

Building a Neural Machine Translation System Using Only Synthetic Parallel Data

Jaehong Park and **Jongyoon Song** and **Sungroh Yoon**

Department of Electrical and Computer Engineering, Seoul National University

{uwanggood, coms1580, sryoon}@snu.ac.kr

Abstract

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. As an efficient alternative to real parallel data, we also present a new type of synthetic parallel corpus. The proposed pseudo parallel data are distinct from 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, which shows not only enhanced results for bidirectional translation tasks but also substantial improvement with the aid of a ground truth real parallel corpus.

1 Introduction

Given the data-driven nature of neural machine translation (NMT), the limited 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 with regard to the data scarcity in NMT. Many studies have found that the pseudo parallel data combined with the real bilingual parallel corpus significantly enhance the quality of NMT models (Sennrich et al., 2015a; Zhang and Zong, 2016b; Cheng et al., 2016b). In addition, synthesized parallel data have played vital roles in many NMT problems such as domain adaptation (Sennrich et al., 2015a), zero-resource NMT (Firat et al., 2016b), and the rare word problem (Zhang and Zong, 2016a).

Inspired by their efficacy, we attempt to train NMT models using only synthetic parallel data.

To the best of our knowledge, building NMT systems with only pseudo parallel data has yet to be studied. Through our research, we explore the availability of synthetic parallel data as an effective alternative to the real-world parallel corpus. The active usage of synthetic data in NMT particularly has its significance in low-resource environments where the ground truth parallel corpora are very limited or not established. Even in recent approaches such as zero-shot NMT (Johnson et al., 2016) and pivot-based NMT (Cheng et al., 2016a), where direct source-to-target bilingual data are not required, the direct parallel corpus brings substantial improvements in translation quality where the pseudo parallel data can also be employed.

Previously suggested synthetic data, however, have several drawbacks to be a reliable alternative to the real parallel corpus. As illustrated in Figure 1, existing pseudo parallel corpora can be classified into two groups: *source-originated* and *target-originated*. The common property between them is that ground truth examples exist only on a single side (source or target) of pseudo sentence pairs, while the other side is composed of synthetic sentences only. The bias of synthetic examples in sentence pairs, however, may lead to the imbalance of the quality of learned NMT models when the given pseudo parallel corpus is exploited in 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 translation model would prevent NMT models from learning solid parameters.

To overcome these shortcomings, we propose a novel synthetic parallel corpus called PSEUDO_{mix}. In contrast to previous works, PSEUDO_{mix} includes both synthetic and real sentences on either side of 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 demonstrate that the proposed PSEUDO_{mix} shows useful properties that make it a reliable candidate for real-world parallel data. In detail, we make the following contributions:

1. PSEUDO_{mix} shows more balanced translation quality compared to existing pseudo parallel corpora in bidirectional translation tasks. For each task, it outperforms both source- and target-originated data when their performance gap is under a certain range.
2. When fine-tuned using real parallel data, the model trained with PSEUDO_{mix} outperforms other fine-tuned models trained with source-originated and target-originated synthetic parallel data, indicating substantial improvement in translation quality.

2 Neural Machine Translation

Given a source sentence $x = (x_1, \dots, x_m)$ and its corresponding target sentence $y = (y_1, \dots, y_n)$, 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 architecture (Kalchbrenner and Blunsom, 2013; Cho et al., 2014b; Sutskever et al., 2014). Thereafter, the attention mechanism (Bahdanau et al., 2014; Luong et al., 2015) has been introduced and successfully addressed the quality degradation of NMT when dealing with long input sentences (Cho et al., 2014a).

In this study, we use the attentional NMT architecture proposed by Bahdanau et al. (2014). In their work, the encoder, which is a bidirectional recurrent neural network, reads the source sentence and generates a sequence of source representations $\mathbf{h} = (\mathbf{h}_1, \dots, \mathbf{h}_m)$. The decoder, which is another recurrent neural network, produces the target sentence one symbol at a time. The log conditional probability thus can be decomposed as follows:

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

where $y_{<t} = (y_1, \dots, y_{t-1})$. As described in Equation (2), the conditional distribution of $p(y_t|y_{<t}, x)$ is modeled as a function of the previously predicted output y_{t-1} , the hidden state of the decoder s_t , and the context vector c_t .

$$p(y_t|y_{<t}, x) \propto \exp\{g(y_{t-1}, s_t, c_t)\} \quad (2)$$

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 sum of source representations $\mathbf{h}_1, \dots, \mathbf{h}_m$. Each weight α_{ti} for \mathbf{h}_i implies the probability of the target symbol y_t being aligned to the source symbol x_i :

$$c_t = \sum_{i=1}^m \alpha_{ti} \mathbf{h}_i \quad (3)$$

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 $\{(x^n, y^n)\}_{n=1}^N$:

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{n=1}^N \log p(y^n|x^n) \quad (4)$$

where θ^* is the optimal parameter.

3 Related Work

In statistical machine translation (SMT), synthetic bilingual data have been primarily proposed as a means to exploit monolingual corpora. By applying a self-training scheme, the pseudo parallel data were obtained by automatically translating the source-side monolingual corpora (Ueffing et al., 2007; Wu et al., 2008). In a similar but reverse way, the target-side monolingual corpora were also employed to build the synthetic parallel data (Bertoldi and Federico, 2009; Lambert et al., 2011). The primary goal of these works was to adapt trained SMT models to other domains using relatively abundant in-domain monolingual data.

Inspired by the successful application in SMT, there have been efforts to exploit synthetic parallel data in improving NMT systems. Source-side (Zhang and Zong, 2016b), target-side (Sennrich et al., 2015a) and both sides (Cheng et al., 2016b) 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., (2015a), domain adaptation of NMT was achieved by fine-tuning trained NMT models using a synthetic parallel corpus. Firat et al. (2016b) 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 (Firat et al., 2016a) models, which enabled zero-resource NMT between the source and target languages. Lastly, synthetic sentence pairs have been utilized to enrich the training examples having rare or unknown translation lexicons (Zhang and Zong, 2016a).

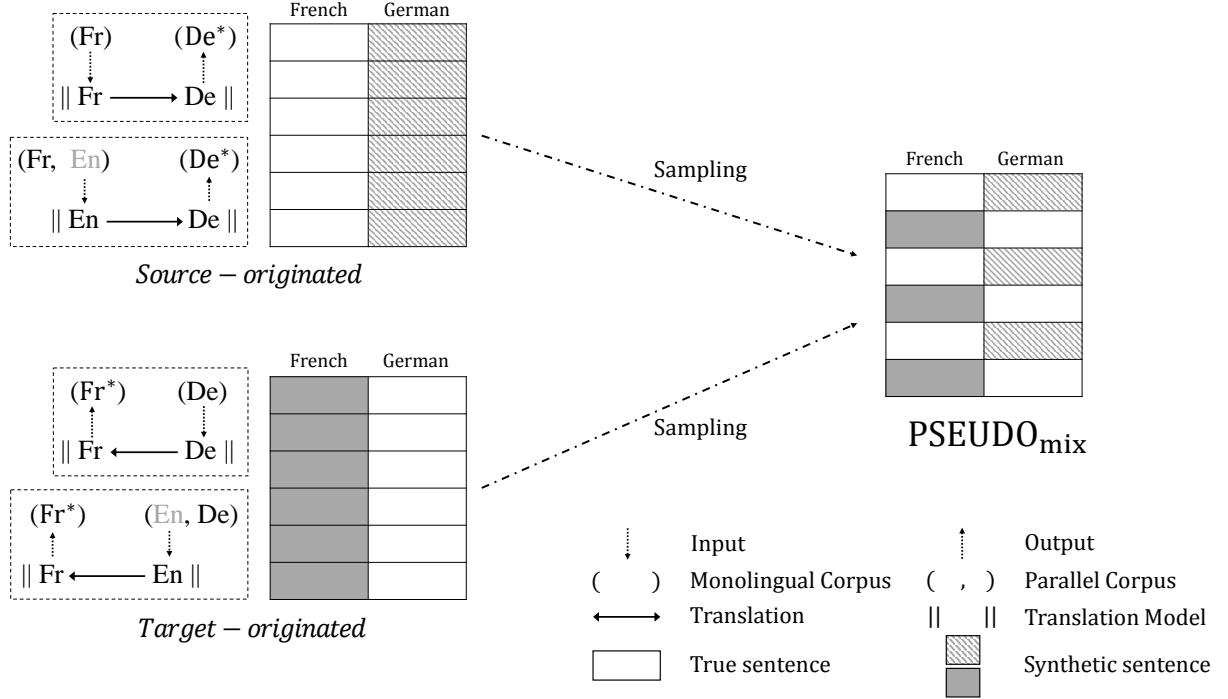


Figure 1: The process of building each pseudo parallel corpus group for French → German translation. * indicates the synthetic sentences generated by translation models. Each of the source-originated and the target-originated synthetic parallel data can be made from French or German monolingual corpora.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

4 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 performance of NMT. In this work, we further extend their application by training NMT with only synthetic data. In certain language pairs or domains where the source-to-target real parallel corpora are very rare or even unprepared, the 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 real parallel data.

4.2 Limits of the Previous Approaches

For a given translation task, we classify the existing pseudo parallel data into the following groups:

- (a) *Source-originated*: The source sentences are from a real corpus, and the associated target sentences are synthetic. The corpus can be formed by automatically translating a source-side monolingual corpus into the target language (Zhang and Zong, 2016a,b). 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.

- (b) *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 (Sennrich et al., 2015a). Similar to the source-originated case, it can be built from a pivot-target bilingual corpus using a pivot-to-source translation model (Firat et al., 2016b).

The process of building each synthetic parallel corpus is illustrated in Figure 1. As shown in Figure 1, the previous studies on pseudo parallel data share a common property: synthetic and ground truth sentences are biased on a single side of sentence pairs. In such a case where the synthetic parallel data are the only or major resource used to train NMT, this may severely limit the availability of the given pseudo parallel corpus. For instance, as will be demonstrated in our experiments, synthetic data showing relatively high quality in one translation task (e.g., French→German) can pro-

Corpus	Size	Avg len	
		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 1: Statistics of the parallel corpora for Fr \leftrightarrow De translation tasks. The notation * denotes the synthetic part of the parallel corpus.

duce poor results in the translation task of the reverse direction (German \rightarrow 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 from which the synthetic sentences originate. Depending on the quality of the mother model, ill-formed or inaccurate synthetic examples could be generated, which would negatively affect the reliability of the resultant synthetic parallel data. In the previous study, Zhang and Zong (2016b) 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 new type of synthetic parallel corpus called PSEUDO_{mix}. Our approach is quite straightforward: For a given translation task, we first build both source-originated and target-originated pseudo parallel data. PSEUDO_{mix} can then be readily built by mixing them together. The overall process of building PSEUDO_{mix} for the French \rightarrow German translation task is illustrated in Figure 1.

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

5 Experiments: Effects of Mixing Real and Synthetic Sentences

In this section, we analyze the effects of the 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 corpora from the identical source-to-target real parallel corpus. Our experiments are performed on French (Fr) \leftrightarrow German (De) translation tasks. Throughout the remaining paper, we use the notation * to denote the synthetic part of the pseudo sentence pairs.

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 (Koehn, 2005), constructing a multi-parallel corpus of Fr-En-De. Then each of the Fr*-De and Fr-De* pseudo parallel corpora 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 \rightarrow De and train another NMT model for En \rightarrow Fr using the WMT’15 En-Fr parallel corpus (Bojar et al., 2015). 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* corpus and mixing them together.

5.2 Data Preprocessing

Each training corpus is tokenized using the tokenization script in Moses (Koehn et al., 2007). We represent every sentence as a sequence of subword units learned from byte-pair encoding (Sennrich et al., 2015b). We remove empty lines and all the sentences of length over 50 subword units. For a fair comparison, all cleaned synthetic parallel data have equal sizes. The summary of the final parallel corpora is presented in Table 1.

5.3 Training and Evaluation

All networks have 1024 hidden units and 500 dimensional embeddings. The vocabulary size is limited to 30K for each language. Each model is trained for 10 epochs using stochastic gradient descent with Adam (Kingma and Ba, 2014). The Minibatch size is 80, and the training set is reshuffled between every epoch. The norm of the gra-

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

Corpus	Fr → De			De → Fr		
	newstest2011	newstest2012	newstest2013	newstest2011	newstest2012	newstest2013
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 2: 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.

dient is clipped not to exceed 1.0 (Pascanu et al., 2013). The learning rate is $2 \cdot 10^{-4}$ 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, beam search is used to approximately find the most likely translation. 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 (Papineni et al., 2002) computed with the multi-bleu.perl script from Moses. For each case, we present average BLEU evaluated on three different models trained from scratch.

5.4 Results and Analysis

5.4.1 A Comparison between 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 (Sennrich et al., 2015a, 2016), its availability becomes restricted in low-resource cases which is our major concern. This is due to the poor quality of the back-translation model built from the 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 the quality of baseline translation models used for generating synthetic corpora.

5.4.2 Effects of Mixing Source- and Target-originated Synthetic Data

From Table 2, we find that the bias of the synthetic examples in pseudo parallel corpora brings

imbalanced quality in the bidirectional translation tasks. 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 of target-originated synthetic data over the source-originated counterparts was formerly investigated in SMT by Lambert et al., (2011). In NMT, it can be explained by the degradation in quality in the source-originated data owing to the erroneous target language model formed by the synthetic target sentences. In contrast, we observe that PSEUDO_{mix} not only produces balanced results for both Fr→De and De→Fr translation tasks but also shows the best or competitive translation quality for each task.

We note that mixing two different synthetic corpora leads to improved BLEU not their intermediate value. To investigate the cause of the improvement in PSEUDO_{mix}, we build additional target-originated synthetic corpora for each Fr↔De translation with a beam of size 3. As shown in Table 3, for the De→Fr task, the new target-originated corpus (c) shows higher BLEU than the source-originated corpus (b) by itself. The improvement in BLEU, however, occurs only when mixing the source- and target-originated synthetic parallel data (b+d) compared to mixing two target-originated synthetic corpora (c+d). The same phenomenon is observed in the Fr→De case as well. The results suggest that real and synthetic sentences mixed on either side of sentence pairs 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 reinforces 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 mixing strategy in phrase-based statistical machine translation (Koehn et al., 2003). We use

Corpus	Fr → De			De → Fr		
	newstest2011	newstest2012	newstest2013	newstest2011	newstest2012	newstest2013
(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 3: Translation results (BLEU) for $\text{Fr} \leftrightarrow \text{De}$ experiments. K denotes the beam size used to generate the corresponding synthetic parallel data. The highest BLEU for each set is bold-faced.

Corpus	Fr → De		De → Fr	
	NMT	SMT	NMT	SMT
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 4: Translation results (BLEU) for $\text{Fr} \leftrightarrow \text{De}$ experiments evaluated on the newstest 2013 set.

Moses (Koehn et al., 2007) and its baseline configuration for training. A 5-gram Kneser-Ney model is used as the language model. Table 4 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 noise robustness in the synthetic examples. It is also notable that the proposed PSEUDO_{mix} outperforms other synthetic corpora in PBSMT. The results clearly show that the benefit of the mixed composition in synthetic sentence pairs is beyond a specific machine translation framework.

6 Experiments: Large-scale Application

The experiments shown in the previous section verify the potential of PSEUDO_{mix} as an efficient alternative to the real parallel data. The condition in the previous case, however, is somewhat artificial, as we deliberately match the sources of all pseudo parallel corpora. In this section, we move on to more practical and large-scale applications of synthetic parallel data. Experiments are conducted on Czech (Cs) \leftrightarrow German (De) and French (Fr) \leftrightarrow 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 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 ground truth parallel corpus.

The suggested scenarios reflect low-resource situations in building NMT systems. In the 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 corpora from the shared translation task of WMT’15 and WMT’16 (Bojar et al., 2016). Using the same pivot-based technique as the previous task, Cs-De* and Fr-De* corpora are built from the WMT’15 Cs-En and Fr-En parallel data respectively. For Cs*-De and Fr*-De, WMT’16 En-De parallel data are 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 real parallel corpora from the Europarl and News Commentary11 dataset. These direct parallel corpora are obtained from OPUS (Tiedemann, 2012). The size of each set of ground truth and synthetic parallel data is presented in Table 5. Given that the training corpus for widely studied language pairs amounts to several million lines, the Cs-De language pair (0.6M) reasonably represents a low-resource situation. On the other hand, the Fr-De language pair (1.8M) is considered to be relatively resource-rich in our experiments. The details of the preprocessing are identical to those in the previous case.

Corpus	Size	Avg length	
		Cs	De
Europarl+NC11	0.6M	23.54	25.49
Cs-De*	3.5M	25.33	26.01
Cs*-De	3.5M	23.31	25.37
PSEUDO _{mix}	3.5M	24.39	25.72

(a) Cs \leftrightarrow De

Corpus	Size	Avg length	
		Fr	De
Europarl+NC11	1.8M	26.18	24.08
Fr-De*	3.7M	26.67	23.71
Fr*-De	3.7M	25.42	24.90
PSEUDO _{mix}	3.7M	26.01	24.33

(b) Fr \leftrightarrow DeTable 5: Statistics of the training parallel corpora for large-scale Cs \leftrightarrow De and Fr \leftrightarrow De translation tasks.

6.3 Training and Evaluation

We use the same experimental settings that we used for the previous case except for the Real Fine-tuning scenario. In the fine-tuning step, we use the learning rate of $2 \cdot 10^{-5}$, 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 task.

6.4 Results and Analysis

6.4.1 A Comparison with Real Parallel Data

Table 6 shows the results of the Pseudo Only scenario on Cs \leftrightarrow De and Fr \leftrightarrow De tasks. For the baseline comparison, we also present the translation quality of the NMT models trained with the ground truth Europarl+NC11 parallel corpora (a). In Cs \leftrightarrow 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 \leftrightarrow 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 \rightarrow De) and 0.49 (De \rightarrow Fr). We list the results on the newstest 2011 and newstest 2012 in the appendix. From the results, we conclude that large-scale synthetic parallel data can perform as an effective alternative to the real parallel corpora, particularly in low-resource language pairs.

6.4.2 Results from the Pseudo Only Scenario

As shown in Table 6, the model learned from the Cs*-De corpus outperforms the model trained with the Cs-De* corpus in every case. This result is slightly different from the previous case, where the target-originated synthetic corpus 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 6 also shows that mixing the Cs*-De corpus with the Cs-De* corpus of worse quality brings improvements in the resulting PSEUDO_{mix}, showing the highest BLEU for

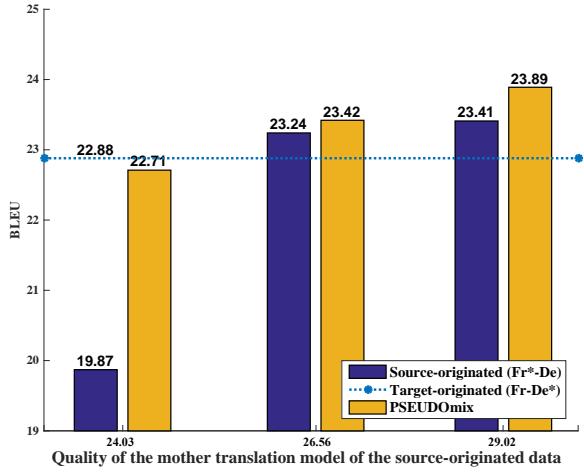


Figure 2: Translation results for the De \rightarrow Fr task on the newstest 2013 set with respect to the quality of the mother model for the source-originated Fr*-De data. The quality of the mother model is evaluated on the En-Fr newstest 2012 set.

bidirectional Cs \leftrightarrow De translation tasks. In addition, PSEUDO_{mix} again shows much more balanced performance in Fr \leftrightarrow De translations compared to other synthetic parallel corpora.

While the mixing strategy compensates for most of the gap between the Fr-De* and the Fr*-De (3.01 \rightarrow 0.17) in the De \rightarrow 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 (En \rightarrow Fr). As illustrated in Figure 2, 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, outperforming both Fr*-De and Fr-De* data. The results indicate that the benefit of the proposed mixing approach becomes much more evident when the quality gap between the source- and target-originated synthetic data is within a certain range.

Baseline		Cs → De		De → Cs	
(a) Europarl+NC11		14.96		12.36	
(b) +Pivot back-trans 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-trans 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 6: Translation results (BLEU) for Pseudo Only and Real Fine-tuning scenarios evaluated on 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.

6.4.3 Results from the Real Fine-tuning Scenario

As presented in Table 6, we observe that fine-tuning using ground truth parallel data brings substantial improvements in the translation qualities of all NMT models. Among all fine-tuned models, PSEUDO_{mix} shows the best performance in all experiments. 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 corpora in the Pseudo Only scenario, it shows higher improvements in the translation quality when fine-tuned with the real parallel data. These results clearly demonstrate the strengths of the proposed PSEUDO_{mix}, which indicate both competitive translation quality by itself and relatively higher potential improvement as a result of the refinement using ground truth parallel corpora.

In Table 6 (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. (2015a) which employs back-translation for data synthesis. Instead of direct back-translation, we used pivot-based back-translation, as we verified the strength 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

PSEUDO_{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 practical 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.

7 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 PSEUDO_{mix} where synthetic and ground truth real examples are mixed on either side of sentence pairs. Experiments show that the proposed PSEUDO_{mix} not only shows enhanced results 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 on the use of synthetic parallel corpora in 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, we plan to explore robust data sampling methods, which would maximize the quality of the mixed synthetic parallel data.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .
- 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.
- Ondřej 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. <http://www.aclweb.org/anthology/W/W16/W16-2301>.
- Ondřej 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. <http://aclweb.org/anthology/W15-3001>.
- Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016a. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928* .
- Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016b. Semi-supervised learning for neural machine translation. *arXiv preprint arXiv:1606.04596* .
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .
- Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073* .
- Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multilingual neural machine translation. *arXiv preprint arXiv:1606.04164* .
- Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558* .
- Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*. volume 3, page 413.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*. volume 5, pages 79–86.
- 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.
- 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.
- 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.
- Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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.
- Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2013. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026* .

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

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

Rico Sennrich, Barry Haddow, and Alexandra Birch.
2016. Edinburgh neural machine translation sys-
tems for wmt 16. *arXiv preprint arXiv:1606.02891*

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014.
Sequence to sequence learning with neural net-
works. In *Advances in neural information process-
ing systems*, pages 3104–3112.

Jörg Tiedemann. 2012. Parallel data, tools and inter-
faces in opus. In *LREC*. volume 2012, pages 2214–
2218.

Nicola Ueffing, Gholamreza Haffari, Anoop Sarkar,
et al. 2007. Transductive learning for statistical ma-
chine translation. In *Annual Meeting-Association
for Computational Linguistics*. volume 45, page 25.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008.
Domain adaptation for statistical machine transla-
tion with domain dictionary and monolingual cor-
pora. In *Proceedings of the 22nd International Con-
ference on Computational Linguistics-Volume 1*. As-
sociation for Computational Linguistics, pages 993–
1000.

Jiajun Zhang and Chengqing Zong. 2016a. Bridging
neural machine translation and bilingual dictionar-
ies. *arXiv preprint arXiv:1610.07272*.

Jiajun Zhang and Chengqing Zong. 2016b. Exploit-
ing source-side monolingual data in neural machine
translation. In *Proceedings of EMNLP*.

Baseline		Cs → De		De → Cs	
(a) Europarl+NC11		13.15		11.16	
(b) +Pivot back-trans 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-trans 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 on 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-trans 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-trans 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 on 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.