Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages

Yunsu Kim∗ Petre Petrov1,2* Pavel Petrushkov2 Shahram Khadivi2 Hermann Ney1
1RWTH Aachen University, Aachen, Germany {surname}@cs.rwth-aachen.de
2eBay, Inc., Aachen, Germany {petrpetrov,ppetrushkov,skhadivi}@ebay.com

Abstract

We present effective pre-training strategies for neural machine translation (NMT) using parallel corpora involving a pivot language, i.e., source-pivot and pivot-target, leading to a significant improvement in source→target translation. We propose three methods to increase the relation among source, pivot, and target languages in the pre-training: 1) step-wise training of a single model for different language pairs, 2) additional adapter component to smoothly connect pre-trained encoder and decoder, and 3) cross-lingual encoder training via autoencoding of the pivot language. Our methods greatly outperform multilingual models up to +2.6% BLEU in WMT 2019 French→German and German→Czech tasks. We show that our improvements are valid also in zero-shot/zero-resource scenarios.

1 Introduction

Machine translation (MT) research is biased towards language pairs including English due to the ease of collecting parallel corpora. Translation between non-English languages, e.g., French→German, is usually done with pivoting through English, i.e., translating French (source) input to English (pivot) first with a French→English model which is later translated to German (target) with a English→German model (De Gispert and Marino, 2006; Utiyama and Isahara, 2007; Wu and Wang, 2007). However, pivoting requires doubled decoding time and the translation errors are propagated or expanded via the two-step process.

Therefore, it is more beneficial to build a single source→target model directly for both efficiency and adequacy. Since non-English language pairs often have little or no parallel text, common choices to avoid pivoting in NMT are generating pivot-based synthetic data (Bertoldi et al., 2008; Chen et al., 2017) or training multilingual systems (Firat et al., 2016; Johnson et al., 2017).

In this work, we present novel transfer learning techniques to effectively train a single, direct NMT model for a non-English language pair. We pre-train NMT models for source→pivot and pivot→target, which are transferred to a source→target model. To optimize the usage of given source-pivot and pivot-target parallel data for the source→target direction, we devise the following techniques to smooth the discrepancy between the pre-trained and final models:

• Step-wise pre-training with careful parameter freezing.
• Additional adapter component to familiarize the pre-trained decoder with the outputs of the pre-trained encoder.
• Cross-lingual encoder pre-training with autoencoding of the pivot language.

Our methods are evaluated in two non-English language pairs of WMT 2019 news translation tasks: high-resource (French→German) and low-resource (German→Czech). We show that NMT models pre-trained with our methods are highly effective in various data conditions, when fine-tuned for source→target with:

• Real parallel corpus
• Pivot-based synthetic parallel corpus (zero-resource)
• None (zero-shot)

For each data condition, we consistently outperform strong baselines, e.g., multilingual, pivoting, or teacher-student, showing the universal effectiveness of our transfer learning schemes.

∗ Equal contribution.
The rest of the paper is organized as follows. We first review important previous works on pivot-based MT in Section 2. Our three pre-training techniques are presented in Section 3. Section 4 shows main results of our methods with a detailed description of the experimental setups. Section 5 studies variants of our methods and reports the results without source-target parallel resources or with large synthetic parallel data. Section 6 draws conclusion of this work with future research directions.

2 Related Work

In this section, we first review existing approaches to leverage a pivot language in low-resource/zero-resource MT. They can be divided into three categories:

1. **Pivot translation (pivoting).** The most naive approach is reusing (already trained) source→pivot and pivot→target models directly, decoding twice via the pivot language (Kauers et al., 2002; De Gispert and Marino, 2006). One can keep N-best hypotheses in the pivot language to reduce the prediction bias (Utiyama and Isahara, 2007) and improve the final translation by system combination (Costa-Jussà et al., 2011), which however increases the translation time even more. In multilingual NMT, Firat et al. (2016) modify the second translation step (pivot→target) to use source and pivot language sentences together as the input.

2. **Pivot-based synthetic parallel data.** We may translate the pivot side of given pivot-target parallel data using a pivot→source model (Bertoldi et al., 2008), or the other way around translating source-pivot data using a pivot→target model (De Gispert and Marino, 2006). For NMT, the former is extended by Zheng et al. (2017) to compute the expectation over synthetic source sentences. The latter is also called teacher-student approach (Chen et al., 2017), where the pivot→target model (teacher) produces target hypotheses for training the source→target model (student).

3. **Pivot-based model training.** In phrase-based MT, there have been many efforts to combine phrase/word level features of source-pivot and pivot-target into a source→target system (Utiyama and Isahara, 2007; Wu and Wang, 2007; Bakhshaei et al., 2010; Zahabi et al., 2013; Zhu et al., 2014; Miura et al., 2015). In NMT, Cheng et al. (2017) jointly train for three translation directions of source-pivot-target by sharing network components, where Ren et al. (2018) use the expectation-maximization algorithm with the target sentence as a latent variable. Lu et al. (2018) deploy intermediate recurrent layers which are common for multiple encoders and decoders, while Johnson et al. (2017) share all components of a single multilingual model. Both methods train the model for language pairs involving English but enable zero-shot translation for unseen non-English language pairs. For this, Ha et al. (2017) encode the target language as an additional embedding and filter out non-target tokens in the output. Lakew et al. (2017) combine the multilingual training with synthetic data generation to improve the zero-shot performance iteratively, where Sestorain et al. (2018) applies the NMT prediction score and a language model score to each synthetic example as gradient weights.

Our work is based on transfer learning (Zoph et al., 2016) and belongs to the third category: model training. On the contrary to the multilingual joint training, we suggest two distinct steps: pre-training (with source-pivot and pivot-target data) and fine-tuning (with source-target data). With our proposed methods, we prevent the model from losing its capacity to other languages while utilizing the information from related language pairs well, as shown in the experiments (Section 4).

Our pivot adapter (Section 3.2) shares the same motivation with the interlingua component of Lu et al. (2018), but is much compact, independent of variable input length, and easy to train offline. The adapter training algorithm is adopted from bilingual word embedding mapping (Xing et al., 2015). Our cross-lingual encoder (Section 3.3) is inspired by cross-lingual sentence embedding algorithms using NMT (Schwenk and Douze, 2017; Schwenk, 2018).

Transfer learning was first introduced to NMT by Zoph et al. (2016), where only the source language is switched before/after the transfer. Nguyen and Chiang (2017) and Kocmi and Bojar (2018) use shared subword vocabularies to work
with more languages and help target language switches. Kim et al. (2019) propose additional techniques to enable NMT transfer even without shared vocabularies. To the best of our knowledge, we are the first to propose transfer learning strategies specialized in utilizing a pivot language, transferring a source encoder and a target decoder at the same time. Also, for the first time, we present successful zero-shot translation results only with pivot-based NMT pre-training.

3 Pivot-based Transfer Learning

Our methods are based on a simple transfer learning principle for NMT, adjusted to a usual data condition for non-English language pairs: lots of source-pivot and pivot-target parallel data, little (low-resource) or no (zero-resource) source-target parallel data. Here are the core steps of the plain transfer (Figure 1):

1. Pre-train a source → pivot model with a source-pivot parallel corpus and a pivot → target model with a pivot-target parallel corpus.
2. Initialize the source → target model with the source encoder from the pre-trained source → pivot model and the target decoder from the pre-trained pivot → target model.
3. Continue the training with a source-target parallel corpus.

If we skip the last step (for zero-resource cases) and perform the source → target translation directly, it corresponds to zero-shot translation.

Thanks to the pivot language, we can pre-train a source encoder and a target decoder without changing the model architecture or training objective for NMT. On the contrary to other NMT transfer scenarios (Zoph et al., 2016; Nguyen and Chi-ang, 2017; Kocmi and Bojar, 2018), this principle has no language mismatch between transferor and transforee on each source/target side. Experimental results (Section 4) also show its competitiveness despite its simplicity.

Nonetheless, the main caveat of this basic pre-training is that the source encoder is trained to be used by an English decoder, while the target decoder is trained to use the outputs of an English encoder — not of a source encoder. In the following, we propose three techniques to mitigate the inconsistency of source → pivot and pivot → target pre-training stages. Note that these techniques are not exclusive and some of them can complement others for a better performance of the final model.

3.1 Step-wise Pre-training

A simple remedy to make the pre-trained encoder and decoder refer to each other is to train a single NMT model for source → pivot and pivot → target in consecutive steps (Figure 2):

1. Train a source → pivot model with a source-pivot parallel corpus.
2. Continue the training with a pivot-target parallel corpus, while freezing the encoder parameters of 1.

In the second step, a target decoder is trained to use the outputs of the pre-trained source encoder as its input. Freezing the pre-trained encoder ensures that, even after the second step, the encoder is still modeling the source language although we train the NMT model for pivot → target. Without the freezing, the encoder completely adapts to the pivot language input and is likely to forget source language sentences.

We build a joint vocabulary of the source and pivot languages so that the encoder effectively represents both languages. The frozen encoder is pre-trained for the source language in the first step, but also able to encode a pivot language sentence in a similar representation space. It is more effective for linguistically similar languages where
many tokens are common for both languages in the joint vocabulary.

3.2 Pivot Adapter

Instead of the step-wise pre-training, we can also postprocess the network to enhance the connection between the source encoder and the target decoder which are pre-trained individually. Our idea is that, after the pre-training steps, we adapt the source encoder outputs to the pivot encoder outputs to which the target decoder is more familiar (Figure 3). We learn a linear mapping between the two representation spaces with a small source-pivot parallel corpus:

1. Encode the source sentences with the source encoder of the pre-trained source→pivot model.
2. Encode the pivot sentences with the pivot encoder of the pre-trained pivot→target model.
3. Apply a pooling to each sentence of 1 and 2, extracting representation vectors for each sentence pair: \((s, p)\).
4. Train a mapping \(M \in \mathbb{R}^{d \times d}\) to minimize the distance between the pooled representations \(s \in \mathbb{R}^{d \times 1}\) and \(p \in \mathbb{R}^{d \times 1}\), where the source representation is first fed to the mapping:

\[
\hat{M} = \arg\min_{M} \sum_{s, p} \|Ms - p\|^2
\]  

(1)

where \(d\) is the hidden layer size of the encoders. Introducing matrix notations \(S \in \mathbb{R}^{d \times n}\) and \(P \in \mathbb{R}^{d \times n}\), which concatenate the pooled representations of all \(n\) sentences for each side in the source-pivot parallel corpus, we rewrite Equation 1 as:

\[
\hat{M} = \arg\min_{M} \|MS - P\|^2
\]  

(2)

which can be easily computed by the singular value decomposition (SVD) for a closed-form solution, if we put an orthogonality constraint on \(M\) (Xing et al., 2015). The resulting optimization is also called Procrustes problem.

The learned mapping is multiplied to encoder outputs of all positions in the final source→target tuning step. With this mapping, the source encoder emits sentence representations that lie in a similar space of the pivot encoder. Since the target decoder is pre-trained for pivot→target and accustomed to receive the pivot encoder outputs, it should process the mapped encoder outputs better than the original source encoder outputs.

3.3 Cross-lingual Encoder

As a third technique, we modify the source→pivot pre-training procedure to force the encoder to have cross-linguality over source and pivot languages; modeling source and pivot sentences in the same mathematical space. We achieve this by an additional autoencoding objective from a pivot sentence to the same pivot sentence (Figure 4).

The encoder is fed with sentences of both source and pivot languages, which are processed by a shared decoder that outputs only the pivot language. In this way, the encoder is learned to produce representations in a shared space regardless of the input language, since they are used in the same decoder. This cross-lingual space facilitates smoother learning of the final source→target model, because the decoder is pre-trained to translate the pivot language.

The same input/output in autoencoding encourages, however, merely copying the input; it is said to be not proper for learning complex structure of the data domain (Vincent et al., 2008). Denoising autoencoder addresses this by corrupting the input sentences by artificial noises (Hill et al., 2016). Learning to reconstruct clean sentences, it encodes linguistic structures of natural language sentences, e.g., word order, better than copying. Here are the
noise types we use (Edunov et al., 2018):

- Drop tokens randomly with a probability $p_{\text{del}}$
- Replace tokens with a $<\text{BLANK}>$ token randomly with a probability $p_{\text{rep}}$
- Permute the token positions randomly so that the difference between an original index and its new index is less than or equal to $d_{\text{per}}$

We set $p_{\text{del}} = 0.1$, $p_{\text{rep}} = 0.1$, and $d_{\text{per}} = 3$ in our experiments.

The key idea of all three methods is to build a closer connection between the pre-trained encoder and decoder via a pivot language. The difference is in when we do this job: Cross-lingual encoder (Section 3.3) changes the encoder pre-training stage (source $\rightarrow$ pivot), while step-wise pre-training (Section 3.1) modifies decoder pre-training stage (pivot $\rightarrow$ target). Pivot adapter (Section 3.2) is applied after all pre-training steps.

## 4 Main Results

We evaluate the proposed transfer learning techniques in two non-English language pairs of WMT 2019 news translation tasks\(^1\): French $\rightarrow$ German and German $\rightarrow$ Czech.

### Data

We used the News Commentary v14 parallel corpus and newstest2008-2010 test sets as the source-target training data for both tasks. The newstest sets were oversampled four times. The German $\rightarrow$ Czech task was originally limited to unsupervised learning (using only monolingual corpora) in WMT 2019, but we relaxed this constraint by the available parallel data. We used newstest2011 as a validation set and newstest2012/newstest2013 as the test sets.

Both language pairs have much abundant parallel data in source-pivot and pivot-target with English as the pivot language. Detailed corpus statistics are given in Table 1.

### Preprocessing

We used the Moses\(^2\) tokenizer and applied true-casing on all corpora. For all transfer learning setups, we learned byte pair encoding (BPE) (Sennrich et al., 2016) for each language individually with 32k merge operations, except for cross-lingual encoder training with joint BPE only over source and pivot languages. This is for modularity of pre-trained models: for example, a French $\rightarrow$ English model trained with joint French/English/German BPE could be transferred smoothly to a French $\rightarrow$ German model, but would not be optimal for a transfer to e.g., a French $\rightarrow$ Korean model. Once we pre-train an NMT model with separate BPE vocabularies, we can reuse it for various final language pairs without wasting unused portion of subword vocabularies (e.g., German-specific tokens in building a French $\rightarrow$ Korean model).

On the contrary, baselines used joint BPE over all languages with also 32k merges.

### Model and Training

The 6-layer base Transformer architecture (Vaswani et al., 2017) was used for all of our experiments. Batch size was set to 4,096 tokens. Each checkpoint amounts to 10k updates for pre-training and 20k updates for fine-tuning.

Each model was optimized with Adam (Kingma and Ba, 2014) with an initial learning rate of 0.0001, which was multiplied by 0.7 whenever perplexity on the validation set was not improved for three checkpoints. When it was not improved for eight checkpoints, we stopped the training. The NMT model training and transfer were done with the OPENNMT toolkit (Klein et al., 2017).

Pivot adapter was trained using the MUSE toolkit (Conneau et al., 2018), which was originally developed for bilingual word embeddings but we adjusted for matching sentence representations.

### Baselines

We thoroughly compare our approaches to the following baselines:

1. **Direct source $\rightarrow$ target**: A standard NMT model trained on given source $\rightarrow$ target parallel.
Table 2: Main results fine-tuned with source-target parallel data.

<table>
<thead>
<tr>
<th></th>
<th>French→German</th>
<th></th>
<th>German→Czech</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>newstest2012</td>
<td>newstest2013</td>
<td>newstest2012</td>
<td>newstest2013</td>
</tr>
<tr>
<td></td>
<td>BLEU [%]</td>
<td>TER [%]</td>
<td>BLEU [%]</td>
<td>TER [%]</td>
</tr>
<tr>
<td>Direct source→target</td>
<td>14.8</td>
<td>75.1</td>
<td>11.1</td>
<td>81.1</td>
</tr>
<tr>
<td>Multilingual many-to-many</td>
<td>18.7</td>
<td>71.9</td>
<td>14.9</td>
<td>76.6</td>
</tr>
<tr>
<td>Multilingual many-to-one</td>
<td>18.3</td>
<td>71.7</td>
<td>13.1</td>
<td>79.6</td>
</tr>
<tr>
<td>Plain transfer</td>
<td>17.5</td>
<td>72.3</td>
<td>15.4</td>
<td>75.4</td>
</tr>
<tr>
<td>+ Pivot adapter</td>
<td>18.0</td>
<td>71.9</td>
<td>15.9</td>
<td>75.0</td>
</tr>
<tr>
<td>+ Cross-lingual encoder</td>
<td>17.4</td>
<td>72.1</td>
<td>15.0</td>
<td>75.9</td>
</tr>
<tr>
<td>+ Pivot adapter</td>
<td>17.8</td>
<td>72.3</td>
<td>15.6</td>
<td>75.3</td>
</tr>
<tr>
<td>Step-wise pre-training</td>
<td>18.6</td>
<td>70.7</td>
<td>15.6</td>
<td>75.0</td>
</tr>
<tr>
<td>+ Cross-lingual encoder</td>
<td>19.5</td>
<td>69.8</td>
<td>16.2</td>
<td>74.6</td>
</tr>
</tbody>
</table>

In Table 2, we report principal results after fine-tuning the pre-trained models using source-target parallel data.

As for baselines, multilingual models are better than a direct NMT model. The many-to-many models surpass the many-to-one models; since both tasks are in a low-resource setup, the model gains a lot from related language pairs even if the target languages do not match.

Plain transfer of pre-trained encoder/decoder without additional techniques (Figure 1) shows a nice improvement over the direct baseline: up to +2.7% BLEU for French→German and +5.2% BLEU for German→Czech. Pivot adapter provides an additional boost of maximum +0.7% BLEU or -0.7% TER.

Cross-lingual encoder pre-training is proved to be not effective in the plain transfer setup. It shows no improvements over plain transfer in French→German, and 0.4% BLEU worse performance in German→Czech. We conjecture that the cross-lingual encoder needs a lot more data to be fine-tuned for another decoder, where the encoder capacity is basically divided into two languages at the beginning of the fine-tuning. On the other hand, the pivot adapter directly improves the connection to an individually pre-trained decoder, which works nicely with small fine-tuning data.

Pivot adapter gives an additional improvement on top of the cross-lingual encoder; up to +0.4% BLEU in French→German and +0.6% BLEU in German→Czech. In this case, we extract source and pivot sentence representations from the same shared encoder for training the adapter.

Step-wise pre-training gives a big improvement up to +1.2% BLEU or -1.6% TER against plain transfer in French→German. It shows the best performance in both tasks when combined with the cross-lingual encoder: up to +1.2% BLEU in French→German and +2.6% BLEU in German→Czech, compared to the multilingual baseline. Step-wise pre-training prevents the cross-lingual encoder from degeneration, since the pivot→target pre-training (Step 2 in Section 3.1) also learns the encoder-decoder connection with a large amount of data — in addition to the source→target tuning step afterwards.

Note that the pivot adapter, which inserts an extra layer between the encoder and decoder, is not appropriate after the step-wise pre-training: the decoder is already trained to correlate well with the pre-trained encoder. We experimented with the pivot adapter on top of step-wise pre-trained models — with or without cross-lingual encoder — but obtained detrimental results.

Compared to pivot translation (Table 5), our
best results are also clearly better in French → German and comparable in German → Czech.

5 Analysis

In this section, we conduct ablation studies on the variants of our methods and see how they perform in different data conditions.

5.1 Pivot Adapter

Table 3: Pivot adapter variations (German → Czech). All results are tuned with source-target parallel data.

<table>
<thead>
<tr>
<th>Adapter Training</th>
<th>BLEU [%]</th>
<th>TER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>18.2</td>
<td>70.7</td>
</tr>
<tr>
<td>Max-pooled</td>
<td>18.4</td>
<td>70.5</td>
</tr>
<tr>
<td>Average-pooled</td>
<td>18.7</td>
<td>70.3</td>
</tr>
<tr>
<td>Plain transfer</td>
<td>18.0</td>
<td>70.9</td>
</tr>
</tbody>
</table>

Firstly, we compare variants of the pivot adapter (Section 3.2) in Table 3. The row “None” shows that a randomly initialized linear layer already guides the pre-trained encoder/decoder to harmonize with each other. Of course, when we train the adapter to map source encoder outputs to pivot encoder outputs, the performance gets better. For compressing encoder outputs over positions, average-pooling is better than max-pooling. We observed the same trend in the other test set and in French → German.

We also tested nonlinear pivot adapter, e.g., a 2-layer feedforward network with ReLU activations, but the performance was not better than just a linear adapter.

5.2 Cross-lingual Encoder

Table 4: Cross-lingual encoder variations (French → German). All results are in the zero-shot setting with step-wise pre-training.

<table>
<thead>
<tr>
<th>Trained on</th>
<th>Input</th>
<th>BLEU [%]</th>
<th>TER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual</td>
<td>Clean</td>
<td>15.7</td>
<td>77.7</td>
</tr>
<tr>
<td></td>
<td>Noisy</td>
<td>17.5</td>
<td>73.6</td>
</tr>
<tr>
<td>Pivot side of parallel</td>
<td>Clean</td>
<td>15.9</td>
<td>77.3</td>
</tr>
<tr>
<td></td>
<td>Noisy</td>
<td>18.0</td>
<td>72.7</td>
</tr>
</tbody>
</table>

Table 4 verifies that the noisy input in autoencoding is indeed beneficial to our cross-lingual encoder. It improves the final translation performance by maximum +2.1% BLEU, compared to using the copying autoencoding objective.

As the training data for autoencoding, we also compare between purely monolingual data and the pivot side of the source-pivot parallel data. By the latter, one can expect a stronger signal for a joint encoder representation space, since two different inputs (in source/pivot languages) are used to produce the exactly same output sentence (in pivot language). The results also tell that there are slight but consistent improvements by using the pivot part of the parallel data.

Again, we performed these comparisons in the other test set and German → Czech, observing the same tendency in results.

5.3 Zero-resource/Zero-shot Scenarios

If we do not have an access to any source-target parallel data (zero-resource), non-English language pairs have two options for still building a working NMT system, given source-English and target-English parallel data:

- **Zero-shot**: Perform source → target translation using models which have not seen any source-target parallel sentences, e.g., multilingual models or pivoting (Section 2.1).
- **Pivot-based synthetic data**: Generate synthetic source-target parallel data using source ↔ English and target ↔ English models (Section 2.2). Use this data to train a model for source → target.

Table 5 shows how our pre-trained models perform in zero-resource scenarios with the two options. Note that, unlike Table 2, the multilingual baselines exclude source → target and target → source directions. First of all, plain transfer, where the encoder and the decoder are pretrained separately, is poor in zero-shot scenarios. It simply fails to connect different representation spaces of the pre-trained encoder and decoder. In our experiments, neither pivot adapter nor cross-lingual encoder could enhance the zero-shot translation of plain transfer.

Step-wise pre-training solves this problem by changing the decoder pre-training to familiarize itself with representations from an already pre-trained encoder. It achieves zero-shot performance of 11.5% BLEU in French → German and 6.5% BLEU in German → Czech (newstest2013), while
<table>
<thead>
<tr>
<th></th>
<th>French→German</th>
<th>German→Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>newstest2012</td>
<td>newstest2013</td>
</tr>
<tr>
<td><strong>BLEU [%]</strong></td>
<td><strong>TER [%]</strong></td>
<td><strong>BLEU [%]</strong></td>
</tr>
<tr>
<td>Multilingual many-to-many</td>
<td>14.1</td>
<td>79.1</td>
</tr>
<tr>
<td>Pivot translation</td>
<td>16.6</td>
<td>72.4</td>
</tr>
<tr>
<td>Teacher-student</td>
<td>18.7</td>
<td>70.3</td>
</tr>
<tr>
<td>Plain transfer</td>
<td>0.1</td>
<td>-</td>
</tr>
<tr>
<td>Step-wise pre-training</td>
<td>11.0</td>
<td>81.6</td>
</tr>
<tr>
<td>+ Cross-lingual encoder</td>
<td>17.3</td>
<td>72.1</td>
</tr>
<tr>
<td>+ Teacher-student</td>
<td><strong>19.3</strong></td>
<td><strong>69.7</strong></td>
</tr>
</tbody>
</table>

Table 5: Zero-resource results. Except those with the teacher-student, the results are all in the zero-shot setting, i.e., the model is not trained on any source-target parallel data. ‘-’ indicates a TER score over 100%.

showing comparable or better fine-tuned performance against plain transfer (see also Table 2).

With the pre-trained cross-lingual encoder, the zero-shot performance of step-wise pre-training is superior to that of pivot translation in French→German with only a single model. It is worse than pivot translation in German→Czech. We think that the data size of pivot-target is critical in pivot translation; relatively huge data for English→Czech make the pivot translation stronger. Note again that, nevertheless, pivoting (second row) is very poor in efficiency since it performs decoding twice with the individual models.

For the second option (pivot-based synthetic data), we compare our methods against the sentence-level beam search version of the teacher-student framework (Chen et al., 2017), with which we generated 10M synthetic parallel sentence pairs. We also tried other variants of Chen et al. (2017), e.g., N-best hypotheses with weights, but there were no consistent improvements.

Due to enormous bilingual signals, the model trained with the teacher-student synthetic data outperforms pivot translation. If tuned with the same synthetic data, our pre-trained model performs even better (last row), achieving the best zero-resource results on three of the four test sets.

We also evaluate our best German→Czech zero-resource model on newstest2019 and compare it with the participants of the WMT 2019 unsupervised news translation task. Ours yield 17.2% BLEU, which is much better than the best single unsupervised system of the winner of the task (15.5%) (Marie et al., 2019). We argue that, if one has enough source-English and English-target parallel data for a non-English language pair, it is more encouraged to adopt pivot-based transfer learning than unsupervised MT — even if there is no source-target parallel data. In this case, unsupervised MT unnecessarily restricts the data condition to using only monolingual data and its high computational cost does not pay off; simple pivot-based pre-training steps are more efficient and effective.

### 5.4 Large-scale Results

We also study the effect of pivot-based transfer learning in more data-rich scenarios: 1) with large synthetic source-target data (German→Czech), and 2) with larger real source-target data in combination with the synthetic data (French→German). We generated synthetic parallel data using pivot-based back-translation (Bertoldi et al., 2008): 5M sentence pairs for German→Czech and 9.1M sentence pairs for French→German. For the second scenario, we also prepared 2.3M more lines of French→German real parallel data from Europarl v7 and Common Crawl corpora.

Table 6 shows our transfer learning results fine-tuned with a combination of given parallel data and generated synthetic parallel data. The real source-target parallel data are oversampled to make the ratio of real and synthetic data to be 1:2. As expected, the direct source→target model can be improved considerably by training with large synthetic data.

Plain pivot-based transfer outperforms the synthetic data baseline by up to +1.9% BLEU or -3.3% TER. However, the pivot adapter or cross-lingual encoder gives marginal or inconsistent improvements over the plain transfer. We suppose that the entire model can be tuned sufficiently well without
Table 6: Results fine-tuned with a combination of source-target parallel data and large synthetic data. French→German task used larger real parallel data than Table 2.

<table>
<thead>
<tr>
<th></th>
<th>French→German</th>
<th>German→Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>newstest2012</td>
<td>newstest2013</td>
</tr>
<tr>
<td></td>
<td>BLEU [%]</td>
<td>TER [%]</td>
</tr>
<tr>
<td>Direct source→target</td>
<td>20.1</td>
<td>69.8</td>
</tr>
<tr>
<td>+ Synthetic data</td>
<td>21.1</td>
<td>68.2</td>
</tr>
<tr>
<td>Plain transfer</td>
<td>21.8</td>
<td>67.6</td>
</tr>
<tr>
<td>+ Pivot adapter</td>
<td>21.8</td>
<td>67.6</td>
</tr>
<tr>
<td>+ Cross-lingual encoder</td>
<td>21.9</td>
<td>67.7</td>
</tr>
<tr>
<td>+ Pivot adapter</td>
<td>22.1</td>
<td>67.5</td>
</tr>
<tr>
<td>Step-wise pre-training</td>
<td>21.8</td>
<td>67.8</td>
</tr>
<tr>
<td>+ Cross-lingual encoder</td>
<td>21.9</td>
<td>67.6</td>
</tr>
</tbody>
</table>

additional adapter layers or a well-curated training process, once we have a large source-target parallel corpus for fine-tuning.

6 Conclusion

In this paper, we propose three effective techniques for transfer learning using pivot-based parallel data. The principle is to pre-train NMT models with source-pivot and pivot-target parallel data and transfer the source encoder and the target decoder. To resolve the input/output discrepancy of the pre-trained encoder and decoder, we 1) consecutively pre-train the model for source→pivot and pivot→target, 2) append an additional layer after the source encoder which adapts the encoder output to the pivot language space, or 3) train a cross-lingual encoder over source and pivot languages.

Our methods are suitable for most of the non-English language pairs with lots of parallel data involving English. Experiments in WMT 2019 French→German and German→Czech tasks show that our methods significantly improve the final source→target translation performance, outperforming multilingual models by up to +2.6% BLEU. The methods are applicable also to zero-resource language pairs, showing a strong performance in the zero-shot setting or with pivot-based synthetic data. We claim that our methods expand the advances in NMT to many more non-English language pairs that are not yet studied well.

Future work will be zero-shot translation without step-wise pre-training, i.e., combining individually pre-trained encoders and decoders freely for a fast development of NMT systems for a new non-English language pair.

Acknowledgments

This work has received funding from the European Research Council (ERC) (under the European Union’s Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS") and eBay Inc. The work reflects only the authors’ views and none of the funding agencies is responsible for any use that may be made of the information it contains.

References


Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. 2017. A teacher-student framework for zero-


