\overline{\beta}_{t,k}$ is the $k-$th element of a vector $\overline{\beta}_t$, which is computed as follows:

$$\overline{\beta}_t = U_o s_i + V_o E y_{i-1} + C_o c_i$$

$U_o \in \mathbb{R}^{2l \times n}$, $V_o \in \mathbb{R}^{2l \times m}$ and $C_o \in \mathbb{R}^{2l \times 2n}$ are learnable decoding matrices.

To train a NMT model, note that log conditional probability of the target sentence y given the source sentence x can be decomposed as follows:

$$\log p(y|x) = \sum_{t=1}^{T_y} \log p(y_t|y_{<t}, x)$$

Given a sentence-aligned parallel corpus of size N, the entire parameter θ of the NMT model is jointly trained to maximize the conditional probabilities of all sentence pairs $\{(x^n, y^n)\}_{n=1}^N$:

$$\theta^* = \arg\max_{\theta} \sum_{n=1}^{N} \log p(y^n|x^n)$$

where θ^* is the optimal parameter.

III. Related Work

In statistical machine translation (SMT), synthetic bilingual data have been primarily proposed as a means of exploiting monolingual corpora. By applying a self-training scheme, the pseudo parallel data was obtained by automatically translating the source-side monolingual corpora [16, 17]. In a similar but reverse way, the target-side monolingual corpora were also employed to build the synthetic parallel data [18, 19]. The primary goal of these works was to adapt trained SMT models to other domain using relatively abundant in-domain monolingual data.

Inspired by the successful application in SMT, there have been many attempts to exploit synthetic parallel data to improve NMT systems. Source-side [7], target-side [5], and both sides [8] of the monolingual data have been used to build synthetic parallel corpora. In their work, the pseudo parallel data combined with a real training corpus significantly enhanced the translation quality of NMT. In Sennrich et al. [5], domain adaptation of NMT was achieved by fine-tuning trained NMT models using a synthetic parallel corpus. Firat et al. [9] attempted to build NMT systems without any direct source-to-target parallel corpus. In their work, the pseudo parallel corpus was employed in fine-tuning the target-specific attention mechanism of trained multi-way multilingual NMT [20] models, which enabled zero-resource NMT between the source and target

languages. Lastly, in Zhang and Zong [10], synthetic bilingual sentence pairs were generated to enrich training examples including rare or unknown translation lexicons.

IV. Synthetic Parallel Data as an Alternative to Real Parallel Corpus

4.1. Motivation

As described in the previous section, synthetic parallel data have been widely used to boost the translation quality of NMT. In this work, we further extend their application by training NMT models with only synthetic parallel data. In certain language pairs or domains where the source-to-target real parallel data are very rare or even unprepared, a model trained with synthetic parallel data can function as an effective baseline model. Once the additional ground truth parallel corpus is established, the trained model can be improved by retraining or fine-tuning using the ground truth parallel data.

4.2. Limits of the Previous Approaches

For a given translation task, we classify the existing synthetic parallel corpora into the following groups based on the composition of sentence pairs:

i) Source-originated: The source sentences are from real data,

and the associated target sentences are synthetic. The corpus can be formed by automatically translating a source-side monolingual corpus into the target language [7, 10]. It can also be built from source-pivot bilingual data by introducing a pivot language. In this case, a pivot-to-target translation model is employed to translate the pivot language corpus into the target language. The generated target sentences paired with the original source sentences form a pseudo parallel corpus.

ii) Target-originated: The target sentences are from a real corpus, and the associated source sentences are synthetic. The corpus can be formed by back-translating a target-side monolingual corpus into the source language [5]. Like the source-originated case, it can also be built from a pivot-target bilingual corpus using a pivot-to-source translation model [9].

Figure 1 illustrates the overall process of building each synthetic parallel corpus. As shown in Figure 1, previous approaches to pseudo parallel data share a common property: synthetic and non-synthetic real sentences are biased to a single side of sentence pairs. Given that synthetic parallel data have been exploited only as a supplementary resource, the bias of the synthetic examples in the pseudo sentence pairs has never been seriously discussed. In such a case where the synthetic parallel data is the only or major resource to build NMT systems, this can severely limit the availability of the

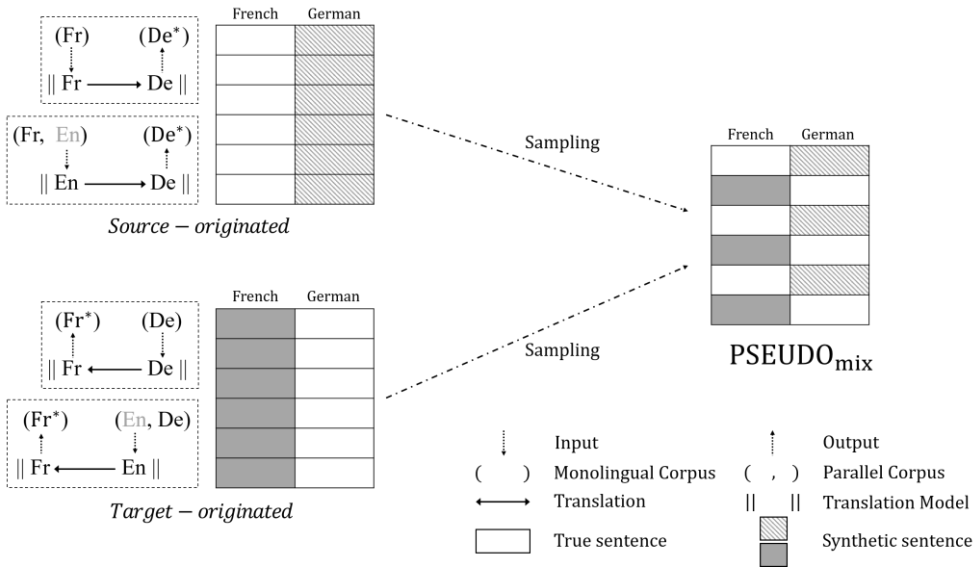

Figure 1. The process of building each pseudo parallel corpus group for French → German translation. * indicates the synthetic sentences generated by translation models. Each of the source−originated and the target−originated synthetic parallel data can be made from French or German monolingual corpora. They can also be built from parallel corpora including English, which is the pivot language.

given pseudo parallel corpus. For instance, as will be presented in our experiments, pseudo parallel data showing relatively high performance in one translation task (e.g. French → German) can produce poor results in the translation task of the reverse direction (German → French).

Another drawback of employing synthetic parallel data in training NMT is that the capacity of the synthetic parallel corpus is inherently influenced by the mother translation model where the synthetic sentences originate. Depending on the quality of the mother model, ill−fored or inaccurate synthetic examples can be generated, which negatively affect the reliability of the resulting synthetic parallel data. In a previous study, Zhang and Zong [7] bypassed this issue by

freezing the decoder parameters while training with the minibatches of pseudo bilingual pairs made from a source language monolingual corpus. This scheme, however, cannot be applied to our scenario as the decoder network will remain untrained during the entire training process.

4.3. Proposed Mixing Approach

To overcome the limitations of the previously suggested pseudo parallel data, we propose a novel synthetic parallel corpus called $\text{PSEUDO}_{\text{mix}}$. Our approach is quite straightforward: for a given translation task, we first build both source- and target-originated pseudo parallel data using the method described in the previous section. $\text{PSEUDO}_{\text{mix}}$ can then be readily built by mixing them together. Figure 1 illustrates the overall process of building $\text{PSEUDO}_{\text{mix}}$ for the French → German translation task. We also present samples of ground truth and synthetic parallel corpora in Table 1.

By mixing source- and target-originated pseudo data, the resulting corpus includes both real and synthetic examples on each side of sentence pairs, which is the most evident feature of $\text{PSEUDO}_{\text{mix}}$. Through this mixing approach, we attempt to lower the discrepancy in the quality of the source and target examples of synthetic sentence pairs, thus enhancing its reliability as a parallel resource. In the following section, we evaluate the actual benefits of a mixed composition in the synthetic parallel data.

Fr	#1. Je crois, pour répondre à Mme Kinnock, que le sentiment sur le rythme et les perspectives de ces accords varie selon les régions auxquelles on s'adresse. #2. Le Parlement se prononce en faveur d'un renforcement de la politique étrangère et de sécurité, tout en vidant le droit de veto de son contenu.
De	#1. Ich glaube, um auf Frau Kinnocks Frage zu antworten, dass es von den jeweiligen Regionen abhängt, wie die Geschwindigkeit der Verhandlungsaufnahme und die Perspektiven der Abkommen wahrgenommen werden. #2. Das Parlament spricht sich für den Ausbau der Außen- und Sicherheitspolitik und für die gleichzeitige Aushöhlung des Vetorechts aus.

(a) Ground truth parallel data

Fr	#1. Je crois, pour répondre à Mme Kinnock, que le sentiment sur le rythme et les perspectives de ces accords varie selon les régions auxquelles on s'adresse. #2. Le Parlement se prononce en faveur d'un renforcement de la politique étrangère et de sécurité, tout en vidant le droit de veto de son contenu.
De	*#1. Als Antwort auf Frau Kinnock bin ich der Meinung, daß die Ansichten über den Rhythmus und die Perspektiven dieser Abkommen je nach Region variieren.* *#2. Dieses Haus fordert eine starke gemeinsame Außenpolitik mit einem hohlen Vetorecht.*

(b) Source-originated synthetic parallel data

Fr	*#1. En réponse à Mme Kinnock, je pense que les sentiments sur le rythme et les perspectives de ces accords varient selon les régions que nous parlons.* *#2. Cette Assemblée demande une ferme politique étrangère commune dotée d'un droit de veto creux.*
De	**#1. Ich glaube, um auf Frau Kinnocks Frage zu antworten, dass es von den jeweiligen Regionen abhängt, wie die Geschwindigkeit der Verhandlungsaufnahme und die Perspektiven der Abkommen wahrgenommen werden.** **#2. Das Parlament spricht sich für den Ausbau der Außen- und Sicherheitspolitik und für die gleichzeitige Aushöhlung des Vetorechts aus.**

(c) Target-originated synthetic parallel data

Fr	*#1. En réponse à Mme Kinnock, je pense que les sentiments sur le rythme et les perspectives de ces accords varient selon les régions que nous parlons.* **#2. Le Parlement se prononce en faveur d'un renforcement de la politique étrangère et de sécurité, tout en vidant le droit de veto de son contenu.**
De	**#1. Ich glaube, um auf Frau Kinnocks Frage zu antworten, dass es von den jeweiligen Regionen abhängt, wie die Geschwindigkeit der Verhandlungsaufnahme und die Perspektiven der Abkommen wahrgenommen werden.** *#2. Dieses Haus fordert eine starke gemeinsame Außenpolitik mit einem hohlen Vetorecht.*

(d) $\text{PSEUDO}_{\text{mix}}$

Table 1. Examples of ground truth and synthetic parallel corpora for French → German translation task. Sentences from real corpora are bold-faced and synthetic sentences generated from translation models are italic formatted. Note that $\text{PSEUDO}_{\text{mix}}$ contains both real and synthetic examples on either side of sentence pairs.

V. Experiments: Effects of Mixing Real and Synthetic Examples

In this section, we analyze the effects of a mixed composition in the synthetic parallel data. Mixing pseudo parallel corpora derived from different sources, however, inevitably brings diversity, which affects the capacity of the resulting corpus. We isolate this factor by building both source- and target-originated synthetic parallel corpora from the identical source-to-target ground truth real parallel corpus. Our experiments are performed on French (Fr) ↔ German (De) translation tasks. The choice of the Fr - De language pair reflects our motivation to assume low-resource environments in NMT. While many public benchmark parallel resources are concentrated on language pairs including English, the size of publicly released parallel corpora for the Fr - De language pair is relatively restricted (~1.8M). Throughout the remaining paper, we use the notation * to denote the synthetic part of the pseudo sentence pairs.

Corpus	Size	Average length	
		Fr	De
Europarl Fr−En−De	1.78M	26.00	23.16
Fr−De*	1.45M	25.56	22.98
Fr*−De	1.45M	25.32	23.46
$PSEUDO_{mix}$	1.45M	25.47	23.26

Table 2. Statistics of the parallel corpora for Fr ↔ De translation tasks. The notation * denotes the synthetic part of the parallel corpus.

5.1. Data Preparation

By choosing English (En) as the pivot language, we perform pivot alignments for identical English segments on Europarl Fr−En and En−De parallel corpora [21], thus constructing a multi−parallel corpus of Fr−En−De. Then each of the Fr*−De and Fr−De* pseudo parallel corpus is established from the multi−parallel data by applying the pivot language−based translation described in the previous section. For automatic translation, we utilize a pre−trained and publicly released NMT model① for En → De and train another NMT model for En → Fr using the WMT'15 En−Fr parallel corpus [22]. A beam of size 5 is used to generate synthetic sentences. Lastly, to match the size of the training data, $PSEUDO_{mix}$ is established by randomly sampling half of each Fr*−De and Fr−De* and mixing them together.

① http://data.statmt.org/rsennrich/wmt16_systems

5.2. Data Preprocessing

Each training corpus is tokenized using the tokenization script in Moses [23]. We represent every sentence as a sequence of subword units learned from byte-pair encoding [24]. We remove empty lines and all sentences with a length of over 50 subword units. For a fair comparison, all cleaned synthetic parallel data have equal sizes. Table 2 presents a summary of the final parallel corpora.

5.3. Training and Evaluation

All networks have 1024 hidden units and 500 dimensional embeddings. Vocabulary size is limited to 30K for each language. Each model is trained for 10 epochs using stochastic gradient descent with Adam [25]. The Minibatch size is 80, and the training set is reshuffled between every epoch. The norm of the gradient is clipped not to exceed 1.0 [26]. The learning rate is 2e-4 in every case.

We use the newstest 2012 set for a development set and the newstest 2011 and newstest 2013 sets as test sets. At test time, the beam search is used to approximately find the most likely translation $\hat{y}$ given a source sentence x.

$$\hat{y} = \arg\max_{y} p(y|x)$$

We use a beam of size 12 and normalize probabilities by the length of the candidate sentences. The evaluation metric is case-sensitive tokenized BLEU [27] computed with the multi-bleu script from Moses. For each case, we present average BLEU evaluated on three different models trained with the same synthetic corpus.

5.4. Results and Analysis

5.4.1. A Comparison between the Pivot-based Approach and Back-translation

Before we choose the pivot language-based method for data synthesis, we conduct a preliminary experiment analyzing both pivot-based and direct back-translation. The model used for direct back-translation was trained with the ground truth Europarl Fr-De data made from the multi-parallel corpus presented in Table 2. On the newstest 2012/2013 sets, the synthetic corpus generated using the pivot approach showed higher BLEU (19.11 / 20.45) than the back-translation counterpart (18.23 / 19.81) when used in training a De → Fr NMT model. Although the back-translation method has been effective in many studies [5], its availability becomes restricted in low-resource cases which is our major concern. This is due to the inferior quality of a back-translation model built from a limited source-to-target parallel corpus. Instead, one can utilize abundant pivot-to-target parallel corpora by using a rich-resource language as the pivot language. This consequently improves the reliability of

Corpus	Fr → De			De → Fr		
	news 2011	news 2012	news 2013	news 2011	news 2012	news 2013
Fr−De*	13.30	13.81	14.89	18.78	19.01	20.32
Fr*−De	13.81	**14.52**	15.20	18.46	18.73	19.82
$PSEUDO_{mix}$	**13.90**	14.50	**15.57**	**18.81**	**19.33**	**20.41**

Table 3. Translation results (BLEU) for Fr ↔ De experiments. The notation * denotes the synthetic part of the parallel corpus. The highest BLEU for each set is bold−faced.

the quality of baseline translation models used for generating synthetic corpora.

5.4.2. Effects of Mixing Source− and Target−originated Synthetic Parallel Data

From Table 3, we find that the bias of the synthetic examples in pseudo parallel corpora brings imbalanced performance to the bidirectional translation tasks. For instance, the Fr*−De corpus reports the highest BLEU for the Fr → De case while showing the lowest performance for De → Fr on the development set. Given that the source− and target−originated classification of a specific synthetic corpus is reversed depending on the direction of the translation, the overall results imply that the target−originated corpus for each translation task outperforms the source−originated data. The preference for target−originated synthetic data over source−originated counterparts was formerly investigated in SMT by Lambert et al. [19]. In NMT, it can be explained by degradation in

Corpus	Fr → De			De → Fr		
	news 2011	news 2012	news 2013	news 2011	news 2012	news 2013
(a) Fr*−De *(K=3)*	13.76	14.43	15.18	−	−	−
(b) Fr*−De *(K=5)*	13.78	**14.49**	15.23	17.76	18.63	19.73
(a) + (b)	13.74	14.38	15.27	−	−	−
(c) Fr−De* *(K=3)*	−	−	−	18.44	18.70	20.32
(d) Fr−De* *(K=5)*	13.36	14.08	15.28	18.18	18.76	20.13
(c) + (d)	−	−	−	18.06	18.63	20.21
(b) + (d)	**13.93**	14.27	**15.53**	**18.52**	**19.04**	**20.33**

Table 4. Translation results (BLEU) for Fr ↔ De translation tasks. K denotes the beam size used to generate the corresponding synthetic parallel data. The highest BLEU for each set is bold−faced.

the quality of source−originated data owing to an erroneous target language model formed by synthetic target sentences. In contrast, we observe that $\text{PSEUDO}_{\text{mix}}$ produces balanced results for both Fr → De and De → Fr translation tasks. Furthermore, we observe that $\text{PSEUDO}_{\text{mix}}$ even shows the best or competitive performance among all synthetic parallel corpora for each task.

We note that mixing two different synthetic parallel data leads to improved BLEU but not intermediate values. To investigate the cause of the BLEU improvement in $\text{PSEUDO}_{\text{mix}}$, we build additional target−originated synthetic corpora for each Fr ↔ De translation with a beam of size 3. We apply the same preprocessing step and again match the size of each synthetic parallel corpus. As shown in Table 4, for the De → Fr task, the new target−originated corpus (c)

Corpus	Fr → De		De → Fr	
	NMT	PBSMT	NMT	PBSMT
Fr−De*	14.89	11.65	20.32	17.46
Fr*−De	15.20	12.06	19.82	17.38
$PSEUDO_{mix}$	**15.57**	**12.19**	**20.41**	**17.79**

Table 5. A comparison between neural machine translation (NMT) and phrase−based statistical machine translation (PBSMT) evaluated on the newstest 2013 set.

shows higher BLEU than the source−originated corpus (b) by itself. The improvement in BLEU, however, occurs only when mixing the source− and the target−originated synthetic parallel data (b + d), compared to mixing two target−originated synthetic corpora (c + d). The same phenomenon is also observed in the Fr → De case as well. The results suggest that real and synthetic sentences mixed on either side of sentence pairs indeed enhance the capability of a synthetic parallel corpus. We conjecture that ground truth examples in both encoder and decoder networks not only compensate for the erroneous language model learned from synthetic sentences but also reinforce patterns of use latent in the pseudo sentences.

5.4.3. A Comparison with Phrase−based Statistical Machine Translation

We also evaluate the effects of the proposed approach in the phrase−based statistical machine translation [28]. We used Moses [23] and its baseline configuration for training, and a 5−gram Kneser−Ney

model as the language model. Table 5 shows the translation results of the phrase-based statistical machine translation (PBSMT) systems. In all experiments, NMT shows higher BLEU (2.44-3.38) compared to the PBSMT setting. We speculate that the deep architecture of NMT provides more robustness to the noise in the synthetic examples. We also note that the proposed $\text{PSEUDO}_{\text{mix}}$ outperforms other synthetic corpora in PBSMT. This result clearly shows that the benefits of the mixed composition in synthetic sentence pairs exist beyond a specific machine translation framework.

VI. Experiments: Large-scale Application

The experiments shown in the previous section verified the potential of $\text{PSEUDO}_{\text{mix}}$ as an efficient alternative to ground truth real parallel data. The conditions in the previous case, however, were somewhat artificial, as we deliberately matched the sources of all pseudo parallel corpora. In this section, we discuss more practical and large-scale applications of synthetic parallel data. Experiments are conducted on Czech (Cs) ↔ German (De) and French (Fr) ↔ German (De) translation tasks.

6.1. Application Scenarios

We analyze the efficacy of the proposed mixing approach in the following application scenarios:

i) Pseudo Only: This setting trains NMT models using only synthetic parallel data without any ground truth real parallel corpus.

ii) Real Fine-tuning: Once the training of an NMT model is completed in the Pseudo Only manner, the model is fine-tuned using only a real parallel corpus.

Corpus	Size	Average length	
		Cs	De
Europarl+NC11	0.6 M	23.54	25.49
Cs−De*	3.5 M	25.33	26.01
Cs*−De	3.5 M	23.31	25.37
$PSEUDO_{mix}$	3.5 M	24.39	25.72

(a) Cs ↔ De

Corpus	Size	Average length	
		Fr	De
Europarl+NC11	1.8 M	26.18	24.08
Fr−De*	3.7 M	26.67	23.71
Fr*−De	3.7 M	25.42	24.90
$PSEUDO_{mix}$	3.7 M	26.01	24.33

(b) Fr ↔ De

Table 6. Statistics of the training parallel corpora for large−scale Cs ↔ De and Fr ↔ De translation tasks.

The suggested scenarios reflect low−resource situations in building NMT systems. During Real fine−tuning, we fine−tune the best model of the Pseudo Only scenario evaluated on the development set.

6.2. Data Preparation

We use the parallel data from the shared translation task of WMT'15 and WMT'16 [29]. Using the same pivot−based technique as the

small-scale task, we build Cs-De* and Fr-De* corpora from the WMT'15 Cs-En and Fr-En parallel data respectively. For Cs*-De and Fr*-De, WMT'16 En-De parallel data is employed. We again use pre-trained NMT models for En → Cs, En → De and En → Fr to generate synthetic sentences. A beam of size 1 is used for fast decoding.

For the Real Fine-tuning scenario, we use ground truth real parallel corpora from the Europarl and News Commentary11 dataset. These direct parallel corpora are obtained from OPUS [30]. The sizes of each set of ground truth and synthetic parallel data is presented in Table 6. Given that the size of the training corpus for widely studied language pairs amounts to several million lines, the Cs-De language pair (0.6 M) reasonably represents a low-resource situation. On the other hand, the Fr-De language pair (1.8 M) is relatively resource-rich in our experiments. The details of the preprocessing are identical to those in the previous case.

6.3. Training and Evaluation

We use the same experimental settings that we used for the previous case. In the fine-tuning step, we use the learning rate of 2e-5 which produced better results. Embeddings are fixed throughout the fine-tuning steps. For evaluation, we use the same development and test sets used in the previous tasks.

Baseline	Cs → De		De → Cs	
(a) Europarl + NC11	13.15		11.16	
(b) + Pivot back-translation corpus	(+3.82) 16.97		(+4.24) 15.40	
Synthetic Corpus	Pseudo Only	Real Fine-tuning	Pseudo Only	Real Fine-tuning
Cs-De*	14.77	(+1.66) 16.43	14.34	(+0.86) 15.20
Cs*-De	16.88	(+0.17) 17.05	15.48	(+0.53) **16.01**
$PSEUDO_{mix}$	16.98	(+0.46) **17.44**	15.66	(+0.17) 15.83

(a) Cs ↔ De

Baseline	Fr → De		De → Fr	
(a) Europarl + NC11	16.14		20.86	
(b) + Pivot back-translation corpus	(+1.26) 17.40		(+1.76) 22.62	
Synthetic Corpus	Pseudo Only	Real Fine-tuning	Pseudo Only	Real Fine-tuning
Fr-De*	15.48	(+1.68) 17.16	20.73	(+2.07) 22.80
Fr*-De	17.15	(+0.54) 17.69	17.60	(+5.47) 23.07
$PSEUDO_{mix}$	16.94	(+0.95) **17.89**	20.11	(+3.11) **23.22**

(b) Fr ↔ De

Table 7. Translation results (BLEU) for Pseudo Only and Real Fine-tuning scenarios evaluated with the newstest 2011 set. For the results of the Real Fine-tuning, the values in parentheses are improvements in BLEU compared to the Pseudo Only setting. The highest BLEU for each translation task is bold-faced.

Baseline	Cs → De		De → Cs	
(a) Europarl + NC11	13.49		10.76	
(b) + Pivot back-translation corpus	(+3.92) 17.41		(+4.54) 15.30	
Synthetic Corpus	Pseudo Only	Real Fine-tuning	Pseudo Only	Real Fine-tuning
Cs-De*	15.26	(+1.81) 17.07	14.08	(+0.79) 14.87
Cs*-De	17.05	(+0.13) 17.18	15.17	(+0.35) 15.52
$PSEUDO_{mix}$	16.97	(+0.57) **17.54**	15.37	(+0.28) **15.65**

(a) Cs ↔ De

Baseline	Fr → De		De → Fr	
(a) Europarl + NC11	16.36		21.45	
(b) + Pivot back-translation corpus	(+1.74) 18.10		(+1.86) 23.31	
Synthetic Corpus	Pseudo Only	Real Fine-tuning	Pseudo Only	Real Fine-tuning
Fr-De*	16.59	(+1.23) 17.82	21.56	(+1.43) 22.99
Fr*-De	17.42	(+0.57) 17.99	18.27	(+5.11) 23.38
$PSEUDO_{mix}$	17.42	(+0.92) **18.34**	21.20	(+2.45) **23.65**

(b) Fr ↔ De

Table 8. Translation results (BLEU) for Pseudo Only and Real Fine-tuning scenarios evaluated with the newstest 2012 set. For the results of the Real Fine-tuning, the values in parentheses are improvements in BLEU compared to the Pseudo Only setting. The highest BLEU for each translation task is bold-faced.

Baseline	Cs → De		De → Cs	
(a) Europarl + NC11	14.96		12.36	
(b) + Pivot back-translation corpus	(+4.02) 18.98		(+4.40) 16.76	
Synthetic Corpus	Pseudo Only	Real Fine-tuning	Pseudo Only	Real Fine-tuning
Cs-De*	16.87	(+1.95) 18.82	15.29	(+1.21) 16.50
Cs*-De	18.62	(+0.40) 19.02	16.51	(+0.45) 16.96
$PSEUDO_{mix}$	18.82	(+0.53) **19.35**	16.79	(+0.68) **17.47**

(a) Cs ↔ De

Baseline	Fr → De		De → Fr	
(a) Europarl + NC11	17.68		22.39	
(b) + Pivot back-translation corpus	(+1.59) 19.27		(+1.93) 24.32	
Synthetic Corpus	Pseudo Only	Real Fine-tuning	Pseudo Only	Real Fine-tuning
Fr-De*	17.57	(+1.65) 19.22	22.88	(+1.42) 24.30
Fr*-De	18.55	(+1.04) 19.59	19.87	(+4.74) 24.61
$PSEUDO_{mix}$	18.98	(+0.87) **19.85**	22.71	(+1.99) **24.70**

(b) Fr ↔ De

Table 9. Translation results (BLEU) for Pseudo Only and Real Fine-tuning scenarios evaluated with the newstest 2013 set. For the results of the Real Fine-tuning, the values in parentheses are improvements in BLEU compared to the Pseudo Only setting. The highest BLEU for each translation task is bold-faced.

6.4. Results and Analysis

6.4.1. A Comparison with Real Parallel Data

Tables 7, 8, and 9 show the results from the Pseudo Only and Real Fine-tuning scenarios for Cs ↔ De and Fr ↔ De translation tasks evaluated on the newstest 2011, 2012, and 2013 sets. For a baseline comparison, we present the translation quality of the NMT models trained with the ground truth Europarl and News Commentary11 parallel corpora (a). In Cs ↔ De, the Pseudo Only scenario shows outperforming results compared to the real parallel corpus by up to 3.86-4.43 BLEU on the newstest 2013 set. Even for the Fr ↔ De case, where the size of the real parallel corpus is relatively large, the best BLEU of the pseudo parallel corpora is higher than that of the real parallel corpus by 1.3 (Fr → De) and 0.49 (De → Fr) on the same test set. From the results, we conclude that large-scale synthetic parallel data can perform as an effective alternative to the real parallel corpus particularly in low-resource language pairs.

6.4.2. Results from the Pseudo Only Scenario

As shown in Table 9, the model learned from the Cs^{*}-De corpus outperforms the model trained with the Cs-De^{*} corpus in every case. The result is slightly different from that of the previous case, where the target-originated data for each translation task reports better results than the source-originated data. This arises from the diversity in the source of each pseudo parallel corpus, which vary in their suitability for the given test set. Table 9 also shows that mixing

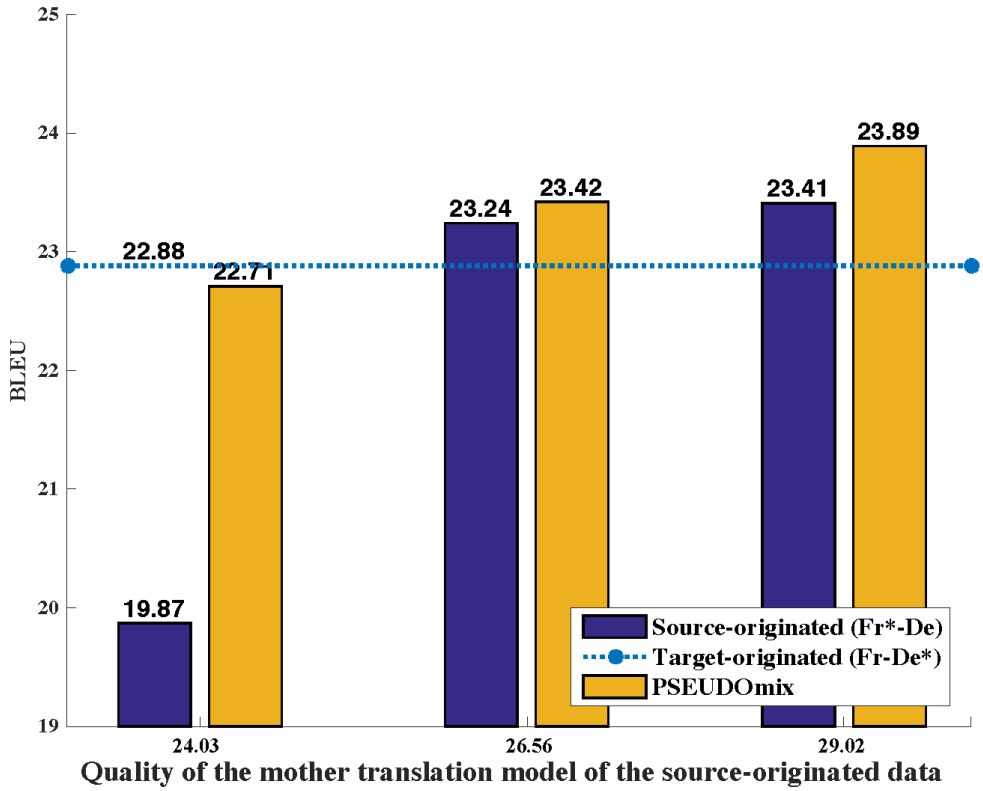

Figure 2. Translation results for the De → Fr task on the newstest 2013 set with respect to the quality of the mother model for the source−originated Fr*−De data. The quality of the mother model is evaluated on the En−Fr newstest 2012 set.

the Cs*−De corpus with the Cs−De* corpus of worse improves the resulting $PSEUDO_{mix}$, showing the highest BLEU for bidirectional Cs ↔ De translation tasks. In addition, $PSEUDO_{mix}$ again shows much more balanced performance in Fr ↔ De translations compared to other synthetic parallel corpora.

While the mixing strategy compensates for most of the BLEU gap between the Fr−De* and the Fr*−De (newstest2013: 3.01 → 0.17) in the De → Fr case, the resulting $PSEUDO_{mix}$ still shows lower BLEU than the target−originated Fr−De* corpus. We thus enhance

the quality of the synthetic examples of the source-originated Fr^*-De data by further training its mother translation model. A larger beam of size 5 is also used to enhance the quality of synthetic sentences. As Figure 2 illustrates, with the target-originated Fr-De^* corpus being fixed, the quality of the models trained with the source-originated Fr^*-De data and $PSEUDO_{mix}$ increases in proportion to the quality of the mother model for the Fr^*-De corpus. Eventually, $PSEUDO_{mix}$ shows the highest BLEU (23.89), outperforming both Fr^*-De (23.41) and Fr-De^* (22.88) data on the newstest 2013 set. The results indicate that the benefit of the proposed mixing approach becomes more evident when the performance gap between the source- and the target-originated synthetic parallel data is within a certain range.

6.4.3. Results of Real Fine-tuning Scenario

As presented in Tables 7, 8, and 9, we observe that fine-tuning using ground truth parallel data substantially improves the qualities of NMT models trained with synthetic parallel corpora. Among all fine-tuned models, the proposed $PSEUDO_{mix}$ shows the best translation quality in almost every experiment. This is particularly encouraging for the case of De → Fr where $PSEUDO_{mix}$ reported lower BLEU than the Fr-De^* data before it was fine-tuned. Even in the case where $PSEUDO_{mix}$ shows comparable results with other synthetic parallel corpora in the Pseudo Only scenario, it shows higher improvements in translation quality when fine-tuned with real parallel data. These results clearly demonstrate the benefits of the proposed $PSEUDO_{mix}$

that indicates both competitive translation quality by itself and relatively higher potential improvement as a result of the refinement using ground truth parallel corpora.

In Tables 7, 8, and 9 (b), we also present the performance of NMT models learned from the ground truth Europarl+NC11 data merged with the target-originated synthetic parallel corpus for each task. This is identical in spirit to the method in Sennrich et al. [5] which employs back-translation for data synthesis. Instead of direct back-translation, we used pivot-based back-translation, as we verified the benefit of the pivot-based data synthesis in low-resource environments. Although the ground truth data is only used for the refinement, the Real Fine-tuning scheme applied to $\text{PSEUDO}_{\text{mix}}$ shows better translation quality compared to the models trained with the merged corpus (b). Even the results of the Real Fine-tuning on the target-originated corpus provide comparable results to the training with the merged corpus from scratch. The overall results support the efficacy of the proposed two-step methods in empirical application: the Pseudo Only method to introduce useful prior on the NMT parameters and the Real Fine-tuning scheme to reorganize the pre-trained NMT parameters using in-domain parallel data.

VII. Conclusion

In this work, we have constructed NMT systems using only synthetic parallel data. For this purpose, we suggest a novel pseudo parallel corpus called $\text{PSEUDO}_{\text{mix}}$ where synthetic and ground truth real examples are mixed on either side of sentence pairs. $\text{PSEUDO}_{\text{mix}}$ can be readily composed by mixing existing pseudo parallel corpora, namely source- and target-originated synthetic parallel data. Experiments show that the proposed $\text{PSEUDO}_{\text{mix}}$ not only shows enhanced translation quality for bidirectional translation but also reports substantial improvement when fine-tuned with ground truth parallel data. Our work has significance in that it provides a thorough investigation of the use of synthetic parallel corpora in a low-resource NMT environment. Without any adjustment, the proposed method can also be extended to other learning areas where parallel samples are employed. For future work, robust data sampling methods to maximize the quality of the mixed synthetic parallel data should be explored.

Bibliography

[1] Nal Kalch brenner and Phil Blunsom. 2013. Recurrent continuous translation models. In Proceedings of EMNLP. volume 3, page 413.

[2] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014a. Learning phrase representations using rnn encoder–decoder for statistical machine translation. arXiv preprint arXiv: 1406.1078

[3] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems. pages 3104–3112.

[4] Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. 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 (Volume 1: Long Papers). Association for Computational Linguistics, Beijing, China, pages 1–10.

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

[6] Minh-Thang Luong and Christopher D Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. arXiv preprint arXiv: 1604.00788.

[7] Jiajun Zhang and Chengqing Zong. 2016a. Exploiting source-side monolingual data in neural machine translation. In Proceedings of EMNLP.

[8] Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016a. Semi-supervised learning for neural machine translation. arXiv preprint arXiv: 1606.04596.

[9] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016a. Zero-resource translation with multi-lingual neural machine translation. arXiv preprint arXiv: 1606.04164.

[10] Jiajun Zhang and Chengqing Zong. 2016b. Bridging neural machine translation and bilingual dictionaries. arXiv preprint arXiv: 1610.07272.

[11] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viegas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-

shot translation. arXiv preprint arXiv: 1611.04558.

[12] Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016b. Neural machine translation with pivot languages. arXiv preprint arXiv: 1611.04927.

[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 1409.0473.

[14] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv: 1508.04025.

[15] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. On the properties of neural machine translation: Encoder-decoder approaches. In Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8).

[16] Nicola Ueffing, Gholamreza Haffari, Anoop Sarkar, et al. 2007. Transductive learning for statistical machine translation. In Annual Meeting-Association for Computational Linguistics. volume 45, page 25.

[17] Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain

adaptation for statistical machine translation with domain dictionary and monolingual corpora. In proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, pages 993-1000.

[18] Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In Proceedings of the fourth workshop on statistical machine translation. Association for Computational Linguistics, pages 182-189.

[19] Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, pages 284-293.

[20] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016b. Multi-way, multilingual neural machine translation with a shared attention mechanism. arXiv preprint arXiv: 1601.01073.

[21] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In MT summit. volume 5, pages 79-86.

[22] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara

Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In Proceedings of the Tenth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Lisbon, Portugal, pages 1–46.

[23] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics, pages 177–180.

[24] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. arXiv preprint arXiv: 1508.07909.

[25] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980.

[26] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2013. How to construct deep recurrent neural networks. arXiv preprint arXiv: 1312.6026.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pages 311–318.

[28] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase–based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Volume 1. Association for Computational Linguistics, pages 48–54.

[29] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In Proceedings of the First Conference on Machine Translation. Association for Computational Linguistics, Berlin, Germany, pages 131–198.

[30] Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In LREC. volume 2012, pages 2214–2218.

[31] Jaehong Park, Jongyoon Song, and Sungroh Yoon. 2017. Building a Neural Machine Translation System Using Only Synthetic Parallel Data. arXiv preprint arXiv: 1704.00253

국문초록

학습된 번역 모델에 의해 생성 가능한 합성 병렬데이터는 최근 인공신경망 기계번역에서 발생하는 다양한 이슈에 효과적인 해결책으로 대두되었다. 이러한 합성 병렬데이터의 효용에 착안하여 본 연구에서는 합성 병렬데이터만을 활용하여 인공신경망 기계번역 시스템을 구축한다. 더불어 본 연구에서는 실제 병렬 데이터의 효과적인 대안이 될 수 있는 새로운 유형의 합성 병렬데이터를 제시한다. 본 연구에서 제안하는 합성 병렬데이터는 실제 문장과 합성된 문장이 병렬 문장 쌍의 양쪽에 혼재되어 있다는 점에서 기존에 제시됐던 합성 병렬데이터와 차별성을 갖는다. 동일한 조건에서 본 연구가 제안하는 합성 병렬데이터로 인공신경망 기계번역 시스템을 학습한 결과, 기존에 제시됐던 합성 병렬데이터로 학습한 경우에 비해 양방향 번역에서 보다 우수하고 안정적인 번역 성능을 나타냈다. 또한 새로운 합성 병렬데이터로 학습한 인공신경망 번역 모델을 실제 병렬데이터로 fine-tuning 할 경우, 기존에 제시된 합성 병렬데이터에 비해 상대적으로 높은 번역 성능의 향상을 확인할 수 있었다.

주요어 : 기계번역, 인공신경망 기계번역, 합성 병렬데이터
학 번 : 2015-20929